UNIVERSITY OF CALIFORNIA

Los Angeles

Relations in Human Cognition

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy in Psychology

by

Nicholas Ichien

2023

ABSTRACT OF THE DISSERTATION

Relations in Human Cognition

by

Nicholas Thomas Ichien

Doctor of Philosophy in Psychology

University of California, Los Angeles, 2023

Professor Hongjing Lu, Chair

Abstract: Human thinking relies on the ability to process *relations* between individuals, kinds, properties, and other relations. Explicit relation processing has been invoked to explain our ability to grasp 'cross-domain' analogies between situations whose similarity is driven by a shared relational structure, rather than any similarities among the relata populating each analog (e.g., between the solar system and an atom) and 'cross-modal' analogies between relata spanning different sensory modalities (e.g., sound and vision); generalize relational schemas, categories whose members share some canonical structure (e.g., *things consisting of elements converging on a central location*); or abstract rule-like sequences (e.g., an A-B-A sequence of syllables). At the same time, explicit relation processing requires that a reasoner simultaneously represent a set of individual relata and then bind them to a relational structure. This ability is slow to develop in childhood, and even among adults, it places high demands on working memory. Relations thus

raise a tension between the expressive advantage they confer and the cognitive cost they impose, and this tension suggests that the human *ability* for relation processing does not imply its inevitable use, especially when less-demanding alternatives are available.

The present dissertation confronts this tension and attempts to specify the computational mechanisms by which human reasoners process relations in the face of their cognitive demands. It presents novel research that clarifies how humans make use of explicitly relational thought instead of nonrelational alternatives. In Chapter 1, I start by examining the role of relations in comparison. Cognitive scientists researching analogy have generalized the processes governing analogical comparison, and the representations of relational structure that it operates on, to all comparison. A consequence of this view is that human reasoners make use of relations whenever they make any comparison. I test this claim and show that whereas relations do tend to underlie comparisons aimed at assessing similarity, they tend not to underlie assessments of difference. This asymmetry is consistent with recent accounts of a representational asymmetry between the relations *same* and *different*, in which *different* is represented as a negation of the relation *same* (i.e., *different* is represented as *not-same*). When judging difference, human reasoners are more likely to shift to simpler non-relational representations to ease working memory capacity.

Having lent support to the claim that explicit similarity judgments do tend to incorporate relational information, I extend this claim to implicit similarity comparisons made during recognition in Chapter 2. When an agent attempts to assess whether they recognize a given stimulus, they make an implicit comparison between the perceptually available stimulus to a representation in memory. I show that when agents make this comparison, they tend to incorporate relational information; indeed, relations are available to serve as cues in human recognition memory.

Finally, in Chapter 3, I examine a cognitive process, generative analogical inference, that integrates human reasoning and memory, investigated relatively independently in Chapters 1 and 2 respectively. I introduce a computational model of this process, in which a reasoner uses their prior knowledge of some familiar source domain to elaborate on some less-familiar target domain. This new model can reproduce human-like inference whether the relational structure that constrains inference is prespecified in the model input, as required by existing inference models, or are unspecified, unlike existing models. Across three simulations, I use comparisons between this model and a non-relational control model to clarify what relations contribute to the inference process. Specifically, relations promote far generalization across semantically distance analogs. My dissertation instantiates a framework for studying human relation processing that acknowledges both the expressive advantages that relations provide and the cognitive costs imposed by processing them.

Keywords: relations, concepts, analogy, representation, reasoning, memory

The dissertation of Nicholas Thomas Ichien is approved.

Keith J. Holyoak

Idan Blank

Silvia A. Bunge

Hongjing Lu, Committee Chair

University of California, Los Angeles

2023

Dedication

To my parents Jerry and Helen. I love you so much.

Table of Contents

Introduction **1**

Overview *4*

Chapter 1: Relations and comparison **7**

Introduction *7*

Experiment 1 *10*

Method 10

Results 15

Computational modeling 19

Model specification and approach 20

Modeling results 22

Discussion 23

Experiment 2 *26*

Method 27

Results and Discussion 30

General Discussion *31*

Chapter 2: Relations and recognition **33**

Introduction *33*

Word Embeddings as Predictors of Analogical Reasoning and Word Recognition 34

Experiment 1 *35*

Method 36

Results 43

Experiment 2 *47*

Method 47

Results 52

Tests of Computational Models *53*

Measures of Word-Pair Similarity 53

Modeling Verbal Analogical Reasoning 55

Modeling Recognition Memory 57

General Discussion *74*

Summary 74

Conclusion 76

Chapter 3: Relations and inference 78

Introduction *78*

Analogical inference 80

Eduction of relations 85

Overview of the chapter 86

Computational modeling *87*

Relation representation in BART 87

Generative inference in BART-Gen 91

Simulation 1: Inference using pre-specified relations *98*

Simulation 1a: Jurgens et al. (2012) 98

Simulation 1b: Contribution of relations in Jurgens et al. (2012) 102

Stimulation 2: Inference with unspecified relations in four-term analogies *105*

Simulation 2a: Jurgens et al. (2012) analogies 105

Simulation 2b: Green et al. (2012) analogies 107

Simulation 3: Inference with unspecified relations in extended analogies *110*

General Discussion and Conclusion *114*

General Discussion **116**

Open issues *116*

List of Figures and Tables

List of Figures

Figure 1: Example trials of the verbal and visual comparison tasks. 10

Figure 2: Human accuracy on unambiguous trials of verbal and visual comparison tasks. 15

Figure 3: Relational response rate on ambiguous trials in verbal and visual comparison tasks. 17

Figure 4: Relation-weight parameter values fit to individual participant data, broken down according to comparison type. 23

Figure 5: Relational response rate on story comparison task in comparison and control conditions. 30

Figure 6: Task structure. 38

Figure 7: Human and model-predicted (i.e., relational and lexical) 'valid' responses on the verbal analogy task in Experiment 1. 43

Figure 8: Human false-alarm rates on the recognition memory task in Experiment 1. 44

Figure 9: Task structure for Experiment 2. 47

Figure 10: Human false-alarm rates and model predictions on the recognition memory task in Experiment 2 51

Figure 11: An illustration of relation similarity model (left top panel) and lexical similarity model (left bottom panel), and the resulting 2-D plot of similarity space derived using each (right panel). 53

Figure 12: Human item-level ‘valid’ response rates on verbal analogy problems in Experiment 1, plotted against z-scored distance (dissimilarity) metrics predicted by the relational model (left) and by the lexical model (right). 56

Figure 13: Human false-alarm rates and model predictions on the recognition memory task in Experiment 1, broken down according to familiar, unfamiliar and unrelated stimulus types. 60

Figure 14: Simulation of model-predicted relational luring effect in Experiment 1 as a function of model parameters. 62

Figure 15: Multidimensional scaling based on for lexical similarity among individual first words in pairs used in the memory task for Experiment 1. Colors indicate word-pair relations (category:exemplar, part:whole, and place:thing). 64

Figure 16: Simulation of model-predicted relational luring effect in Experiment 2 as a function of model parameters 67

Figure 17: Mean lexical and relational cosine distances (scaled between 0 and 1) between familiar and unfamiliar word pairs and intact word pairs within each stimulus set used in Experiment 2. 69

Figure 18: Human false-alarm rates from Popov et al. (2017), Experiment 1, and model predictions on the recognition memory task, broken down according to stimulus type. 72

Figure 19: Simulation of model-predicted relational luring effect as a function of model parameters for Popov et al. (2017), Experiment 1. 73

Figure 20: An illustration of three-stage model of BART for learning a relation from word pairs. 89

Figure 21: An illustration of using relation vector to capture distributed representations of semantic relations in BART. 90

Figure 22: Proportion of human-generated responses to two sentence completion problems, constructed from the same statement. These statements are based on the word pair bread:flour and the relation X is made out of Y. 100

Figure 23: Results from Simulation 1a with generative relation problems (e.g., robin is a type of ?), showing median ranks for the most frequent human-generated response, among all human-generated responses across the task (lower ranks indicate better performance). 101

Figure 24: BART-Gen's contribution index to accurate inference in Jurgens et al. (2012). 103

Figure 25: Recall rate as a function of top-k ranked model responses including the most frequent human responses. 106

Figure 26: Recall accuracy plotting the proportion of problems where the most frequent human responses were among top-k ranked responses for BART-Gen (red), Word2vec parallelogram model (solid blue), BERT sentence (dashed blue), and BERT analogy (dotted blue) for k = [1,100]. 109

Figure 27: Distance between model-generated embeddings and the correct answer for each analogy problem, as stipulated in Turney's (2008) materials for BART-Gen (red) and Word2vec parallelogram (blue). Top panel shows performance on scientific analogies from Holyoak and Thagard (1994) and bottom panel on metaphorical analogies from Lakoff and Johnson (2003). Lower distance indicates better performance. 113

Figure 28: Difference in the semantic distance between BART-Gen's generated embedding and the correct answer and the distance between the parallelogram model's generated embedding and the correct answer, as a function of the semantic distance between source and terms. 114

List of Tables

Table 1: Example set of story stimuli drawn from Gentner et al. (1993) 28

Table 2: Properties of each stimulus type used during recognition memory task. 40

Table 3: Example of a stimulus set used in Experiment 2, adapted from Popov et al. (2017). 49

Table 4: Stimuli in Experiment 2 generated from the set presented in Table 2, adapted from Popov et al. (2017) 50

Table 5: GCM parameters fit to human data and fit-model performance for relational similarity (rel) and lexical similarity (lex) 59

Table 6: Log and luring-specific model evidence for GCM using relational similarity (rel) and lexical similarity (lex) averaged over a wide range of the model parameter space (c = [0,50], k = [.1,1]) 61

Table 7: Examples of Jurgens analogy problems. A:B pairs are listed in the leftmost column, and C-terms are listed in the middle column. Most frequent human responses, along with the percentage of participants generating that response are in the rightmost column. 105

Table 8: Examples of Green analogy problems. 107

Table 9: Examples of Turney extended analogies. 111

Acknowledgments

Chapter 1 of this dissertation is an updated version of Ichien, Lin, Holyoak, & Lu (2023) add was supported by NSF Grants IIS-1956441 awarded to H.L. and BCS-2022369 awarded to K.J.H. Chapter 2 has been published in Ichien, Alfred, Baia, Kraemer, Holyoak, Bunge, & Lu (2023) and was supported by NSF Grants BCS-2022477, 2022357, and 2022369, respectively awarded to S.A.B., D.J.M.K., and K.J.H with H.L. My co-authors and I thank Ven Popov for providing raw data reported in Popov et al. (2017). Chapter 3 is an updated version of Ichien, Kan, Holyoak, & Lu (2022) and was supported by NSF Grant BCS-1827374 awarded to K.J.H and IIS-1956441 award to H.L. This dissertation research was supported by UCLA's Dissertation Year Fellowship.

Vita

Education

May 2019	M.A., University of California, Los Angeles (UCLA) Psychology
May 2016	M.Sc., London School of Economics and Political Science (LSE) Philosophy of the Social Sciences Dissertation: *An ameliorative conceptualization of psychiatric disorders*
May 2012	B.A., New York University (NYU) Psychology

Awards / Honors

2022	Cognitive Area Research Award ($6,000), UCLA
2022	Dissertation Year Fellowship ($20,000), UCLA
2019	Graduate Summer Research Mentorship Program ($6,000), UCLA
2018, 2021	Edwin W. Pauley Fellowship ($15,000), UCLA
2016	Distinction for dissertation, LSE
2012	Magna cum Laude, NYU
2012	Founder's Day Award, NYU
2009-2012	Dean's List, NYU

Publications

Peer-reviewed journal articles

Ichien, N.*, Liu, Q.*, Fu, S., Holyoak, K. J., Yuille, A. L., & Lu, H. (under review at *Cognitive Science*). Analogical reasoning with three-dimensional objects: A comparison of deep learning and compositional models.
*****equal author contribution

Holyoak, K. J., **Ichien, N.**, & Lu, H. (under review at *Creativity Research Journal*). Analogy and the generation of ideas.

Stamenkovic, D., Milenkovic, K., **Ichien, N.**, & Holyoak, K. J. (2023). An individual-differences approach to poetic metaphor: Impact of aptness and familiarity. *Metaphor and Symbol*.

Ichien, N., Alfred, K. L., Baia, S., Kraemer, D. J. M., Holyoak, K. J., Bunge, S. A., & Lu, H. (2023). Relational and lexical similarity in analogical reasoning and recognition memory: Behavioral evidence and computational evaluation. *Cognitive Psychology*.

Holyoak, K. J., **Ichien, N.**, & Lu, H. (2022). From semantic vectors to analogical mapping. *Current Directions in Psychological Science*

Lu, H., **Ichien, N.**, & Holyoak, K. J. (2022). Probabilistic analogical mapping with semantic relation networks. *Psychological Review*.

Ichien, N., Lu, H., & Holyoak, K. J. (2022). Predicting patterns of similarity among abstract semantic relations. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *48*(1), 108–121.

Stamenković, D., **Ichien, N.**, & Holyoak, K. J. (2020). Individual Differences in Comprehension of Contextualized Metaphors. *Metaphor and Symbol*, *35*(4), 285-301.

Kadambi, A., **Ichien, N.**, Qiu, S., & Lu, H. (2020). Understanding the visual perception of awkwardness: How greetings go awry. *Attention, Perception, & Psychophysics*. *82*, 2544-2557.

Ichien, N., Lu, H., & Holyoak, K. J. (2020). Verbal analogy problem sets: An inventory of testing materials. *Behavior Research Methods*. *52*(5): 1803-1816.

Peng, Y., **Ichien, N.**, & Lu, H. (2020). Causal actions enhance perception of continuous body movements. *Cognition*. *194*, 104060.

Stamenković, D., **Ichien, N.**, & Holyoak, K. J. (2019). Metaphor comprehension: An individual differences approach. *Journal of Memory and Language*. *105*, 108-118.

Peer-reviewed book chapters

Ichien, N., & Cheng, P. (2022). Revisiting Hume in the 21st century: How forming generalizable causal beliefs is possible when causal relations are inherently unobservable. In A. Wiegmann & P. Willemsen (Eds.), *Advances in Experimental Philosophy of Causation*, London, UK: Bloomsbury Press.

Peer-reviewed conference proceedings papers

Ichien, N., Lin, N.**, Holyoak, KJ. J., & Lu, H. (2023). Asymmetry in the complexity of the relations same and different produces an asymmetry in similarity and difference judgments. *Proceedings of the 45th Annual Meeting of the Cognitive Science Society*. Sydney, Australia: Cognitive Science Society.

Ichien, N., Kan, A.**, Holyoak, K. J., & Lu, H. (2022). Generative inferences in relational and analogical reasoning: A comparison of computational models. *Proceedings of the 44th Annual Meeting of Cognitive Science Society*. Toronto, Canada: Cognitive Science Society.
**advisee

Ichien, N., Alfred, K. L., Baia, S. R., Kraemer, D. J. M., Bunge, S. A., Lu, H., & Holyoak, K. J. (2022). Relation representations in analogical reasoning and recognition memory. *Proceedings of the 44th Annual Meeting of Cognitive Science Society*. Toronto, Canada: Cognitive Science Society.

Snefjella, B., **Ichien, N.**, Holyoak, K. J., & Lu, H. (2022). Predicting human judgments of relational similarity: A comparison of computational models based on vector representations of meaning. *Proceedings of the 44th Annual Meeting of Cognitive Science Society*. Toronto, Canada: Cognitive Science Society.

Lindford, B., **Ichien, N.**, Holyoak, K. J., & Lu, H. (2022). Impact of semantic representations on analogical mapping with transitive relations. *Proceedings of the 44th Annual Meeting of Cognitive Science Society*. Toronto, Canada: Cognitive Science Society.

Ichien, N.*, Liu, Q.*, Fu, S., Holyoak, K. J., Yuille, A., & Lu, H. (2021). Visual analogy: Deep learning versus compositional models. *Proceedings of the 43rd Annual Meeting of Cognitive Science Society*. Vienna, Austria: Cognitive Science Society.
*equal author contribution

Ichien, N., Lu, H., & Holyoak, K. J. (2019). Individual differences in judging similarity between semantic relations. *Proceedings of the 41st Annual Meeting of the Cognitive Science Society*. Montreal, Canada: Cognitive Science Society.

Peng, Y., **Ichien, N.**, & Lu, H. (2019). Perception of continuous movements from causal actions. *Proceedings of the 41st Annual Meeting of the Cognitive Science Society*. Montreal, Canada: Cognitive Science Society.

Lu, H., Liu, Q, **Ichien, N.**, Yuille, A., & Holyoak, K. J. (2019). Seeing the meaning: Vision meet semantics in solving pictorial analogy problems. *Proceedings of the 41st Annual Meeting of the Cognitive Science Society*. Montreal, Canada: Cognitive Science Society.

Introduction

Human intelligence relies on the ability to flexibly represent various forms of structured information. An agent's navigation and adaptive manipulation of their physical environment benefits not only from representations of the objects that populate it, but also from representations of the spatial and functional organization among those objects, of the constituent parts of those objects and their internal organization, and of the hierarchical organization relating object parts, whole objects, and groups of objects (Bapst et al., 2019; P. Battaglia et al., 2016; Biederman, 1987; Kubricht et al., 2017). Similarly, the ability to comprehend and productively generate meaningful linguistic, logical, and algebraic expressions is made much more efficient by the representation of those expressions as constructed from constituent concepts serving particular roles in those expressions, which in turn may be recursively embedded in more complex expressions (Fodor & Pylyshyn, 1988; Marcus, 2003). Finally, useful generalizations that humans exploit in everyday reasoning and problem solving, as well as those leading to technological innovation and scientific discovery, abstract across instances defined at least in part by similar conceptual or visual structures (Gick & Holyoak, 1980; Holyoak & Thagard, 1994a; Kittur et al., 2019).

At the core of the ability to represent structured information is the ability to represent *relations* between individuals, kinds, properties, and other relations. Psychological descriptions of relations have been guided by formal analyses that represent them by a predicate-argument structure (Doumas & Hummel, 2005; Gentner, 1983; Halford et al., 1998; Hummel & Holyoak, 1997, 2003). Under these analyses, explicit relation processing necessitates holding a relation in mind (e.g., *X is taller than Y*) and dynamically (i.e., temporarily) binding that relation to some set of discrete relata (e.g., *Jane* and *Jack*) that each fill a role (e.g., *X*, the taller filler, and *Y*, the shorter filler) in the relation to form some evaluable (i.e., true or false) expression integrating the two (e.g., *Jane is taller than Jack*). (Hummel, 2010, 2011). This dynamic binding enables the same relation to be bound to multiple sets of relata, which permits representation of other evaluable expressions involving the same relation but novel combinations of relata (e.g., *Jack is taller than Judy* and *Jane is taller than Judy*). Relations constitute one kind of mental representation that instantiates a language of thought (LOT; Fodor, 1979; Quilty-Dunn et al., 2022), and so evidence for explicit relation representation in human cognition lends support for the hypothesis that the mind implements a language-like representational system that operates across semantic domains and sensory modalities and whose constituents are able to represent abstract content (e.g., *X is the same as Y*), serve as arguments to logical operations (e.g., negation), and have a predicate-argument structure (as described above).

The proposal that human reasoners actually possess this ability to process relations explicitly seems necessary to explain their facility relative to other animal species in reasoning with relations (Halford et al., 1998; Penn et al., 2008), as well as their ability to grasp cross-domain analogies between situations for which similarity is driven by a shared relational structure, rather than any similarities among the relata populating each analog (e.g., between the solar system and an atom; Gentner, 1983). Explicit relations also underpin cross-modal analogies between sets of relata spanning different sensory modalities (e.g., sound and vision; Hafri et al., 2023; Weinberger et al., 2022). They also allow generalization to form relational schemas— categories whose members share some canonical structure (e.g., *things consisting of elements converging on a central location*; Gick & Holyoak, 1983)—as well as abstract rule-like sequences (e.g., an A-B-A sequence of syllables; Marcus et al., 1999). Moreover, people readily dissociate featural similarity among individual relata from relational similarity among sets of relata (Bassok & Medin, 1997; Medin et al., 1990). Relations prime each other such that processing one set of relata (e.g., *bear:cave*) facilitates subsequent processing of other sequentially presented sets of relata if they instantiate the same relation (e.g., *bird:nest* vs. *bird:desert*; Estes & Jones, 2006; Popov & Hristova, 2015; Spellman et al., 2001). Furthermore, human reasoning is sensitive to variations in a relation's arity—the number of relata that relation connects—suggesting that people continue to process individual relata as discrete fillers when bound to a given relation (Andrews & Halford, 2002; Armstrong, 1978; Halford et al., 1998; Kroger et al., 2004).

The need to simultaneously represent a set of individual relata and then bind them to a relational structure is cognitively demanding, which explains in part why relation processing is relatively slow to develop in childhood (Cowan, 2001; Hummel & Holyoak, 2003; Morrison et al., 2011). Even among adults, explicitly relational thought continues to place high demands on working memory (Bunge et al., 2005; Green et al., 2010; Halford et al., 1998; Kroger et al., 2002, 2004; Waltz et al., 2000), and relation processing takes longer than processing the individual relata to which relations are bound (Goldstone & Medin, 1994).

Relations thus raise a tension between the expressive advantages they confer and the cognitive costs they impose. This tension suggests that the human *ability* for relation processing does not imply its inevitable use, especially when cognitively cheaper alternatives are available. Lurking behind invocations of relational (and, more generally, symbolic) thought is a cacophony of voices affirming the empirical adequacy of non-relational, association-based approaches to human cognition (Leech et al., 2008; Mikolov, Sutskever, et al., 2013; Rescorla & Wagner, 1972; Rumelhart & Abrahamson, 1973; Shanks & Dickinson, 1988). The present dissertation confronts this tension, attempting to clarify the conditions under which human reasoners actually process relations in despite its cognitive demands, and presents novel research that clarifies how humans make use of explicitly relational thought instead of its nonrelational alternatives.

Overview

In Chapter 1, I assess the usage of relations in comparison judgments. A longstanding proposal made by cognitive scientists attempting to clarify how humans make analogical comparisons is that this ability necessitates relation processing, and the most successful computational models incorporate explicit relation representations (Gentner & Forbus, 2011; Holyoak, 2012). Analogy pervades human cognition, including concept and category learning (Carey, 2011; Gentner & Kurtz, 2005; Goldwater & Schalk, 2016), language processing (Ambridge, 2020; Goldwater, 2017; Martin & Doumas, 2020), social reasoning (Hoyos et al., 2020; Kalkstein et al., 2020), explanation (Edwards et al., 2019; Hoyos & Gentner, 2017), and problem solving (Gick & Holyoak, 1980, 1983). It therefore seems possible that *all* comparison is analogical comparison (Gentner & Markman, 1994; Markman & Gentner, 1993a, 1993b; Sagi et al., 2012). I test this claim in light of recent research suggesting that comparison is not the relational monolith that an analogical imperialist would suggest it is. Instead, I provide evidence for an alternative hypothesis that comparisons assessing difference are distinct from those assessing similarity; and that the former involve *negation* and are thus more complex than the latter (i.e., *different* is implemented as *not-same*; Hochmann, 2021; Hochmann et al., 2016, 2018). Specifically, I show that assessments of difference are often too demanding to additionally incorporate relational information, whereas the relative simplicity of assessing similarity makes relational processing more prevalent.

In Chapter 2, I move on to test the presence of relation processing during recognition, which involves an implicit comparison between a representation of a given stimulus and a memory representation. Having shown in Chapter 1 that human reasoners tend to incorporate relational information when explicitly assessing similarity, I examine whether humans use relations as cues when assessing the similarity between what is perceptually available and what is encoded in episodic memory. Popov et al. (2017) argued for the existence of such relation processing, demonstrating an effect that they called *relational luring* in which people sometimes falsely recognize novel word-pair stimuli (e.g., *pipe : water*) when they are analogous to a studied word pair (e.g., *artery : blood*), extending evidence for language-of-thought-style representations to episodic memory (Mahr & Schacter, 2023). In their demonstration of relational luring, Popov et al. did not directly test whether relation processing is, in fact, *necessary* for explaining this effect. I show that a well-established computational model of old/new recognition, Robert Nosofsky's Generalized Context Model (GCM; Nosofsky, 1986), can weakly reproduce this effect when it operates over non-relational lexical representations, but requires explicit relation representations to reproduce the effect strongly and robustly. These simulations encourage caution when attributing relation processing on the basis of a given behavioral phenomenon, especially when demonstrated using verbal materials in which non-relational representations may contain unexpectedly "relational" content. However, they ultimately do support Popov et al.'s original claim that relational luring constitutes evidence that relations are encoded and retrieved in episodic memory.

Finally, in Chapter 3, I introduce a computational model that incorporates relations as constraints on generative analogical inference (i.e., the process of generating a response to a problem such as the analogy *up:down :: fast:?*). The process of analogical generation requires cognitive process that integrate reasoning and memory, which were investigated separately in Chapters 1 and 2, respectively. This model goes beyond existing models of analogical inference, which all rely on explicit relations to be pre-specified in their input (Burstein, 1983; Carbonell, 1983, 1993; Falkenhainer et al., 1989; Halford et al., 1994; Hofstadter & Mitchell, 1994; Holyoak et al., 1994; Holyoak & Thagard, 1989; Hummel & Holyoak, 2003; Keane & Brayshaw, 1988; Kokinov, 1994). In contrast, the proposed model can simulate inference with or without relations being pre-specified in its input. In assuming that relations governing analogical inference are pre-specified, existing models of analogical inference make the logically prior assumption that humanlike inference requires explicit representations of relations. But in principle, analogical inference could be performed without explicit relation representations, and this proposal has been explored to a small extent with limited success (Leech et al., 2008; Mikolov, Sutskever, et al., 2013; Peterson et al., 2020; Rumelhart & Abrahamson, 1973). In this chapter, I compare the proposed model, which operates on explicit relation representations, with control models lacking such representations. In a series of simulations, I systematically show that relations contribute to inference by enabling generalization across semantic domains; allowing analogies across a source and target that are robust to variations in the similarity or degree of association between the source and target (Doumas & Hummel, 2005; Gentner, 1983; Holyoak, 2012).

Overall, my dissertation tests the usage of explicit relations during cognitive processes emphasizing reasoning (Chapter 1), memory (Chapter 2), and an integration of the two (Chapter 3). Ultimately, I find evidence for the usage of relations in explicit comparisons assessing similarity but not difference, in implicit comparisons involved in recognition, and in analogical inference. More importantly, my dissertation instantiates a perspective on relation processing that takes seriously the tension between its expressive value and its cognitive demands.

Chapter 1: Relations and comparison

Introduction

A naïve construal of *similarity* and *difference* is that one is the inverse of the other: As things become more similar, they become less different. Cognitive scientists, however, have demonstrated that human reasoners process the two relations in a way that violates this inverse relation. Specifically, people tend to use different information when judging what makes things similar than when judging what makes things different (Bassok & Medin, 1997; Medin et al., 1990; Simmons & Estes, 2008; Tversky, 1977). For example, Medin et al. (1990) asked participants to select which of two options was more visually similar to or more different from a standard. Across trials, one option was relationally more similar to the standard and the other was more featurally similar. Participants tended to select the relationally similar option as both more similar and more different from the standard. Bassok and Medin (1997) found the same asymmetry using verbal stimuli. Broadly, these findings indicate that people tend to consider relations more heavily when judging similarity than when judging difference. However, the reason for this asymmetry remains unclear.

One attempt to explain this phenomenon invokes structure mapping theory (Gentner, 1983). Under this hypothesis, assessments of similarity and difference both consist in analogical comparison and involve the same comparison process of structural alignment, in which representations of entity features and their structural relations are placed into one-to-one correspondence (Gentner & Markman, 1994; Markman, 1996; Markman & Gentner, 1993a; Sagi et al., 2012). The asymmetry observed by Medin et al. (1990) is hypothesized to arise from an asymmetry in the relevant output of this comparison process. Whereas all commonalities contribute to similarity judgments, differences are split into *alignable* differences (i.e., those filling corresponding roles within a shared relational structure) and *nonalignable* differences (i.e., those not based on corresponding roles). For example, in a comparison between a car and a bicycle, wheel number would be an alignable difference (i.e., 4 vs. 2), whereas window number would be a nonalignable difference because this feature is only applicable to cars and not bicycles.

Proponents of this explanation noted that the featurally-similar option in the study by Medin et al. (1990) did not involve a salient relation, so that any relational difference between it and the standard did not constitute an alignable difference, and was therefore ignored in difference comparisons. However, later work found that both alignable and nonalignable differences contribute to judgments of difference, and that the latter actually exerted a *greater* influence than the former (Estes & Hasson, 2004). This results casts strong doubt on the core assumption that allowed structural alignment theory to potentially account for asymmetries in similarity and difference judgments.

As an alternative explanation, I propose this asymmetry emerges from a representational asymmetry between the relations *same* and *different*. Whereas assessing similarity involves a relatively straightforward comparison of degree of *sameness*, assessing difference involves a more complex comparison of *not-sameness*. As a result, assessments of difference involve greater processing demands than do assessments of similarity. This analysis has been used to explain the well-established developmental lag between children's understanding of the concepts *same* vs. *different* (Hochmann, 2021; Hochmann et al., 2016, 2018).

In general, processing of negation tends to place additional cognitive load on human reasoning. For example, determining the truth of a proposition including a negated expression (e.g., "star isn't above the plus") takes longer than a matched positive expression (e.g., "star is below the plus") (P. A. Carpenter & Just, 1975; Clark & Chase, 1972). Introducing more negation into sentences makes them more difficult to interpret (e.g., “Because he often worked for hours at a time, **no one** believed that he was **not capable** of sustained effort”) (Sherman, 1976). Previous research has shown that processing negation often involves multiple steps, including processing the affirmative components of negated phrases before processing the entire phrase (Hasson & Glucksberg, 2006). Although the complexity of negation is most pronounced when an explicit negative such as *not* is used, processing difficulty is also increased for expressions that incorporate implicit negation (e.g., words such as *few*, *little*, or *deny*) (Clark, 1976).

When human reasoners compare entities, they tend to do so on the basis of both features of individual entities, and also relations between entities and their component parts. Importantly, processing and comparing relational information is more cognitively demanding than processing featural information (Bunge et al., 2005; Green et al., 2010; Halford et al., 1998; Kroger et al., 2002, 2004; Waltz et al., 2000). It follows that incorporating relational information will be particularly demanding when the task also involves negation. As a consequence, difference judgments—which involve implicit negation—are less likely to take account of relational information.

In the following experiment, I tested this processing-demand hypothesis for both verbal comparisons between word pairs and visual comparisons between sets of geometric shapes. For both types of stimuli, I measured participants’ sensitivity to featural and relational information in a 2-alternative forced-choice task, in which participants selected which of two options was more similar to or more different from a standard. In order to directly examine the relative difficulty of similarity and difference judgments, I included unambiguous comparisons, in which one option was unambiguously more similar to a standard than the other based either on features or on relations I predicted that even for unambiguous trials, participants would have greater difficulty in detecting relational difference compared to relational similarity. I also included ambiguous comparisons, for which either of the options might be selected depending on whether features or relations are emphasized (Bassok & Medin, 1997; Medin et al., 1990). I predicted that when judging difference as compared to similarity, participants would tend to base their choices on features rather than relations.

Experiment 1

Method

Participants. Participants were 184 undergraduates (M_{age} = 20.70, SD_{age} = 3.73, range = [18, 51]) at UCLA. This sample consisted of 3 nonbinary, 128 female, and 51 male participants; 2 participants did not report their gender. All participants completed experimental tasks online to obtain partial course credit in a psychology class. The study was approved by the Institutional Review Board at UCLA.

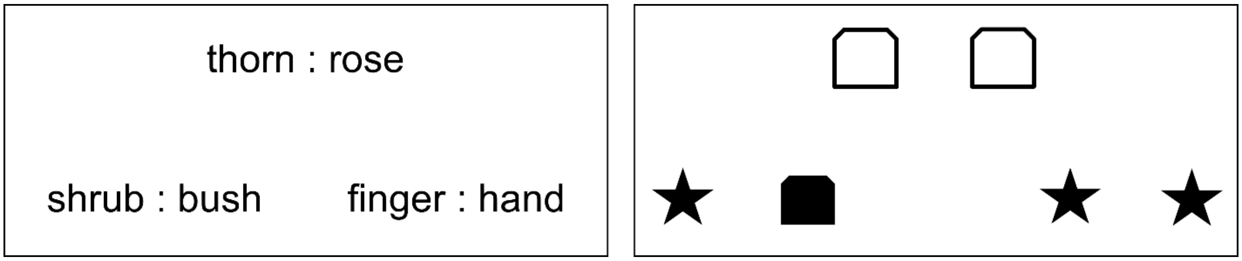


Figure 1: Example trials of the verbal and visual comparison tasks.
In both examples, the left bottom option is more featurally similar to but more relationally different from the standard at the top, whereas the right option is more featurally different from but more relationally similar to the standard.

Comparison tasks. All participants completed two comparison tasks, a verbal task featuring word-pair stimuli and a visual task featuring geometric shape stimuli. On each trial, participants were presented with a standard at the top of the screen and two options on either side at the bottom of the screen. Figure 1 shows an example trial of the verbal task on the left and the visual task on the right. Some participants were instructed to select which option was more *similar* to the standard across both tasks, whereas other participants were asked to select which was more *different* from the standard across both tasks.

Each comparison task consisted of 24 trials, presented in a random order. Of these, 6 *unambiguous* trials included one option that was unambiguously more similar to the standard than the other. On half of the unambiguous trials, the similar option was more featurally similar to the standard than the other option, whereas both options were equally relationally similar to the standard. I refer to these as *featural* trials, The other 3 unambiguous trials were *relational* trials. On these, the similar option was more relationally similar to the standard, whereas both options were equally featurally similar to the standard.

Unambiguous trials enabled us to compare the difficulty of incorporating featural and relational information in similarity and difference judgments. Failure to select the similar option on featural trials would reflect a difficulty with incorporating featural similarity, whereas failure to select the similar option on relational trials would reflect a difficulty with incorporating relational similarity. I expected that relational trials would be more cognitively demanding, and hence prove more difficult for participants judging difference as compared to similarity. On the other hand, since featural trials could be successfully completed without any relation processing, performance for difference versus similarity judgments was expected to be more equal.

The remaining 18 trials consisted of one option that was more featurally similar to but relationally different from the standard (FS/RD; e.g., the left option of both trials depicted in Figure 1) than the other option, which was more featurally different from but relationally similar to the standard (FD/RS; e.g., the right options of both trials in Figure 1). I refer to these trials as *ambiguous* trials because they were constructed so that selecting either option was valid, depending on a participant’s criteria for judging similarity or difference. I used these trials to compare participants’ preferential weighting of featural or relational information in their similarity and difference judgments. Selecting the FS/RD option as more similar indicates a preferential weighting of featural information, whereas selecting it as more different indicates a preferential weighting of relational information, and vice versa for selecting the FD/RS option. I hypothesized that since difference judgments require more complex comparisons than similarity judgments, participants would weight featural similarity more heavily for difference judgments in order to ease the cognitive demands of the comparison. I therefore predicted that participants would select the FD/RS option with greater frequency than the FS/RD option when judging both similarity and difference: Similarity participants would select the former option on the basis of relational similarity, whereas difference participants would do so on the basis of featural difference.

For the verbal task, featural similarity was determined by the semantic similarity among the individual words in each word pair. The left panel of Figure 1 shows an example of an ambiguous trial of the verbal task. The individual words composing the standard (*thorn* and *rose*) and those composing the right option (*shrub* and *bush*) all refer to concepts related to garden plants, and thus are more semantically similar than the words composing the left option (*finger* and *hand*), which are generally less semantically similar to those in the standard.

Relational similarity was determined by the semantic relation instantiated by each word pair. Returning to the left panel of Figure 1, the standard (*thorn:rose*) and the left option (*finger:hand*) both instantiate the semantic relation *part-of*, and are thus more relationally similar to each other than the standard is to the right option (*shrub:bush*), which most saliently instantiates an *instance-of* relation (which does not match the standard’s relation). In addition to *part-of* and *instance-of* relations, verbal comparison trials featured *antonym* (e.g., *love:hate*), *synonym* (e.g., *big:large*), *category coordinate* (e.g., *broom:mop*), and *located-in* (e.g., *grill:patio*) relations.

Importantly, all FS/RD options in the verbal comparison task saliently instantiated one of the relations listed above. For instance, on one trial, participants were given the standard *hoof:horse* and asked to choose between the FS/RD option *goat:cow* and the FS/RS option *wheel:bicycle*. All three word pairs instantiate some binary semantic relation (either *part-of* or *category coordinate*). Accordingly, the relations constitute an alignable difference. Structure mapping theory therefore predicts that mismatching relations (e.g., between *hoof:horse* and *goat:cow*) will contribute to difference judgments just as much as do mismatching features (Gentner & Markman, 1994; Markman, 1996). Structure mapping theory thus predicts symmetric responding for similarity and difference judgments on ambiguous trials: Participants should select all options with the same frequency, regardless of whether they are judging similarity or difference.

For the visual comparison task, featural similarity was determined by a shared salient visual feature among individual objects, either *shape* (as with the left option in the right panel of Figure 1) or *shading*. Relational similarity was determined by the visual relation instantiated by each set of shapes. Most of the visual comparison trials were comparable to the one presented in the right panel of Figure 1, where the standard and the FD/RS option (right) instantiated the *same* relation and each consisted of repetitions of different shapes, while the FS/RD option (left) violated the standard's *same* relation but instantiated a *same-shading* relation and shared one object of the same shape as the standard. Other visual relations featured in this task included *symmetry*, consisting of two identical objects reflected about a vertical axis; *ABA sequences* consisting of three objects, of which the first and last were identical to each other; *ABC sequences* consisting of three unique objects; and *AABB sequences* consisting of two repetitions of different objects. I acknowledge that some FS/RD options in the visual comparison task may not have been interpreted as instantiating a relation, so performance on this test does not constitute as strong a test of the structure mapping theory as does the verbal comparison task.

Ravens Progressive Matrices. Following the verbal comparison task, all participants completed an abridged, 12-problem version of the Ravens Advanced Progressive Matrices (RPM) (Arthur et al., 1999). On each problem in this task, participants are presented with a 3x3 array of simple geometric objects, with the object in the bottom-right corner of the array missing, and they are asked to select which one of 8 options best completes the pattern instantiated by the incomplete array. Carpenter et al. (1990) showed that individual differences in performance on these visual reasoning problems predict differences in the ability to induce abstract relations between objects and to maintain a hierarchy of problem goals and subgoals in working memory. I used this test as a measure of individual differences in general reasoning ability. Since my key manipulation of comparison type (similarity vs. difference) was between-subjects, I included RPM score as a covariate in analyses, in order to compare performance on similarity versus difference judgments after controlling for any individual differences in general reasoning ability.

Procedure. All participants completed a verbal comparison task and a visual comparison task in a counterbalanced order, and then completed the Ravens Progressive Matrices.

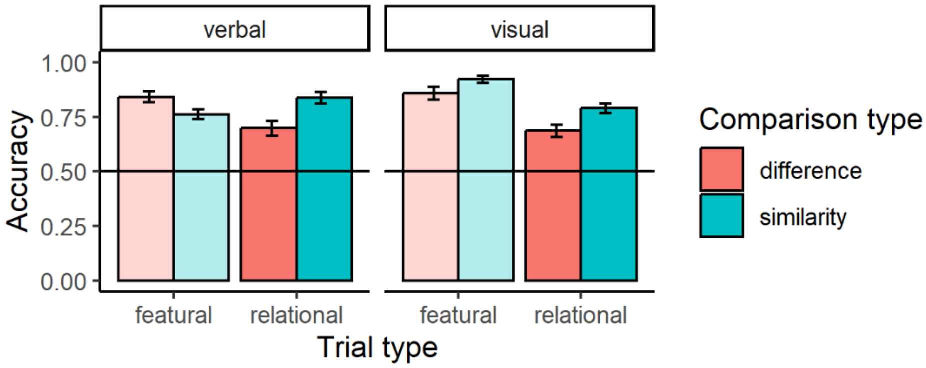


Figure 2: Human accuracy on unambiguous trials of verbal and visual comparison tasks.
Accuracy is broken down according to trial type (featural vs. relational) and comparison type (difference vs. similarity). Error bars reflect ± standard error of the mean, and horizontal line reflects chance performance.

Results

Performance on unambiguous trials. Performance on unambiguous trials across conditions is depicted in Figure 2. Overall, participants performed well on unambiguous trials. Those making similarity judgments (n = 98) frequently selected the more similar option for both the verbal task (M_{sim} = .80, SD_{sim} = .17) and the visual task (M_{sim} = .86, SD_{sim} = .14). Those making difference judgments (n = 86) frequently selected the more different option across both tasks (verbal: M_{diff} = .77, SD_{diff} = .21; visual: M_{diff} = .77, SD_{diff} = .22). Hereafter, I refer to the responses described above as ‘accurate’ responses. Of particular interest was the relative accuracy with which similarity and difference participants completed relational trials. I hypothesized that assessing relational difference is more overtly cognitively demanding than assessing relational similarity, and so I predicted that participants making difference judgments would perform less accurately on relational trials than those making similarity judgments. On the other hand, I did not anticipate a corresponding performance difference for featural trials.

I used the *glmer* function from version 1.1.26 of the LME4 R package (Bates et al., 2015) in R version 4.1.1 (R. Core Team, 2021) to fit a logistic mixed-effects model to performance on unambiguous trials. I defined a full model including *participant* and *comparison problem* as random intercept effects; *comparison task* (*verbal* vs. *visual*), *comparison type* (*similarity* vs. *difference*) and *trial type* (*featural* vs. *relational*), as well as an interaction between the last two as fixed effects. As discussed previously, I included *RPM score* as a covariate, along with *task order* (*verbal first* vs. *visual first*) and *trial number*. The latter two variables respectively account for any impact of task order and any potential improvement in performance across trials within each task.

I used likelihood-ratio tests to compare this full model to reduced models that omitted a term of interest but that was otherwise equivalent to the full model. First, I tested whether performance generally differed across verbal and visual tasks. To do so, I fit a reduced model to the data that lacked the *comparison task* term but that was otherwise equivalent to the full model. I used a likelihood ratio test to compare the full model to the reduced model and found that removing the *comparison task* term did not increase model prediction error, ΔAIC = -1.40, χ^2 (1) = .65, p = .420. This result indicates that verbal and visual tasks did not differ in their overall difficulty.

Next, I tested the processing-demand hypothesis's prediction that relational trials would be more difficult for participants judging difference than for those judging similarity. In order to do to so I compared the full model to a reduced model that lacked the *judgment type x trial type* interaction term (but that retained the individual terms for *judgment type* and *trial type*). Dropping the interaction term did increase model prediction error, ΔAIC = 10.7, χ^2 (2) = 14.66, p < .001, indicating that performance differences between participants making similarity judgments and difference judgments varied across featural and relational trials. To examine this interaction further, I used the *emmeans* and *pairs* functions from version 1.8.4 of the emmeans R package (Lenth, 2023) to compare the relevant estimated marginal means of the full model. Across verbal and visual tasks, similarity participants (M = .81, SD = .18) outperformed difference participants

(*M* = .69, *SE* = .22) on relational trials, *z* = 4.81, *p* < .001, but not on featural trials, *z* = .04, *p* = .966 (similarity: *M* = .84, *SD* = .14; difference: *M* = .84, *SD* = .20). This result supports the processing-demand hypothesis's prediction that difference judgments involve more complex comparisons than similarity judgments, which particularly impact relational trials. Notably this difference in performance persisted even after I accounted for individual differences in reasoning ability by including *RPM score* as a covariate in the full model. Indeed, a likelihood ratio test comparing the full model and a reduced model that lacked the *RPM score* term showed that removing that term increased model prediction error, $\Delta AIC = 13.5$, χ^2 (1) = 15.56, *p* < .001. Thus, even though general reasoning ability influenced performance on unambiguous trials, comparison type impacted performance specifically on relational trials, over and above individual differences in this ability.

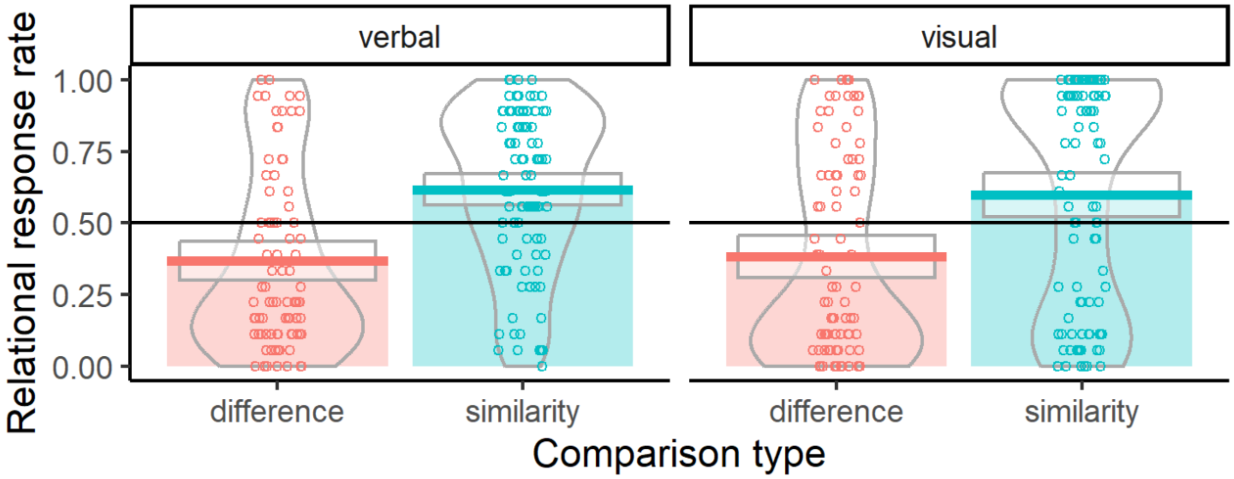


Figure 3: Relational response rate on ambiguous trials in verbal and visual comparison tasks.
Response rates are broken down according to comparison type (difference vs. similarity). Unfilled circles each reflect an individual participant's response rates, dark lines reflect mean response rates, box boundaries reflect ± standard error of the mean, and horizontal line corresponds to indiscriminate selection of relational versus featural options.

Relational responding on ambiguous trials. Having confirmed that detecting relational difference was more difficult on unambiguous trials than was detecting relational similarity, I went on to examine participants' preferential weighting of featural and relational information in ambiguous comparisons for which the two kinds of information are pitted against each other. Overall, participants selected the FD/RS option more often regardless of whether they were judging similarity (*M* = .61, *SD* = .29) or difference (*M* = .62, *SD* = .26). Notably, selecting this option implies different criteria based on comparison type: Selecting FD/RS as more similar implies an emphasis on *relational* similarity, whereas selecting that option as more different implies an emphasis on *featural* difference. In order to assess participant responses across comparison types (similarity vs. difference), I grouped responses according to whether they indicated an emphasis on *relational* information. I thus compared responses in which similarity participants selected the FD/RS option and in which difference participants selected the FS/RD option, and refer to these as *relational* responses.

As with unambiguous trials, I fit logistic mixed-effects models to predict relational responses on ambiguous trials. I defined a full model including *participant* and *comparison problem* as random intercept effects; *comparison task* (*verbal* vs. *visual*), *comparison type* (*similarity* vs. *difference*) as fixed effects; and *RPM score*, *task order* (*verbal first* vs. *visual first*), and *trial number* as covariates.

As was done for unambiguous trials, I used likelihood-ratio tests to compare this full model to reduced models that omitted a term of interest but that was otherwise equivalent to the full model. First, I compared the full model to a reduced model omitting the *comparison task* term. I found that dropping this term did not reduce model prediction error, ΔAIC = -2.0, χ^2 (1) = .01, *p* =

.930. This result indicates that relational responding did not differ across verbal and visual comparison tasks.

Next, I compared relational response rates for similarity judgments and difference judgments, to test my main prediction that participants will preferentially weight relational information more when judging similarity than when judging difference. Indeed, dropping the *comparison type* term from the full model did increase prediction error, $\Delta AIC = 33.3$, χ^2 (1) = 35.31, $p < .001$, which confirms the prediction that relational response rates were affected by comparison type on ambiguous trials. As on unambiguous trials, this effect on ambiguous trials held even after I accounted for individual differences in reasoning ability by including *RPM score* as a covariate in the full model. Omitting *RPM score* from the full model also increased model prediction error, $\Delta AIC = 2.6$, χ^2 (1) = 4.60, $p = .032$. Thus even though individual differences in reasoning ability predicted relational responding on ambiguous trials, my manipulation of comparison type impacted responses over and above these individual differences.

This result disconfirms structure mapping theory's hypothesis that both similarity and difference judgments are based on the same inputs to a structural alignment process (Gentner, 1983; Gentner & Markman, 1994; Markman & Gentner, 1993a; Sagi et al., 2012). According to that theory, similarity judgments are based on all commonalities, whereas differences are sensitive to alignable but not nonalignable differences. In the present study, however, all relational differences on the verbal task (and possibly the visual task) were alignable, so structure mapping theory erroneously predicts symmetric responding across similarity and difference judgments.

Computational modeling

In order to formally characterize the human comparison process on ambiguous trials, I attempted to predict responses of individual participants on the verbal comparison task using a computational model. This model includes a weighting mechanism that controls the relative contribution of relational and featural information to a comparison judgment. I predicted that this weighting mechanism would create the observed asymmetry by altering the emphasis on relational information between similarity and difference judgments. Moreover, the computational model operates entirely on semantic representations of words and relations generated by machine learning, avoiding any hand-coding or reliance on experimenters' intuitions. The same framework could be applied to visual judgments, given an appropriate front-end to generate representations of visual stimuli.

Model specification and approach

Recall that the comparison task dissociated featural and relational information, and that the verbal task involved comparisons between word pairs (e.g., *love:hate* and *spouse:partner*). I operationalized featural information as individual word meanings (e.g., *love, hate*, *wide*, and *narrow*) and relational information as semantic relations holding between paired words (e.g., *antonym-of*, *synonym-of*). I present a computational model that incorporates semantic representations of both individual words and relations between them.

In order to represent individual word meanings, I used pre-trained *Word2vec* word embeddings (Mikolov, Chen, et al., 2013), which represent word meanings as high-dimensional vectors of length 300. These vectors constitute the hidden layer of activation within a neural network trained to predict patterns of text in sequence as they appear in a large corpus consisting of Google News articles (100 billion words). Despite their sole reliance on the statistical distribution of text in their training corpora, these word embeddings and others constitute psychological models of semantic memory in that they preserve the similarity structure of individual word meanings in a psychologically realistic way. These embeddings have been used to successfully model a number of cognitive processes beyond similarity judgments, including human memory search, categorization, and decision making (Bhatia & Aka, 2022; Günther et al., 2019). To compute lexical similarity, the meaning of a word pair is represented by a simple aggregate of the semantic vectors of the two individual words. I use *A* to denote the first word in a word pair and *B* to represent the second word in a word pair. I compute the featural similarity between two word pairs i and j as the cosine similarity between concatenated word vectors constituting *i*, $[f_{A_i} f_{B_i}]$, and those constituting *j*, $[f_{A_j} f_{B_j}]$:

$$sim_{feat_{ij}} = 1 - cos\ \left([f_{A_i} f_{B_i}], [f_{A_j} f_{B_j}]\right). \qquad (1)$$

To compute relational similarity, I used representations generated by *Bayesian Analogy with Relational Transformations* (BART), a learning model that has been used to predict human analogy performance and graded judgments of relational similarity (Ichien et al., 2022; Lu et al., 2012, 2019). BART assumes that specific semantic relations between words are coded as distributed representations over a set of abstract relations. The BART model takes concatenated pairs of Word2vec vectors as input, and then uses supervised learning with both positive and negative examples to acquire representations of individual semantic relations. I use a version of BART that was trained on two datasets consisting of human-generated word pair examples in order to learn a total of 270 semantic relations (Jurgens et al., 2012; Popov et al., 2017).

After learning, BART calculates a relation vector consisting of the posterior probability that a word pair instantiates each of the learned relations. BART uses its pool of learned relations to create a distributed representation of the relation(s) between any two paired words A and B. The posterior probabilities calculated for all learned relations form a 270-dimensional relation vector R_{AB}, in which each dimension codes how likely a word pair instantiates a particular relation. The relational similarity between word pairs i and j is computed as the cosine similarity of the corresponding relation vectors R_i and R_{j}.

Having characterized both featural and relational similarity, I now combine these components simply as a weighted sum in a computational model of comparison,

$$sim_{ij} = \alpha(sim_{feat_{ij}}) + (1-\alpha)sim_{rel_{ij}} \qquad (2)$$

$$diff_{ij} = -\alpha(sim_{feat_{ij}}) - (1-\alpha)sim_{rel_{ij}}, \qquad (3)$$

where α is a free parameter that reflects the degree to which a comparison weights relational information. I refer to this as the relation-weight parameter. Note that both similarity and difference judgments are based on a computation of similarity: difference judgments simply negate the output of that computation.

Modeling results

I used the model to generate trial-level predictions for each participant. I fit the relation-weight parameter to each participant's data by maximizing the accuracy with which the model predicted each response. If multiple values of the relation-weight parameter predicted a participant's data equally well, I took the mean of those parameter values. Overall, the fit model predicted participant responses just as well across similarity judgments (M_{Acc} = .64; SD_{Acc} = .09) and difference judgments (M_{Acc} = .64; SD_{Acc} = .08). The value of the fit relation-weight parameter predicted the rate with which similarity participants selected FD/RS options (Spearman's ρ = .82, p < .002), and the rate with which difference participants selected FS/RD options (Spearman's ρ = .73, p < .001).

I predicted that the value of the relation-weight parameter would be greater when fit to participants making similarity judgments than when fit to those making difference judgments. Figure 4 shows the distribution of the parameter, broken-down according to comparison type. A

Mann-Whitney U test confirmed what is clear from visual inspection: Fit relation-weight parameters were reliably greater for similarity participants than for difference participants, *W* = 2540.5, *p* < .001. This result further supports my main claim: similarity judgments prompt greater reliance on relational information than do difference judgments. Moreover, these simulations support the validity of my manipulation of featural and relational similarity.

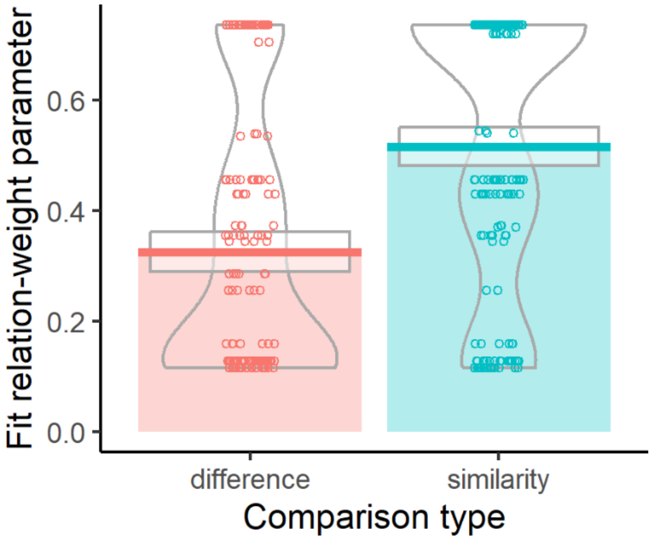


Figure 4: Relation-weight parameter values fit to individual participant data, broken down according to comparison type.

Discussion

For both visual and verbal comparisons, I showed that (1) human reasoners have greater difficulty processing relational difference than they do relational similarity, and (2) they tend to weight relational information more heavily when judging similarity than when judging difference. Moreover, this asymmetry could be accounted for by a computational model of comparison based on machine-generated vector representations for both words and their semantic relations. When fit to human data at the level of individual participants, this model tends to weight relational information more heavily when fit to similarity judgments than when fit to difference judgments.

Contrary to the prediction derived from structure-mapping theory, the asymmetry between similarity and difference judgments was obtained even though all relational differences in my verbal stimulus set were alignable. Notably, this explanation diverges from an alternative account based on structure-mapping theory (Gentner, 1983; Gentner & Markman, 1994; Markman, 1996; Markman & Gentner, 1993a; Sagi et al., 2012), a prominent theory of comparison. Instead of emphasizing the respective cognitive demands of similarity and difference judgments as I do, structure-mapping theory emphasizes a dissociation in the output of a unified comparison process: Whereas the similarity between two entities is based on all commonalities shared among those entities, the difference between two entities privileges *alignable* differences (i.e., those filling corresponding roles within a shared relational structure; e.g., the number of wheels in a car and bicycle) and ignores *nonalignable* differences (i.e., those not based on corresponding roles; e.g., that a car has windows but a bicycle does not). Putative demonstrations that similarity judgments incorporate more relational information than do difference judgments (e.g., Bassok & Medin, 1997; Medin et al., 1990), are instead attributed to relational information's contingent status as a nonalignable difference. However, contrary to the prediction derived from structure-mapping theory, the present study demonstrated an asymmetry between similarity and difference judgments, even though all relational differences in the verbal stimulus set were alignable. I acknowledge that in the present study, I did not directly test whether nonalignable differences contribute to difference judgments. However, when Estes and Hasson (2004) did precisely this, comparing the influence of alignable and nonalignable differences on comparison judgments, they showed not only that nonalignable differences impacted both similarity and difference judgments but also that they actually had greater, not lesser impact than did alignable differences. Overall, this set of findings provides convergent evidence for the claim that assessments of difference are more cognitively demanding than assessments of sameness (Hochmann, 2021; Hochmann et al., 2016, 2018).

Beyond the relative emphasis on alignable and nonalignable differences in structure mapping theory and the processing-demand hypothesis defended here, a more general way to express the difference between these two accounts of comparison is where they locate the dissociation between similarity and difference comparisons. Structure-mapping theory proposes that judgments of both similarity and difference involve in a unified comparison process, structural alignment, that consistently operates over the same representations (i.e., representations of relational structure) (Gentner, 1983; Gentner & Markman, 1994). Any divergence between similarity and difference judgments is then attributed to asymmetries in the usage of the *output* of the comparison process (i.e., all commonalities vs. alignable differences). On the other hand, the processing-demand hypothesis proposes that comparisons of similarity and difference operate on distinct representations; comparisons of similarity tend to operate on representations that incorporate more relational information than do comparisons of difference. And so, the present explanation locates the dissociation between similarity and difference judgments observed in Experiment 1 both in the comparison process (i.e., *same*(X,Y) versus *not-same*(X,Y)) and the representations it operates over. Structure-mapping theory and my processing-demand hypothesis thus make distinct predictions about the extent that any asymmetry in similarity and difference judgments reflects the representations compared in order to arrive at those judgments: Whereas the processing-demand hypothesis proposes a direct link between this response asymmetry and the representations compared, structure-mapping theory proposes no such link. I test these competing hypotheses in the next experiment.

Experiment 2

In Experiment 2, I manipulate whether or not participants are prompted to process experimental stimuli before comparing them. This manipulation is intended to vary the extent to which the processes involved in generating stimulus representations occurs separately from or simultaneously with comparison, with these processes being more separated in participants given a pre-comparison processing step and more simultaneous in participants lacking that step.

Under the processing-demand hypothesis, any distinctions among representations subserving similarity and difference judgments should be diluted in participants who are prompted to generate representations of stimuli prior to making a comparison, relative to those not prompted to do so. Put differently, a pre-comparison processing step should yield fairly crystallized stimulus representations to then be compared, and these representations are expected to be more insulated from the constraints involved in actually comparing them. Specifically with respect to difference judgments, a reasoner is less constrained to represent stimuli non-relationally when they are processed prior to comparing them, relative to when they are first processed *during* comparison. This hypothesis thus predicts the asymmetry in similarity and difference judgments observed in Experiment 1 *only* for participants lacking a pre-processing step and who are therefore more likely to generate stimulus representations while also comparing them. In contrast, structure-mapping theory hypothesizes that asymmetries in similarity and difference judgments do not directly reflect the representations that comparison operates over, and it thus predicts that the manipulation described above will have no effect on response patterns.

In order for the proposed manipulation to produce the desired effect strongly enough to clearly test the predictions mentioned above, experimental stimuli must be complex enough such that generating a stable representation of them requires somewhat extensive processing. Using the highly simplified stimuli used in Experiment 1 would likely render any representational difference across presentation conditions (i.e., with pre-comparison processing versus without it) too subtle to detect from overt behavior. Instead, I use story stimuli originally used in Gentner et al. (1993) to dissociate the impact of relational similarity on analogical retrieval from that on analogical inference. Generating a stable representation of story content necessitates fairly extensive processing, and these stimuli are much more likely than those used in Experiment 1 to reflect any effect of the presentation manipulation discussed above. Moreover, results supporting the processing-demand hypothesis would extend the asymmetry demonstrated in Experiment 1 to more complex and naturalistic stimuli. These stimuli consist of story sets, each including one story that is analogous to a standard story and another story that is disanalogous but superficially similar to the standard, and so I use these sets to respectively emphasize relational and featural similarity in the two response options constituting the same triad task used in Experiment 1.

Method

Participants. Participants were 129 undergraduates (M_{age} = 20.61, SD_{age} = 3.03, range = [18, 37]) at UCLA. The sample consisted of 3 nonbinary, 107 female, and 17 male participants; 2 participants did not report their gender. All participants completed experimental tasks online to obtain partial course credit in a psychology class. The study was approved by the Institutional Review Board at UCLA.

Comparison task. Participants completed a story comparison task, in which they were asked to compare sets of three story stimuli drawn from Gentner et al. (1993). As in Experiment 1, participants were asked to compare a standard to a relational option and a featural option in order to select either which was more similar or which was more different. The relational option (labeled “analogy match” in the original materials) consisted of characters (e.g., a bird and a hunter versus a pair of nations) and individual events (e.g., gift giving of feathers versus gift giving of supercomputers) that were superficially dissimilar to those in the standard but that played roles in an overall plot structure that matched the standard (e.g., an act of kindness leads to a reciprocal act of kindness). In contrast, the featural option (labeled “mere-appearance match” in the original materials) consisted of characters and individual events that were superficially similar to those in the standard but that played roles in different plots structures from the standard (e.g., an act of kindness fails to elicit a reciprocal response versus the plot structure mentioned above). On each trial, the standard was always presented first at the top of the screen, and the two options were presented next on either side of the bottom of the screen. Which side the relational and featural option appeared was randomized across trials. Once presented, each story remained on the screen for the rest of the trial.

Table 1: Example set of story stimuli drawn from Gentner et al. (1993)

Story conditions	**Story examples**
Standard	Karla, an old hawk, lived at the top of a tall oak tree. One afternoon, she saw a hunter on the ground with a bow and some crude arrows that had no feathers. The hunter took aim and shot at the hawk but missed. Karla knew the hunter wanted her feathers so she glided down to the hunter and offered to give him a few. The hunter was so grateful that he pledged never to shoot at a hawk again. He went off and shot deer instead.
Relational	Once there was a small country called Zerdia that learned to make the world's smartest computer. One day Zerdia was attacked by its warlike neighbor, Gagrach. But the missiles were badly aimed and the attack failed. The Zerdian government realized that Gagrach wanted Zerdian computers so it offered to sell some of its computers to the country. The government of Gagrach was very pleased. It promised never to attack Zerdia again.
Featural	Once there was an eagle named Zerdia who donated a few of her tailfeathers to a sportsman so he would promise never to attack eagles. One day Zerdia was nesting high on a rocky cliff when she saw the sportsman coming with a crossbow. Zerdia flew down to meet the man, but he attacked and felled her with a single bolt. As she fluttered to the ground Zerdia realized that the bolt had her own tailfeathers on it.

On each trial, participants were presented with the standard at the top of the screen and were instructed to read the story carefully. They were given 10 seconds before they could proceed to see the two options. Participants were assigned to one of two presentation conditions, a comparison condition and a control condition, which differed in the way that the two options were presented. In the comparison condition, the two options were presented at the same time, directly after participants had proceeded from reading the standard. Once the two options were revealed, participants were instructed to compare and judge which option was more similar or more different from the standard.

In the control condition, each option was revealed one at a time: After being given at least 10 seconds to read the standard, participants were given at least 10 more seconds to read one option on the left side of the screen, before they could proceed to read the option of the right side of the screen for at least another 10 seconds. After having read all three stories, participants were finally asked to enter their responses as to which option was more similar to or different from the standard. The control condition gave participants an opportunity to process each option in isolation before comparing them to the standard. In contrast, the comparison condition required participants to process each option while comparing them to the standard. For both conditions, whether the relational or featural option appeared on the right or left side of the screen was randomized across trials. Crossing presentation condition (comparison versus control) with decision type (similarity versus difference) yielded four conditions, and both factors were manipulated between subjects.

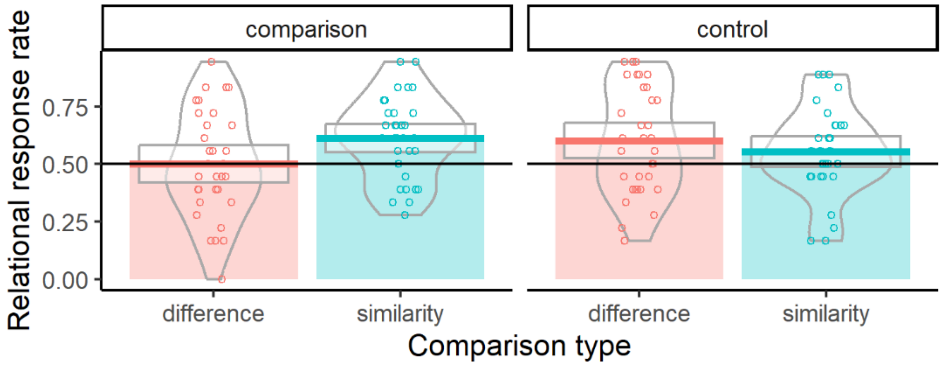


Figure 5: Relational response rate on story comparison task in comparison and control conditions. Response rates are broken down according to comparison type (difference vs. similarity). Unfilled circles each represent an individual participant's response rates, dark lines reflect mean response rates, box boundaries reflect ± standard error of the mean, and horizontal line corresponds to indiscriminate selection of relational versus featural options.

Results and Discussion

As in Experiment 1, I used the *glmer* function from version 1.1.26 of the LME4 R package (Bates et al., 2015) in R version 4.1.1 (R. Core Team, 2021) to fit a logistic mixed-effects model to performance on this comparison task. I defined a full model including *participant* and *comparison problem* as random intercept effects; with a *group* (*comparison* vs. *control*) x *comparison type* (*similarity* vs. *difference*) interaction term, as well as individual group and *comparison type* terms as fixed effects, as well as *trial number* as a covariate, which accounts for any systematic change in strategy across trials within a task.

I used a likelihood-ratio test to compare this full model to reduced models that omitted the *group* x *comparison task* interaction term but that was otherwise equivalent to the full model. As predicted by the processing-demand hypothesis, removing the interaction term increased model prediction error, $\Delta AIC = 2.80$, χ^2 (1) = 4.82, p = .03. Contrary to structure-mapping theory, this result confirms that whether or not participants processed story stimuli in a pre-comparison step had an impact on the response pattern among similarity and difference judgments. Therefore, responses on the triad task reflect differences in stimulus processing, and do not *merely* reflect differences in the usage of comparison output.

I used the *emmeans* and *pairs* functions from version 1.8.4 of the emmeans R package (Lenth, 2023) to compare the relevant estimated marginal means of the full model and test the difference compare relational responding in similarity and difference judgments for each group in relational. Similarity participants (*M* = .61, *SD* = .18) had higher rates of relational responding than difference participants (*M* = .51, *SE* = .23) when asked to simultaneously read and compare stories to the standard (*z* = 2.11, *p* = .035) in the comparison condition, but this difference did not hold for participants asked to read and then compare stories to the standard (similarity: *M* = .55, *SD* = .19; difference: *M* = .60, *SD* = .23; *z* = 1.02, *p* = .307) in the control condition. This result confirms the prediction of the processing-demand hypothesis that processing story stimuli *during* comparison elicited asymmetric responding. This difference was eliminated when participants were given an opportunity to read and process each story prior to comparing them.

General Discussion

Overall, the set of findings from Experiments 1 and 2 using a range of stimulus types (visual images, verbal words, and stories) provide convergent evidence for the claim that assessments of difference are more cognitively demanding than assessments of sameness (Hochmann, 2021; Hochmann et al., 2016, 2018), which in turn affects the way that human reasoners actually represent the items they compare. Because of the greater demand imposed by difference judgments, human reasoners represent the items about which they make this type of judgment in a more shallow or non-relational way, but the effects of this heightened demand can be mitigated by processing a given stimulus *before* comparing it. At any rate, this dissociation may ultimately be rooted in a representational asymmetry in the relations *same* and *different*, such that people process *different* as a negation of *same*.

I acknowledge that previous demonstrations of asymmetries between similarity and difference judgments that are not obviously explained by the representational asymmetry between *same* and *different* defended here (Simmons & Estes, 2008; Tversky, 1977). In addition to considering features and internal structural relations of stimuli (as studied here), human reasoners also tend to incorporate external relations or thematic relatedness *between* stimuli (i.e., association based on co-occurrence in some context; e.g., between *dog* and *leash*) in similarity judgments (Bassok & Medin, 1997), but they tend to do so less in difference judgments (Golonka & Estes, 2009; Simmons & Estes, 2008). Galonka and Estes (2009) argued this this asymmetry arises because thematic relatedness introduces commonalities between thematic associates without reducing the relevant differences between them. However, asking participants to complete a larger number of comparison trials (~60), and reminding participants of task instructions throughout the experimental session, has been shown to eliminate the effect of thematic relatedness on similarity judgments (Honke & Kurtz, 2019). Future work might clarify the impact of thematic relatedness on similarity and difference judgments, and whether any persisting asymmetry between the two might be explained in terms of the representational asymmetry discussed here.

Having shown in the present chapter that human reasoners tend to incorporate relational information when explicitly assessing similarity, I move on to examine the presence of relation processing in implicit comparisons made during recognition. Specifically, I examine whether humans use relations as cues when assessing the similarity between what is perceptually available and what is encoded in episodic memory.

Chapter 2: Relations and recognition

Introduction

If explicit relation representations impact human reasoning, then it may be possible to detect their influence in other cognitive tasks that do not directly involve reasoning. It has been reported that relation similarity can impact episodic memory in recognition tasks, giving rise to a phenomenon termed *relational luring* (Popov et al., 2017). In a typical experiment, participants were shown a sequence of word pairs to commit to memory, and at test were asked to indicate that a given word pair was ‘old’ if they had seen that exact word pair previously in the sequence, ‘recombined’ if it was a novel combination of individual words that they had seen before, or ‘new’ if they had not previously seen either the full word pair or its constituent words. Popov et al. showed that participants were more likely to misclassify ‘recombined’ word pairs as ‘old’, and took longer to correctly identify ‘recombined’ word pairs, when the pair instantiated a relation made familiar by previously presented pairs, as compared to word pairs that did not instantiate the same relation as a prior word pair. Moreover, the degree to which ‘recombined’ word pairs were misclassified, and correct responses were delayed, increased linearly with the number of instances of that relation a participant had seen previously. If a given relation is encoded explicitly as an item in memory, then relational luring is consistent with prior work showing that repeated presentations of a given item increase the likelihood of recognizing that item on a subsequent presentation (Challis & Sidhu, 1993; Reder et al., 2000).

Relational luring constitutes an example of false memory based on semantic similarity, extending massive evidence for semantic effects on false memory for individual words (e.g., Roediger & McDermott, 1995). However, relational luring has the distinctive property that it appears to arise from specific pairings of words, rather than the individual words in the pair. On the face of it, relational luring is naturally explained by assuming that an explicit representation of a semantic relation becomes increasingly familiar as it is activated by exposure to specific instances. The accrued familiarity of the relation then serves as a cue that tends to lead to false recognition of recombined word pairs instantiating the same relation. Thus, relational luring has been interpreted as providing evidence for the role of explicit relations in guiding recognition memory (Popov et al., 2017). However, this assumption has never been formally tested in a computational model of recognition memory, nor compared against alternative possibilities based on non-relational semantic analyses. The present paper fills this gap.

Word Embeddings as Predictors of Analogical Reasoning and Word Recognition

Advances in natural language processing (NLP) have generated representations of individual word meanings (e.g., Devlin et al., 2019; Mikolov, Chen, et al., 2013; Pennington et al., 2014), referred to as *word embeddings*. These representations are high-dimensional vectors that constitute hidden layers of activation within neural network models trained to predict patterns of text in sequence as they appear in large corpora. Word embeddings have been used to predict human judgments of lexical similarity and probability (for a review see Bhatia & Aka (2022); for a discussion of and response to critiques of embeddings as psychological models, see Günther et al. (2019)).

Crucially, word embeddings may capture rich aspects of conceptual meaning that go beyond surface features and direct category relations. For example, Utsumi (2020) was able to extract information from embeddings sufficient to predict the values of about 500 words on most of 65 semantic features (e.g., the extent to which something is *social*) for which neurobiological correlates have been identified. Such successes raise the possibility that relational luring might be explicable in terms of lexical overlap based solely on embeddings for word pairs, without necessarily involving explicit relation representations. In particular, embeddings might capture information about characteristic relational *roles* that concepts play (Goldwater et al., 2011; Jones & Love, 2007; Markman & Stilwell, 2001). For example, concatenated embeddings for the word pair *nurse:hospital* might include features that implicitly encode the facts that *nurse* is a human occupation and that *hospital* is a work location, perhaps creating a basis for relational luring.

In the present study I build on recent theoretical developments in which embeddings have been used to learn relation representations that can provide a basis for analogical reasoning. A number of alternative methods can be used to define similarity between word pairs. In the present study, I examine alternative methods that take the same embeddings as inputs, extracted using Word2vec (Mikolov et al., 2013). All these methods compute word-pair similarity based on cosine similarity (a measure well-suited for high-dimensional spaces). Critically, relation representations can either be based on explicit re-representations within a new relational space (i.e., a representational space in which the dimensions code abstract semantic relations such as *hypernym*, *antonym*, and *cause*; Ichien et al., 2022; Lu et al., 2012, 2022), or can be implicit in the raw word embeddings (Mikolov et al., 2013; Pennington et al., 2014).

Experiment 1

I first report an experiment designed to elicit relational luring. Rather than studying word pairs in the context solely of a memory task (Popov et al., 2017), participants were exposed to word pairs while making specific judgments about them (so that the encoding of these word pairs for a subsequent memory task was more incidental in nature). The first encoding task, involving relatedness judgments, required participants to decide whether the two words in a pair were related. Because relatedness judgments do not require identification of any specific relation, they can potentially be made using an implicit relation representation. The second encoding task, verbal analogical reasoning, required participants to decide whether or not an analogy in *A:B :: C:D* format was valid. Evaluating analogies requires attention to the similarity of the specific relations linking the *A:B* and the *C:D* word pairs, and hence is likely to depend on explicit relation representations (consistent with previous computational modeling; Lu et al., 2019). Each task was followed by a test of recognition memory, which included conditions designed to potentially elicit relational luring. By comparing memory performance following the relatedness and verbal analogy tasks, I sought to test whether relational luring depends on determining the particular semantic relations holding between word pairs (as evoked by the verbal analogy task), or whether a more generic assessment of whether a discernible relation exists between word pairs (as evoked by the relatedness task) is sufficient.

Critically, both the analogy task and the subsequent recognition memory task can be modeled using the same alternative measures of word-pair similarity. Specifically, I compare a measure of *relational* similarity between explicit relation representations with a measure of *lexical* similarity between individual word meanings. Based on previous findings, I predicted that the measure based on relational similarity would prove most effective for the analogy task. The key question is whether recognition memory will be best predicted by the same relational measure of word-pair similarity, or whether a dissociation will be observed between the analogical reasoning and recognition memory tasks. Procedures and analyses for all experiments were pre-registered on AsPredicted (#66576). All materials and analysis scripts are available on OSF (https://osf.io/vmn4z/?view_only=02dbe0d6beba4d2f8b0fd5002b693019).

Method

Participants. Participants were 111 undergraduates (M_{age} = 20.12, SD_{age} = 1.94) at either the University of California, Los Angeles (UCLA) (*n* = 93) or at Dartmouth College (*n* = 18).

Across the entire sample, participants were 81 female, 20 male, 1 nonbinary, and 9 gender not reported. All participants completed experimental tasks online to obtain partial course credit in a psychology class. The study was approved by the Institutional Review Boards at UCLA and at Dartmouth College. Participants were self-assessed proficient English speakers, and 82% were native English speakers. All analyses excluded data from 18 participants whose median correct response time, number of omitted responses, and/or *d'* were 2.5 standard deviations away from the sample mean on any task (final sample size: 93).

Procedure. All participants completed two blocks, each of which included three tasks. The first task in each block was an incidental encoding task involving either relatedness judgments (first block) or analogical reasoning (second block). The second task in each block was a demanding task involving visuospatial reasoning (a short form of Raven's Progressive Matrices); for my current purposes, this served as a distractor task. The third task in each block was a recognition memory task. The assignment of word pairs to each block was counterbalanced across participants. Participants were first shown a list of all the tasks they would be completing during the experimental session and thus made aware before starting the experiment that they would be completing memory tasks. Importantly, participants were not directly told that the relatedness and verbal analogy tasks were at all related to the memory tasks. The entire test session lasted approximately one hour. Figure 6 presents the sequence of tasks that each participant completed during an experimental session.

Prior to beginning the relatedness task, participants were shown examples of related and unrelated word pairs and then completed seven practice trials. Prior to beginning the verbal analogy task, participants were shown examples of valid (e.g., *carpenter:hammer* and *nurse:syringe*) and invalid analogies (*loop:ice* and *bowl:cereal*), and then completed four practice trials. Neither the individual words in the practice trials, nor the relations instantiated by them, overlapped with the word pairs used in the actual encoding tasks.

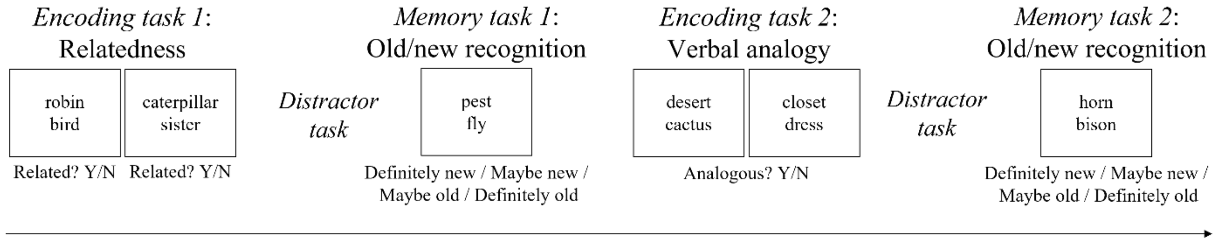


Figure 6: Task structure.
Participants completed six tasks, divided into two blocks (columns) of three tasks each. Task order was fixed. The two blocks of tasks were the same except for the encoding task, with assignment of specific word pairs counterbalanced across the two sets.

Materials and Encoding Tasks. In the relatedness task, participants were presented with a sequence of word pairs and asked to judge whether each pair was comprised of words that were semantically related (e.g., *footwear:boot*) or not (e.g., *mascara:spoon*). Word pairs were semantically related 90% of the time. In the verbal analogy task, participants were sequentially presented with two word pairs on each trial, and were asked to judge whether each set constituted a valid analogy (e.g., *fin:shark* and *wing:butterfly*) or not (e.g. *device:calculator* and *thorn:rose*). Valid analogies were shown on 54% of trials.

Both encoding tasks involved word pairs that instantiated one of three abstract semantic relations: *category:exemplar* (e.g., *bird:robin*), *part:whole* (e.g., *toe:foot*), and *place:thing* (e.g., *store:groceries*), or else were not semantically related (e.g., *mascara:spoon*). To create stimuli for these tasks, a total of 200 word pairs were constructed out of 400 unique words. Words were selected based on three sets of norms, for *concreteness*, *prevalence*, *frequency*, and also on *length*. Word *concreteness* is the extent that a given word refers to something that exists in reality and of which one can have immediate sensory (visual, auditory, gustatory, tactile, or olfactory) experience. I used concreteness norms presented by Brysbaert et al. (2014), which were collected as ratings on a 5-point scale from 1 (abstract) to 5 (concrete). Word pair stimuli were eliminated from this study if either of its two words had a mean concreteness rating lower than 4. Word *prevalence* is the proportion of people who know that word. I used prevalence ratings presented by Brysbaert et al. (2019), which consisted of *z*-scores such that words received negative prevalence ratings if fewer than 50% of people said they knew those words. Word-pair stimuli were eliminated from this study if either of the two words in a pair had a prevalence rating lower than 2.

The word pairs were evenly distributed across two 100 word-pair lists, one used for the relatedness task and the other used for the analogy task; which of the two lists was used for which encoding task was counterbalanced across participants. Within each list of 100 word pairs, 10 unrelated pairs consisted of words with no discernible semantic relation between them. The remaining 90 pairs were evenly distributed across the three abstract semantic relations (i.e., 30 word pairs per relation). Participants saw one list during the relatedness task and the other list during the verbal analogy task; which list was presented during each task was counterbalanced across participants. The analogy task appears to require explicit comparison of relations; hence this task was always placed in the second block (i.e., after the relatedness task), so as to avoid priming an explicit strategy of identifying abstract relations in the relatedness task (which potentially could be performed using a more implicit strategy of simply assessing the presence versus absence of any relation).

Each encoding task consisted of two blocks with a self-paced break between them. Each word pair within a given list was presented once during each block, and in each block word pairs were presented in a different order. Thus, each block of the relatedness task consisted of 100 trials (with one word pair shown per trial), yielding 200 trials in total. Each block of the verbal analogy task consisted of 50 trials (with two word pairs shown per trial), yielding 100 trials in total. In each encoding task, participants saw each word pair twice across the two blocks.

Memory Tasks. Following each encoding task and the intervening distractor task, participants completed an old/new recognition task in which they were presented with a sequence of 54 word pairs. Each word pair was constructed from individual words that participants had seen during their prior encoding task. Thus, each individual word was familiar to participants; however, they were recombined into new word pairs on 2/3 of the trials (i.e., 36 trials). Participants were asked to identify whether or not they had seen that exact combination of words in the previous encoding task, as well as to rate how confident they were in their judgment using a four-point scale: "Definitely New", "Maybe New", "Maybe Old", and "Definitely Old". The specific word pairs differed across the memory tasks in the two blocks. Participants were given a brief tutorial on the memory task prior to beginning each such task. None of the individual words or relations instantiated in this tutorial overlapped with those used in the actual task.

Table 2: Properties of each stimulus type used during recognition memory task.

Type of test word pairs	Previously studied individual words?	Previously studied word pairs?	Previously studied abstract relations?	Valid relation?
intact	✓	✓	✓	✓
familiar	✓		✓	✓
unfamiliar	✓			✓
unrelated	✓			

A total of 108 word pairs were used for the memory tasks, with each word pair drawn from one of four types (see Table 2). The first type, *intact*, consisted of “old” word pairs that were shown during the encoding task (relation identification or analogy). The other three types of word pairs were “new” pairs. All of these were constructed by recombining words that had appeared in the immediately prior encoding task, so that individual words were now paired differently, generating novel word pairs distinct from those used in the encoding task. More specifically, *relationally* *familiar* word pairs consisted of recombined word pairs instantiating the same relations as the word pairs presented during the encoding tasks (i.e., *part:whole*, *category:exemplar*, and *place:thing*). *Relationally unfamiliar* word pairs consisted of recombined word pairs instantiating a relation type (A *is similar to* B) to which participants had not been exposed in the encoding phase. These word pairs were formed using concepts with overlapping salient attributes (e.g., *bartender:cashier*), and hence were relationally similar to one another, but not with respect to any of the three relations included in the encoding tasks. Finally, *unrelated* word pairs consisted of recombined word pairs that were not semantically related in any discernible way (e.g., *cookbook:remote*). For intact pairs, responses of either “Maybe Old” or "Definitely Old" were scored as correct. The other three types of trials consisted of word pairs that were not used in either encoding task; either “Maybe New” or "Definitely New" were scored as correct responses. Among the 54 word pairs tested in the recognition memory task, 18 pairs were intact, 18 pairs were relationally familiar, 9 pairs were relationally unfamiliar, and 9 pairs were unrelated.

To generate “new” pairs by recombining words in the encoding tasks, another relevant factor (in addition to controlling relations instantiated by word pairs) that varied among the recognition stimuli was consistency of word position between the encoding tasks and the memory task (i.e., assignment of a given word to first versus second position in a pair for study versus test pairs). Popov et al. (2017) constructed their stimulus set using a large number of different relations with a few exemplar word pairs of each, enabling them to keep the position of any word in the test pairs the same as its position in the encoding tasks. In contrast, because my study used a small number of relations (three) in the encoding tasks and a large number of exemplar word pairs per relation (30 pairs per relation), it was impossible to maintain the same position for all words between the encoding tasks and the memory test.

In Experiment 1, for test pairs used in the memory task, the position of at least one word was preserved from its position in a study pair most often for intact pairs (100%), followed by familiar (95%), unfamiliar (84%), and unrelated pairs (66%). These differences in word positions across test pair types reflect the fact that word position naturally correlates with the role that a word plays in a relation (e.g., in a *category:exemplar* pair such as *food:spaghetti*, *food* fills the category role and *spaghetti* fills the exemplar role). For the three relations included in the encoding tasks (*part:whole*, *category:exemplar*, and *place:thing*), the terms occupying the first position in study pairs, and thus assigned to the first role in the corresponding relations (i.e., *part*, *category*, and *place* roles) often had to be assigned to the same role in familiar test pairs. Moreover, when words assigned to the second role of familiar test pairs (e.g., *butterfly* assigned to the exemplar role in the *category:exemplar* test pair *insect:butterfly*) was assigned to a different role from its study pair (e.g., *wing:butterfly*), that word usually still occupied the second position in its relation (e.g., the *whole* role in the *part:whole* relation). Thus, word position tended to be preserved for both words in familiar test pairs. On the other hand, in creating unfamiliar test word pairs that instantiated the *similar* relation, I were often forced to combine words that each filled the same role in different study pairs. For example, the unfamiliar test pair *tail:fin* was generated using words each assigned to the *part* role in *part:whole* study pairs (*tail:skunk* and *fin:shark*). This role-matching constraint tended to yield a position change of one word from study pairs to unfamiliar test pairs, whereas the position of the other word was usually consistent between study and test. In general, playing the same role within a relational structure tends to increase the similarity between distinct entities (Jones & Love, 2007).

Results

Encoding Tasks. Overall, participants performed well on both of the encoding tasks: relatedness task, M_{Acc}= .94, SD_{Acc}=.03; verbal analogy task, M_{Acc} = .76, SD_{Acc} = .12. Figure7 shows human accuracy in identifying valid and invalid analogy in the encoding task. Note that the false alarm rate for unrelated word pairs on the relatedness task was low (M_{FA} = .18, SD_{FA} = .16), yielding a high d-prime ($M_{D'}$ = 2.80; $SD_{D'}$ = .66). Thus, even though 90% of the trials involved semantically related word pairs, participants completed the task as instructed, and did not achieve their high accuracy by simply classifying all word pairs as related.

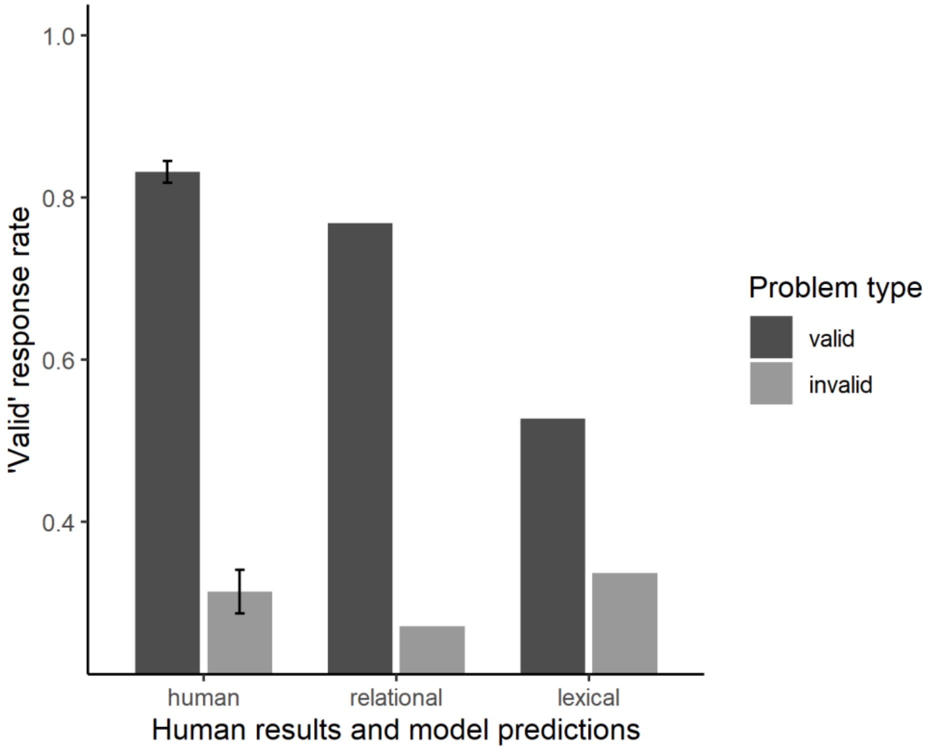


Figure 7: Human and model-predicted (i.e., relational and lexical) 'valid' responses on the verbal analogy task in Experiment 1.
Darker bars represent hits on valid analogies, and lighter bars represent false alarms on invalid analogies. Error bars reflect ± 1 standard error of the mean for human responses.

Recognition Memory. Participants showed good overall performance in recognizing studied word pairs following both encoding tasks (relatedness: M_{Acc} = .81, SD_{Acc} = .12; verbal analogy: M_{Acc} = .80, SD_{Acc} = .13). They correctly recognized old word pairs with responses of either “Maybe Old” or “Definitely Old”, exhibiting a high hit rate (relatedness: M_{Hit} = .90, SD_{Hit} = .12; verbal analogy: M_{Hit} = .86, SD_{Hit} = .14). However, they also sometimes misrecognized recombined word pairs (familiar, unfamiliar, or unrelated), exhibiting a substantial false alarm rate (relatedness: M_{FA} = .24, SD_{FA} = .16; verbal analogy: M_{FA} = .24, SD_{FA} = .17). Figure 8 shows that across encoding tasks, false alarms (i.e., mistakenly judging recombined new word pairs as studied old pairs) were more frequent for relationally familiar word pairs (relatedness: M_{FA} = .33, SD_{FA} = .19; verbal analogy: M_{FA} = .30, SD_{FA} = .19) than relationally unfamiliar word pairs (relatedness: M_{FA} =.21, SD_{FA} = .20; verbal analogy: M_{FA} = .22, SD_{FA} = .22), and for unfamiliar than unrelated pairs (relatedness: M_{FA} =.09, SD_{FA} = 16; verbal analogy: M_{FA} = .11, SD_{FA} = .19). The higher false alarm rate for familiar than unfamiliar pairs is consistent with the relational luring phenomenon reported by Popov et al. (2017).

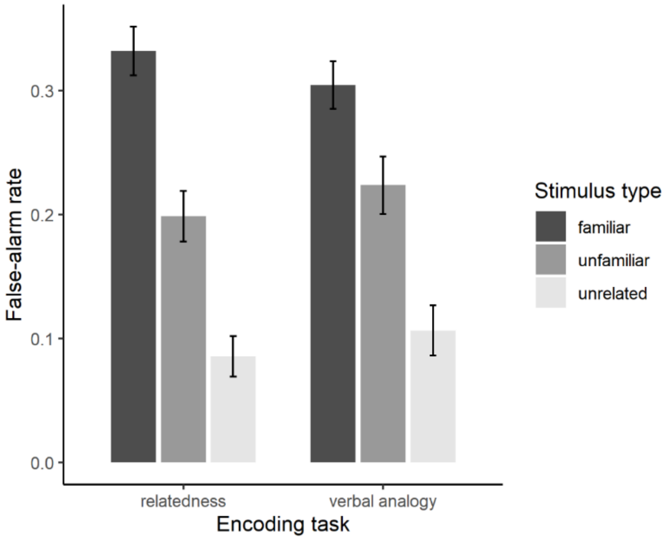


Figure 8: Human false-alarm rates on the recognition memory task in Experiment 1.
False alarm rates are broken down by relatedness and verbal analogy encoding task and by familiar, unfamiliar, and unrelated stimulus types. Error bars reflect ± 1 SEM.

To statistically test whether Experiment 1 replicated the relational luring effect, while controlling for other potential covariates, I analyzed false alarm data using logistic mixed-effects models. I used the *glmer* function from version 1.1.26 of the LME4 package (Bates et al., 2015), using R version 4.1.1 (R Core Team, 2021) to define logistic mixed-effects models of the data. Normed values on concreteness, prevalence, frequency, and length were treated as covariates. Since each of these metrics characterize individual words, I took the mean of a given metric for the two words constituting each word pair. For example, the word pair *food:salad* would have a concreteness of 4.89 because *food* has a concreteness rating of 4.80 and *salad* 4.97.

I defined a full model including *participant* and *word pair* as random effects and the following fixed effects: *stimulus type* (*familiar* vs. *unfamiliar* vs. *unrelated*) and *prior encoding task* (*relation detection* vs. *verbal analogy*), with the following covariates: *within-block trial number, concreteness*, *prevalence*, *frequency*, and *word pair length*. I first examined the effect of prior encoding task on false alarms by defining a reduced model that lacked the *prior encoding task* term but that was otherwise identical to the full model. Removing this term did not increase model prediction error, $\Delta AIC = 2.0, \chi^2(1) = 0.04, p = .85$. This finding reveals that participants did not differ reliably in their false alarm rates across the two encoding tasks (relation detection or verbal analogy). Contrary to my expectation, the relation detection task (which might be performed using more implicit processing of relations) was just as effective as the analogy task in producing false alarms on the recognition task.

Consistent with previous work (Popov et al., 2017), I hypothesized that participants would make false alarms more often to relationally familiar than relationally unfamiliar word pairs (i.e., showing a relational luring effect). In order to test this hypothesis, I fit a reduced model that removed the *stimulus type* fixed-effect term but that was otherwise identical to the full model, and then compared the prediction error between this reduced model and the full model. Indeed, I found that removing the stimulus type term from the full model increased prediction error, $\Delta AIC = 33.6$, $\chi^2(2) = 37.68, p < .001$. Inspecting the fit parameters of the full model, I also found that model predictions of false alarm rates for *familiar* word pairs were reliably higher than those for *unfamiliar* word pairs, $\beta = 0.86$, $z = 3.31$, $p < .001$, indicating that participants were more likely to false alarm on relationally familiar than relationally unfamiliar word pairs. I also found that predictions of false alarm rates for *unfamiliar* word pairs were reliably higher than those for *unrelated* word pairs, $\beta = 1.06$, $z = 3.31$, $p < .001$, indicating that the mere presence of a semantic relation induced participants to make false alarms more often. Moreover, the fact that this effect held across both prior encoding tasks indicates that detecting relations within the relatedness task was sufficient to elicit relational luring.

Experiment 1 thus yielded a higher false alarm rate for relationally familiar than unfamiliar pairs, consistent with the relational luring phenomenon reported by Popov et al. (2017). Their study maintained the same word positions between study and test pairs, whereas my study varied word positions between the encoding tasks and the memory test task. In this study, differences in false alarm rates between the familiar and unfamiliar types could potentially be due to the correlated differences in word position consistency. In a further analysis, I fit a linear mixed-effect model of false alarm data using the full model described above, but with the added covariate of the number of words in the same position from study to test for each word pair (0 vs. 1 vs. 2). I found that omitting both the stimulus type term, $\Delta AIC = 15.9$, $\chi^2(2) = 19.96, p < .001$, and the word position term, $\Delta AIC = 5.8$, $\chi^2(1) = 19.96$, $p = .005$, increased model prediction error. Inspecting fit parameters, I found that familiar word pairs did not reliably induce higher false alarm rates than unfamiliar word pairs after accounting for word position, $\beta = 0.52$, $z = 1.93$, $p =$

$.054$, but that unfamiliar word pairs still induced higher false alarm rates than unrelated word pairs, $\beta\ = 0.95, z = 3.09, p = .002$.

Experiment 2

Although Experiment 1 demonstrated relational luring, I was unable to rule out the possibility that the observed false alarm differences might be attributable to variations in consistency of word positions. Moreover, the previous experiment consistently used *category:exemplar*, *part:whole,* and *place:thing* as the familiar relations during the memory tasks and *similar* as the unfamiliar relation, and so I were unable to show that the observed luring effect generalized beyond this particular comparison of relations. In order to address these issues with Experiment 1, I carried out a follow-up experiment using materials adapted from Popov et al. (2017). These materials perfectly preserved word position for all stimuli between study and test phases, and they enabled us to counterbalance the particular relations that served as familiar and unfamiliar relations across participants.

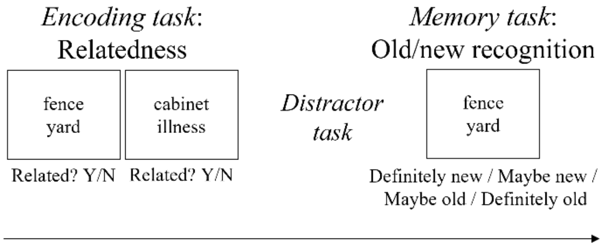


Figure 9: Task structure for Experiment 2.
Participants completed three tasks in a fixed order.

Method

Participants. Participants were 106 UCLA undergraduates (M_{age} = 20.92, SD_{age} = 4.24). Across the entire sample, participants included 92 female, 12 male, 1 nonbinary, and 1 gender not reported. All participants completed experimental tasks online to obtain partial course credit in a psychology class. The study was approved by the Institutional Review Boards at UCLA. All analyses excluded data from 8 participants whose median correct response time, number of omitted responses, and/or *d'* were 2.5 standard deviations away from the sample mean on any task (final sample size: 98).

Procedure. Because relatedness judgments and solving verbal analogies both proved sufficient to induce relational luring in Experiment 1, I employed relatedness judgments as the sole encoding task for Experiment 2. Thus, in contrast to Experiment 1, all participants in Experiment 2 completed a single block of three tasks: Relatedness judgments served as the encoding task, RPM problems served as the distractor task, and old/new recognition served as the memory task. As in Experiment 1, participants were first shown a list of all the tasks they would be completing during the experimental session (and thus made aware before starting the experiment that they would be completing a memory task but were not directly told that the relatedness task would be related to the memory task). The entire test session lasted approximately half an hour. Figure 9 presents the sequence of tasks that each participant completed during an experimental session.

Prior to beginning the relatedness task, participants were shown six examples of related and unrelated word pairs and then completed six practice trials. As with Experiment 1, neither the individual words in the practice trials, nor the relations instantiated by them, overlapped with the word pairs used in the actual encoding task.

Materials and Tasks. Word pair stimuli were adapted from English translations of Bulgarian stimuli used in Experiment 1 of Popov et al. (2017), and originally generated by participants in a study by Popov and Hristova (2015). All stimuli were based on a pool of 84 semantically-related word pairs. To create the present stimulus set, I edited Popov et al.'s translated stimuli in a few ways. I reversed word pairs that formed a common English bigram (e.g., *eye:sight* became *sight:eye*), replaced low-frequency words with more commonly-used associates (e.g., *schnitzel:calf* became *steak:cow*), replaced English words that were translated from multiple distinct Bulgarian words (e.g., *teacher:student* and *professor:student* became *teacher:student* and *parent:child*), and replaced words yielding an unclear semantic relation with more obvious relata (e.g., *soup:plate* became *soup:bowl*, which was then reversed to avoid a common bigram, ultimately yielding *bowl:soup*).

Table 3: Example of a stimulus set used in Experiment 2, adapted from Popov et al. (2017).

ID	Word pair	Relation
A	*atom:nucleus*	*object:center*
B	*planet:core*	
X	*bottle:cork*	*object:closure*
Y	*jar:lid*	

Each of the 84 word pairs had an analogous word pair (e.g., *atom:nucleus* and *planet:core*), and each of these 42 pairs of analogous word pairs was grouped with another pair of analogous word pairs (e.g., *atom:nucleus* and *planet:core* were matched with *bottle:cork* and *jar:lid*), yielding 21 stimulus sets (see Table 3 for an example). These stimulus sets were used to counterbalance across participants which stimuli were assigned to the encoding task and memory task. For a given participant, one word pair from each stimulus set was omitted (e.g., *jar:lid*), and individual words were swapped between two remaining disanalogous word pairs within each set (e.g., *planet:core* and *bottle:cork*), yielding two unrelated word pairs (e.g., *planet:cork* and *bottle:core*) for the encoding task. The final remaining word pair in that set was left intact, and served as a related word pair (e.g., *atom:nucleus*) for the encoding task. For that same participant, individual words in disanalogous word pairs were swapped back, yielding two “new” word pairs for the memory task (e.g., *planet:core* and *bottle:cork*), and the third word pair was again left intact and served as an “old” word pair (e.g., *atom:nucleus*) for the memory task. Of the two “new” word pairs generated from each stimulus set, one was analogous to the “old” word pair (e.g., *plant:core*) and thus served as a *relationally familiar* stimulus, while the other was not (e.g., *bottle:cork*) and thus served as a *relationally unfamiliar* stimulus. Table 4 shows the encoding-task and memory-task stimuli generated from a single stimulus set for two distinct participants.

Table 4: Stimuli in Experiment 2 generated from the set presented in Table 2, adapted from Popov et al. (2017)

Participant	Encoded pair	Encoded condition	Memory pair	Memory condition
1	*atom:nucleus*	*related*	*atom:nucleus*	*intact*
	planet:cork	*not related*	*planet:core*	*familiar*
	bottle:core	*not related*	*bottle:cork*	*unfamiliar*
2	*jar:lid*	*related*	*jar:lid*	*intact*
	atom:cork	*not related*	*bottle:cork*	*familiar*
	bottle:nucleus	*not related*	*atom:nucleus*	*unfamiliar*

This scheme yielded 8 distinct lists of 63 word pairs for each of the encoding and memory tasks; which participants saw which lists was randomized. For the encoding task, 21 word pairs were semantically related (33%), and the remaining 42 were not semantically related (66%) (yielding a better balance between related and unrelated words than did the relatedness task in Experiment 1). For the memory task, 21 word pairs were *intact* (“old” word pairs seen during the relatedness task) and the remaining 42 were “new”, of which 21 were *relationally familiar* and the other 21 were *relationally unfamiliar*. Trial order for the encoding tasks was also counterbalanced such that participants assigned to a given trial list were presented either with one randomized sequence of word pairs or its reverse.

Trial order for the memory tasks was more constrained. Note that each ‘new’ but relationally familiar word pair (e.g., *planet:core*) had an analogous ‘old’ counterpart (e.g., *atom:nucleus*). In contrast to Experiment 1, each ‘old’ word pair exemplified a unique semantic relation (e.g., *object:center*) during the encoding task. Accordingly, between the old word pair and its relationally familiar counterpart, whichever appeared first during the memory task constituted participants’ first exposure to that semantic relation during the task. Popov et al. (2017) found that correct response times were reliably higher (and false alarms were numerically higher) for relationally familiar word pairs than relationally unfamiliar word pairs only when relationally familiar word pairs served as the first instance of their semantic relation during the memory task—that is, when they appeared *before* their 'old' analogs but not when they appeared *after*. (I replicated this finding in a pilot study.) It seems likely that participants would notice (at least implicitly) that a given relation was "used up" once it had occurred once, and hence would avoid making false alarms to further instantiations of the same relation. In order to avoid this complication due to stimulus ordering, I generated a single trial order for each memory task list with the constraint that relationally familiar and the relationally unfamiliar word pairs drawn from the same stimulus set both appeared before their corresponding 'old' word pair. I counterbalanced whether the relationally familiar word pair appeared before or after its corresponding relationally unfamiliar word pair within each list. Otherwise, the trial order for each list was randomized.

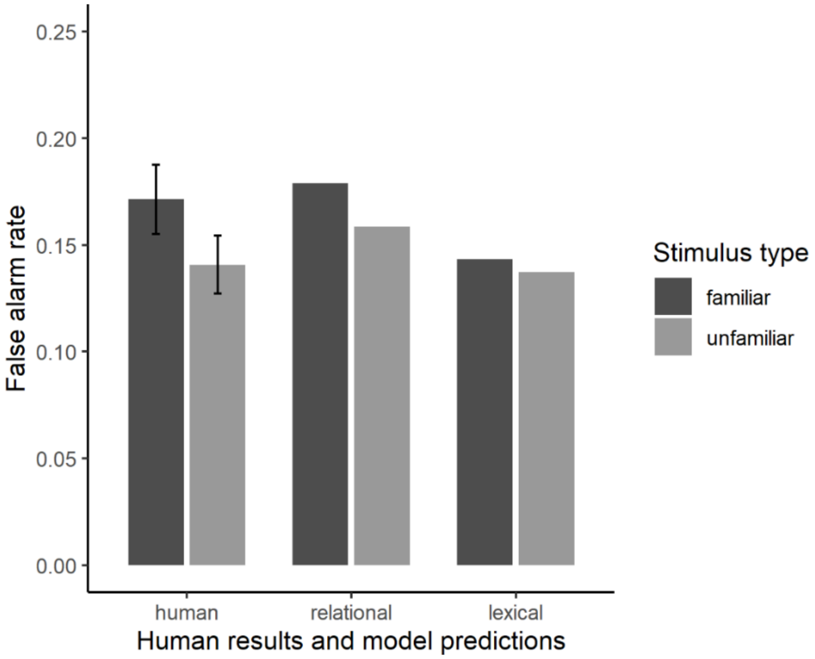


Figure 10: Human false-alarm rates and model predictions on the recognition memory task in Experiment 2
False alarms rates are broken down according to stimulus type (relationally familiar and relationally unfamiliar word pairs). Error bars reflect ±1 SEM.

Results

Encoding Task. Overall, participants performed well on the encoding task: M_{Acc}= .90, SD_{Acc}= .06, with a low rate of false positive judgments (M_{FA} = .09, SD_{FA} = .06).

Memory Task. Overall, participants performed well on the memory task: M_{Acc}= .76, SD_{Acc}= .113, with a moderately high false-alarm rate (M_{FA} = .16, SD_{FA} = .14). I also found that false alarms were more frequent for relationally familiar word pairs (M_{FA} = .17, SD_{FA} = .16) than relationally unfamiliar word pairs (M_{FA} =.14, SD_{FA} = .13). As in Experiment 1, I fit logistic mixed-effects models to the human false-alarm data. I defined a full model including *participant* and *word pair* as random effects, *stimulus type* (*familiar* vs. *unfamiliar*) as a fixed effect, with the following covariates: *trial number, concreteness, prevalence*, *frequency*, and *word pair length*. I found that omitting the stimulus type term reliably increased model prediction error, $\Delta AIC = 8$, $\chi^2(1) = 9.97$, $p = .002$, indicating that participants were more likely to false alarm on relationally familiar than relationally unfamiliar word pairs (see Figure 10). Hence, despite various methodological differences between the present study and the experiment reported by Popov et al. (2017), I obtained the same basic finding: higher false alarm rate for familiar than unfamiliar pairs. Experiment 2 also demonstrated relational luring using materials in which word position was held constant across study and test stimuli. Notably, the magnitude of this luring effect (.03) is smaller than that demonstrated in Experiment 1 (.11). While there are a number of important differences between the two experiments (e.g., the number of relations, the number of word pair examples of each relation, the particular relations used for each condition), I suspect that the word position confound that is present in Experiment 1 but controlled for in Experiment 2 is primarily responsible for the difference in effect magnitude.

Tests of Computational Models

Measures of Word-Pair Similarity

To predict performance on both the analogy task and the recognition memory task, I compared two measures of similarity between word pairs: (1) *relational*: similarity of word pairs based on the similarity of the explicit relation between the two words in each individual pair; (2) *lexical*: similarity of word pairs computed directly from the similarities of the individual words in each pair. I implemented specific versions of both possibilities, all rooted in 300-dimensional word embeddings created by Word2vec.

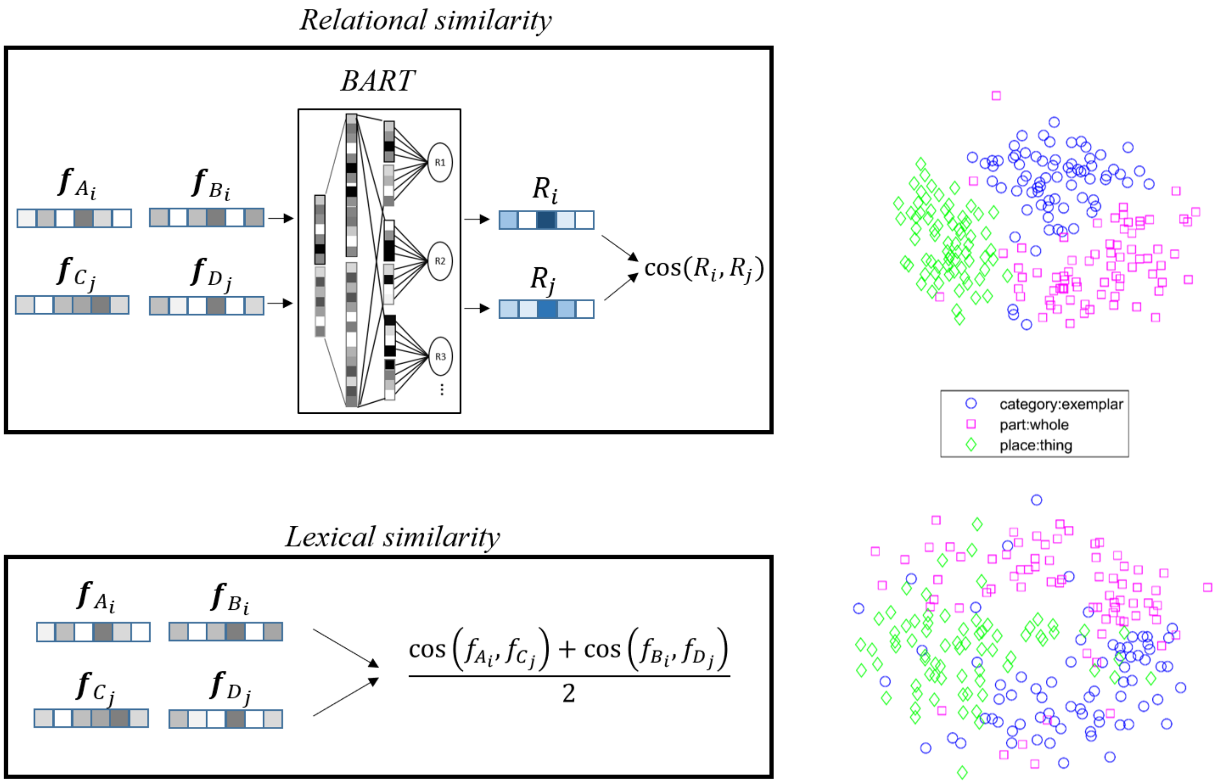


Figure 11: An illustration of relation similarity model (left top panel) and lexical similarity model (left bottom panel), and the resulting 2-D plot of similarity space derived using each (right panel).
The scatter plots of similarity spaces are derived from 216 word-pair stimuli instantiating *category:exemplar* (blue circles), *part:whole* (magenta squares), and *place:thing* (green diamonds) relations. Plotted stimuli on the right consist of related word pairs used for encoding tasks (180 total) and relationally familiar recombinations used for memory tasks (36 total).

As shown in Figure 6 top panel, to compute relational similarity I used relation vectors generated by *Bayesian Analogy with Relational Transformations* (BART; Lu et al., 2012, 2019). BART assumes that specific semantic relations between words are coded as distributed representations over a set of abstract relations. The BART model takes concatenated pairs of Word2vec vectors as input, and then uses supervised learning with both positive and negative examples to acquire representations of individual semantic relations.

After learning, the BART-based relational model calculates a relation vector consisting of the posterior probability that a word pair instantiates each of the learned relations (for details of the training procedure, see Ichien et al., 2022), as shown in Figure 11 left top panel. The relational model uses its pool of 270 learned relations to create a distributed representation of the specific relation between any two paired words *A:B* and *C:D*. The posterior probabilities calculated for all learned relations form a 270-dimensional relation vector R_i for the *A:B* word pair and relation vector R_j for the *C:D* word pair, where each dimension codes how likely a word pair instantiates a particular learned relation. The distance between word pairs i and j is computed as the cosine distance between corresponding relation vectors R_i and R_j :

$$d_{Rel_{ij}} = \cos\,(R_{i,}, R_j). \qquad (4)$$

As shown in Figure 11 left bottom panel, to compute lexical similarity the meaning of a word pair is represented by the two individual semantic vectors respectively representing each word. I use f_A , f_B to denote the semantic vector for the two words in a word pair *A:B*, and f_C, f_D to denote the semantic vector for the words in pair *C:D*. I compute the distance between word pairs i and j as the mean of the distances between f_{A_i} and f_{C_j} and between f_{B_i} and f_{D_j}:

$$d_{Lex_{ij}} = \frac{cos\left(f_{A_i}, f_{C_j}\right) + cos\left(f_{B_i}, f_{D_j}\right)}{2}. \qquad (5)$$

This representation is nonrelational, coding word pairs solely in terms of the meanings of the individual words (as determined by their Word2vec embeddings).

To provide a preliminary sense of how well the two basic measures of word-pair similarity (relational and lexical) capture the categorical distinctions among the three relation types used in the encoding tasks for Experiment 1 (*category:exemplar, part:whole,* and *place:thing*), Figure 11 in the right panels plots 216 word pairs (180 related word pairs used for the encoding tasks and 36 relationally familiar recombinations used for the memory tasks) on a 2-dimensional projection of the similarity space derived using each of the two measures. From visual inspection, it is clear that the relational measure (top) separates the three types of pairs into clusters corresponding to semantic categories more clearly than does the lexical measure (bottom); however, the lexical measure also predicts relation type to some extent, as the three clusters are somewhat separated (despite overlaps across relation categories).

Modeling Verbal Analogical Reasoning

Performance on the verbal analogy task in Experiment 1 was modeled directly by the BART-based relational model, which in addition to learning relations (as described above), can also be used to predict behavioral (Lu et al., 2019) and neural (Chiang et al., 2021) responses to analogy problems. In order to predict yes/no decisions about analogy problems, I computed cosine distances between representations of the *A:B* and *C:D* word pairs, and then searched for a decision threshold that generate the best model performance, such that word pairs with distances below the threshold indicate a valid analogy and those above indicate an invalid analogy. In calculating distance for the purpose of solving analogy problems, I used relational and lexical similarity metrics. Based on prior modeling of verbal analogical reasoning (Lu et al., 2019; Chiang et al.,

2021) and of explicit judgments of relation similarity (Ichien et al., 2022), I predicted that the model based on relational similarity would best predict human judgments on the explicit analogy task.

Figure 7 (see above) presents the proportion of 'valid’ responses for models as well as humans, broken down by valid analogies (darker bars) and invalid analogies (lighter bars). Overall, the BART-based relational model achieved higher accuracy (.75), nearly matching human proportion correct (.76). The alternative model based on lexical (non-relational) similarity performed poorly (.59 correct).

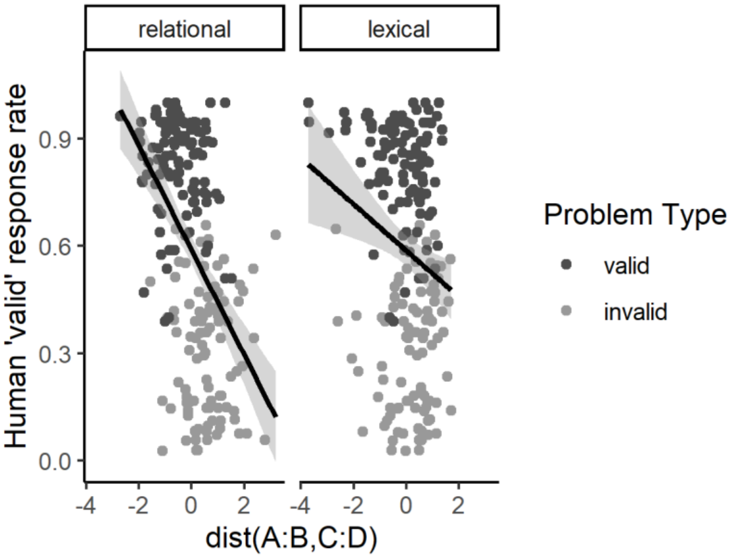


Figure 12: Human item-level ‘valid’ response rates on verbal analogy problems in Experiment 1, plotted against z-scored distance (dissimilarity) metrics predicted by the relational model (left) and by the lexical model (right). Each point represents a single analogy problem, and point shade reflects whether a problem features a valid analogy (dark grey) or an invalid analogy (light grey). The scatter plots were overlaid with a fitted regression line.

An item-level analysis corroborated these results. I used the *cocor* package in R to test the difference between the extent that each similarity measure correlated with the frequency with which human reasoners judged each analogy as valid (Diedenhofen & Musch, 2015). A Dunn and Clark's (1969) *z*-test showed that relational similarity was more highly correlated with human responses (r = .47) than was lexical similarity (r = .21; z = 3.68, p < .001). Figure 12 presents scatter plots of human item-level responses and z-scored model predictions based on each dissimilarity metric. Because this item-level analysis is based purely on dissimilarity predictions generated using each model, its results are independent of the decision threshold that was fit to maximize model accuracy in the analogy task. These simulation results thus confirm previous findings showing that the relational model based on explicit representations of semantic relations outperforms the alternative model based on lexical similarity in tasks involving verbal analogy, as well as explicit judgments of relation similarity (Chiang et al., 2021; Ichien et al., 2022; Lu et al., 2019).

Modeling Recognition Memory

To provide a formal account of relational luring, I adapted an established model of recognition memory, the Generalized Context Model (GCM; Nosofsky, 1988, 1991; Nosofsky & Zaki, 2003). GCM predicts old/new recognition judgments, and is closely related to several other successful models of episodic memory and categorization (e.g., Anderson, 1991; Kruschke, 1992; Love et al., 2004). If a version of GCM is able to account for relation-based false alarms, I will have demonstrated that this phenomenon is one of many that can be explained within a unified theoretical framework for exemplar-based recognition and categorization.

In the version of GCM implemented here, I assume that recognition of a given word pair on a memory task is based on a comparison of similarities between that word pair and all word pairs presented during the prior encoding task (as described below). The probability with which a participant will classify a word pair i as one they had seen during the encoding task is given by

$$P(old|i) = \frac{F_i}{F_i + k}, \qquad (6)$$

where k is a parameter representing a criterion for recognition, and F_i is the familiarity of word pair i, defined as:

$$F_i = \sum_{j \in J} s_{ij}. \qquad (7)$$

Here, J is the set of word pairs shown during the encoding task, and s_{ij} is the similarity between word pair i in the memory task and each word pair j from the encoding task. This similarity follows an exponential decay function (Shepard, 1987) of the psychological distance d_{ij} between word pairs i and j,

$$s_{ij} = e^{-cd_{ij}}, \qquad (8)$$

where c is a scaling parameter representing the rate of decline in similarity with psychological distance between word pairs. When GCM is fit to data from individual participants, c is typically interpreted as a measure of a participant's memory sensitivity (i.e., the extent to which they can discriminate between word pairs in memory, with higher values of c indicating greater sensitivity; Nosofsky, 1988). This interpretation of c is appropriate when comparing among parameter values within a *fixed* representational space. In contrast, the present simulations fit the model to group-level data, varying the representations for word pairs over which the model operates (details below). Therefore in my simulations, c (because it varies across different types of representations) is naturally interpreted as the discriminability between word-pair items within a given representational space. Because representational spaces can vary according to arbitrary scaling properties, I scaled all model-generated distance values between 0 and 1. As these representations are high-dimensional, I adopt cosine distance to compute d_{ij}, rather than the

Minkowski power formula typically used in previous work (e.g., Nosofsky, 1988, 1991; Nosofsky & Zaki, 2003).

As the above equations make clear, GCM must be grounded on some measure of similarity between word pairs. I compared the two measures described above (relational and lexical) within the basic GCM framework.

Table 5: GCM parameters fit to human data and fit-model performance for relational similarity (rel) and lexical similarity (lex)

	c		***k***		**log-likelihood**		**RMSD**		**spearman**	
	rel	**lex**	**rel**	**lex**	**rel**	**lex**	**rel**	**lex**	**rel**	**lex**
Exp. 1	15.5	10.0	.20	.20	-5013	-5063	.163	.169	.794	.764
Exp. 2	15.5	10.5	.30	.20	-2836	-2768	.159	.149	.658	.665
Popov et al. Exp. 1	11.5	8.0	.40	.40	-1288	-1271	.199	.192	.669	.644

Simulation results for Experiment 1. First, I modeled human recognition memory performance for Experiment 1. Because I found no reliable differences in either false alarm rates or overall accuracy across the two encoding tasks, I simulated the data obtained by averaging responses across them. For this simulation, model predictions were $P(old|i)$ for each word pair item; human judgments were the response proportions with which human participants judged a word pair item to be either "Maybe old" or "Definitely old". I first ran the GCM using each of the two variants of similarity (relational vs. lexical) to fit item-level human data for all 54 word pairs tested in the recognition memory task. I used a binomial distribution as the likelihood function to fit the scaling parameter c and criterion parameter k that maximized the log-likelihood. Table 5 summarizes fit model parameters, maximum log-likelihood, and RMSD and spearman correlations between fit model predictions and item-level human data. Figure 8 presents false-alarm rates for model-generated $P(old|i)$ predictions using the fitted parameters, as well as human data, broken down by type of recombined word pairs. Crucially, using either of the alternative similarity calculations, GCM predicts the relational luring effect observed in the human data: higher false alarm rates for relationally familiar than for relationally unfamiliar word pairs. While Figure 13 only shows false alarm rates to clearly highlight that human and model-predicted luring effects, both models also clearly discriminate between intact word pairs and recombined lures, predicting much higher hit rates for intact word pairs than false alarm rates for recombined lures, as observed in the human data (human: M_{Hit} = .88, SD_{Hit} = .10, M_{FA} = 24, SD_{FA} = .15; relational: M_{Hit} = .79, M_{FA} = .26; lexical: M_{Hit} = .79, M_{FA} = .28).

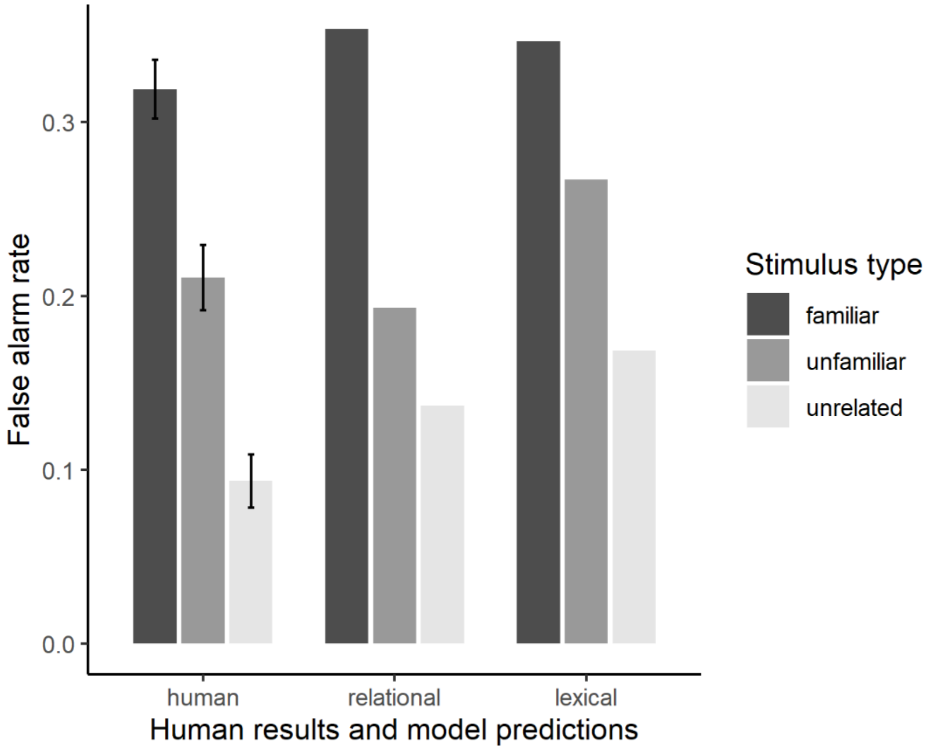


Figure 13: Human false-alarm rates and model predictions on the recognition memory task in Experiment 1, broken down according to familiar, unfamiliar and unrelated stimulus types.
Error bars reflect ±1 SEM.

Next, I assessed the robustness of the relational and lexical models to variations in the two model parameters: GCM's scaling parameter c and its criterion parameter k. Specifically, I examined the space of parameters and item-level deviation between model predictions and human responses using all 54 test word pairs. To provide a quantitative comparison of the model's robustness to predicting human data with each similarity metric, I computed the log model evidence (Friel & Wyse, 2012; Hoeting et al., 1999) by averaging the log likelihood that each model predicts the proportion of human participants who judged each word pair as old, over a range of the model parameter space (c = [0,50] with a stepsize of 0.5; k = [.1,1] with a stepsize of 0.1). I selected this range of parameters to capture both the maximum log-likelihood model predictions of overall human data, as well as the maximum model-predicted luring effect for the current simulation, as well as simulations of Experiment 2 and Popov et al. Experiment 1 discussed below.

The computation of log model evidence assumes a uniform prior for parameters. The log model evidence calculation uses the same binomial likelihood function that I used for model fitting. As shown in Table 6, I found that the log model evidence for the relational similarity metric was E_{log} = -1.324 x 10^4, substantially greater than that for lexical similarity, E_{log} = -1.569 x 10^4. This analysis provides converging evidence that the relational model provides a more robust account of the human data than does the lexical model.

Table 6: Log and luring-specific model evidence for GCM using relational similarity (rel) and lexical similarity (lex) averaged over a wide range of the model parameter space (c = [0,50], k = [.1,1])

	E_{log}		E_{luring}	
	rel	**lex**	**rel**	**lex**
Exp. 1	-1.324 x 10^4	-1.569 x 10^4	2.533	2.355
Exp. 2	-6.738 x 10^3	-8.131 x 10^3	3.310	3.291
Popov et al. Exp. 1	-2.961 x 10^3	-3.614 x 10^3	3.643	3.631

I also examined the range of parameters in models that generate the effect of relational luring. In this analysis I focused on model judgments for two types of test pairs, relationally familiar and relationally unfamiliar word pairs. I identified the parameter combinations for which each model (relational or lexical) predicts more false alarms for relationally familiar than relationally unfamiliar word pairs. The results of this analysis are depicted in Figure 13, where reddish cells indicate paired values of c and k with which models predict a false-alarm difference (i.e., mean $P(old|i)$ for relationally familiar word pairs is greater than mean $P(old|i)$ for relationally unfamiliar word pairs). Examination of the parameter range displayed in Figure 14 clearly reveals that within the GCM framework, relational similarity is a more robust predictor of the relational luring effect than is lexical similarity. That is, relational similarity yields the predicted difference (i.e., luring effect) across a larger set of parameter values than does lexical similarity (hence there are many more dark cells in the left panel than in the right panel).

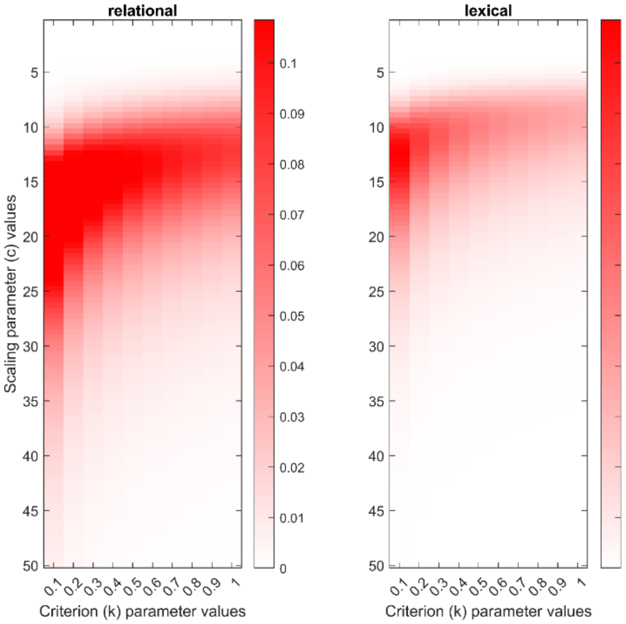


Figure 14: Simulation of model-predicted relational luring effect in Experiment 1 as a function of model parameters. Each cell represents a combination of values for GCM's scaling parameter c (*y*-axis) and its criterion parameter k (*x*-axis), respectively. Given the pair of parameter values for each cell, cell color represents the model-predicted difference of false alarm rates between familiar word pairs and unfamiliar word pairs (i.e., relational luring effect). Redder cells indicate a greater magnitude of model-predicted luring effect. The highest intensity of red corresponds to the magnitude of the luring effect observed in human data.

To provide a quantitative comparison of the robustness with which model predicts relational luring using each similarity metric, I computed the luring-specific model evidence as the marginal likelihood that each model predicts the mean luring effect (i.e., greater false alarms to familiar than unfamiliar test items) observed in human data, averaged across the same range of the parameter space that I used to compute log model evidence (c = [0,50] with a stepsize of 0.5; k = [.1,1] with a stepsize of 0.1). The luring-specific model evidence computation assumes a uniform prior for parameters. For each combination of parameters, likelihood of observing mean human luring effect was calculated using a Gaussian distribution centered at the model-predicted luring effect with the standard deviation SD_{luring} = .1240, which was observed derivation of luring effect among human participants. Model evidence was computed as the marginal likelihood by averaging the likelihood probabilities across the parameter space. As shown in Table 6, I found that the luring-specific model evidence for the relational similarity metric (E_{luring} = 2.533) was greater than that for lexical similarity (E_{luring} = 2.354). The greater robustness for the relational model in predicting the luring effect is consistent with the finding that relational similarity yields clearer separation of word pairs based on the three semantic relations than does lexical similarity (see Figure 11, right panels).

Even though the relational model was able to generate the luring effect more robustly than the lexical model, it is somewhat surprising that the lexical model was able to generate the relational luring effect at all. Since the lexical model only has access to similarities among individual word meanings, how was it able to reproduce this putatively relational effect? The intuitive explanation is that some lexical properties are shared by words that serve the same semantic role in word pairs instantiating a relation. For examples, the *category* words in *category:exemplar* relations (e.g., *reptile*, *food*, or *clothing*) tend to be superordinate categories and abstract words, the *part* words in a *part:whole* relations (e.g., *fang*, *wall*, *lobe*) tend to be objects that do not commonly exist on their own but as parts of a larger structure, and the *place* words in *place:thing* relations (e.g., *pond*, *bakery*, *chapel*) are necessarily locations.

Figure 15 shows a multidimensional scaling result derived from lexical similarity between *individual* Word2vec embeddings for the first words in related word pairs used in the memory task from Experiment 1. This plot illustrates that words filling the first roles in *category:exemplar*, *part:whole,* and *place:thing* relations tend to form discernible clusters, reflecting their tendency to have constraining lexical features. Thus, the lexical model's ability to capture the relational luring effect (shown in the bottom-right panel of Figure 11) is largely based on high similarity among first words in relationally familiar and intact word pairs. The second words in the pairs did not form clusters corresponding to the three relations.

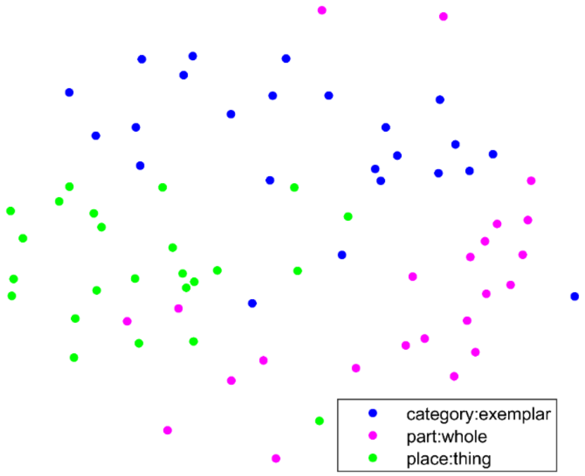


Figure 15: Multidimensional scaling based on for lexical similarity among individual first words in pairs used in the memory task for Experiment 1. Colors indicate word-pair relations (category:exemplar, part:whole, and place:thing).

Simulation results for Experiment 2. Using the same model-fitting procedure as for Experiment 1, I optimized GCM parameters with the maximum log-likelihood fit to the item-level human data for each similarity metric, using a binomial likelihood function. Figure 10 (above) presents false-alarm rates for model-generated $P(old|i)$ predictions using the fitted parameters, as well as human data, for familiar and unfamiliar word pairs, and the figure shows that, as in Experiment 1, both relational and lexical similarity predict a higher false alarm rate for familiar than unfamiliar word pairs. Moreover, both predict much higher hit rates for intact word pairs than false alarm rates for recombined lures, in line with the human data (human: M_{Hit} = .82, SD_{Hit} = .22, M_{FA} = .16, SD_{FA} = .14; relational: M_{Hit} = .78, M_{FA} = .17; lexical: M_{Hit} = .84, M_{FA} = .17).

Experiment 2 used more tightly controlled stimuli than Experiment 1, holding constant word position across study and test pairs and counterbalancing which relations contributed to relational familiarity during the memory task across participants. Likely as a result, the difference in the human false-alarm rates between relationally familiar and unfamiliar word pairs was much smaller in Experiment 2 than in Experiment 1, and both models were able to capture this because both lexical and relational similarity are sensitive to word position: Lexical similarity between word pairs is based on similarity computed between words in the same position only, and the relation representation entering into relational similarity is sensitive to word position such that the relation representation for *dog:animal* is different from that for *animal:dog*, and the former is more similar to *car:vehicle* than is the latter.

Although both models predicted the luring effect in Experiment 2, as well as a smaller effect in Experiment 2 than in Experiment 1, the luring effect generated within the relational similarity metric was much more similar in magnitude to the human effect than that generated within the lexical similarity metric. Moreover, as shown in Figure 11, this was the case across a wide range of parameters: the relational metric robustly produced a human-like luring effect, as shown by the strip of red cells in the left panel, while the lexical metric failed to produce luring effects of comparable magnitude at all, as shown by the lack of any bright red cells in the right panel. Importantly, because Experiment 2 eliminated the word-position confound in Experiment 1, the increased false alarm rate to relationally familiar word pairs compared to relationally unfamiliar word pairs in Experiment 2 more unambiguously reflects *relational* luring than does the comparable data from Experiment 1. Thus, the relational model's unique success in reproducing a luring effect of similar magnitude to humans in Experiment 2 provides particularly strong evidence for the importance of relation representations in recognition memory.

In order to quantitatively examine differences between the two models, I used the same analysis of log model evidence as in Experiment 1 to account for human data from all 63 test word pairs in Experiment 2. As shown in Table 6, I found greater model evidence for the relational model (E_{log} = -6.738 x 10^3) than for the lexical model (E_{log} = -8.131 x 10^3). As for Experiment 1, I went also computed influence of parameter variations on model-predicted relational luring effect. Even more than was the case for Experiment 1, the relational similarity metric predicted relational luring across a greater range of parameter variations than did the lexical metric. Using the same analysis for luring-specific model evidence as in Experiment 1, as Table 6 shows, I found that the model evidence for the luring effect observed in the human data (M_{luring} = .0306, SD_{luring} = .1174) was greater for the relational model (E_{luring} = 3.310) than for the lexical model (E_{luring} = 3.291). I acknowledge that while the luring-specific model evidence for the relational model is greater than that for the lexical model, the magnitude of this difference is much smaller than that observed in Experiment 1. Still, given the large difference in log model evidence between the two models, I maintain that relational similarity more robustly accounts for human data across a wide range of parameter values.

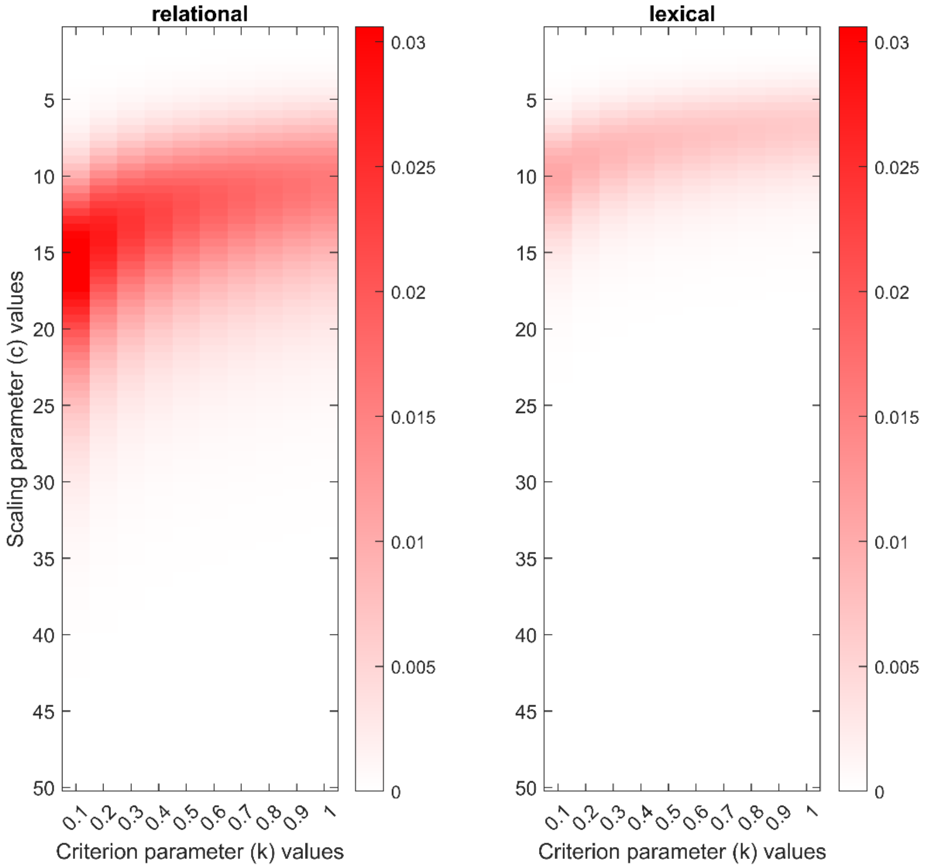


Figure 16: Simulation of model-predicted relational luring effect in Experiment 2 as a function of model parameters Red color of cells indicates magnitude of model-predicted luring effect. The highest intensity of red corresponds to the magnitude of the luring effect in human data.

Given that the materials used in Experiment 2 involved more relation types and were more well-controlled than those used in Experiment 1, it may seem even more puzzling that the lexical model could reproduce the luring effect at all. In order to clarify this issue, I compared relational and lexical similarity between word pair items within this dataset. Recall that all test pairs within each participant's 63-item stimulus list belonged to one of 21 stimulus sets. For each set there was a triplet consisting of an *intact* "old" word pair that was shown during the encoding task (e.g., *atom:nucleus*), and two "new" word pairs not shown during the encoding task. One was a *relationally familiar* word pair that was analogous to the intact word pair (e.g., *planet:core*) and the other was a *relationally unfamiliar* word pair that was disanalogous to the intact word pair (e.g., *bottle:cork*). (See Memory Pair column of Table 4 for two examples of *intact*, *relationally familiar*, and *relationally unfamiliar* triplets generated from the same stimulus set.) I computed the relational and lexical distances between each relationally familiar and each relationally unfamiliar word pair and its corresponding intact word pair. Figure 16 shows the average cosine distances across all such unique triplets used in Experiment 2. While it would be expected that the relational distance between familiar and intact word pairs should be much smaller than that between unfamiliar and intact word pairs, it is striking that lexical distances yield the same pattern.

The explanation for the lexical model's ability to predict relational luring in Experiment 2 is broadly consistent with the explanation for Experiment 1. Words serving the same role in analogous word pairs (e.g., *atom* and *planet*; *nucleus* and *core*) are more similar to each other in Word2vec space than words in disanalogous word pairs (e.g., *atom* and *bottle*; *nucleus* and *cork*). Indeed, this analysis shows that lexical similarity and relational similarity overlap more than might be expected, and that this overlap enabled the lexical model to reproduce the seemingly relational phenomenon of relational luring. These findings thus confirm that embeddings produced by Word2vec capture important aspects of word meaning related to typical relational roles.

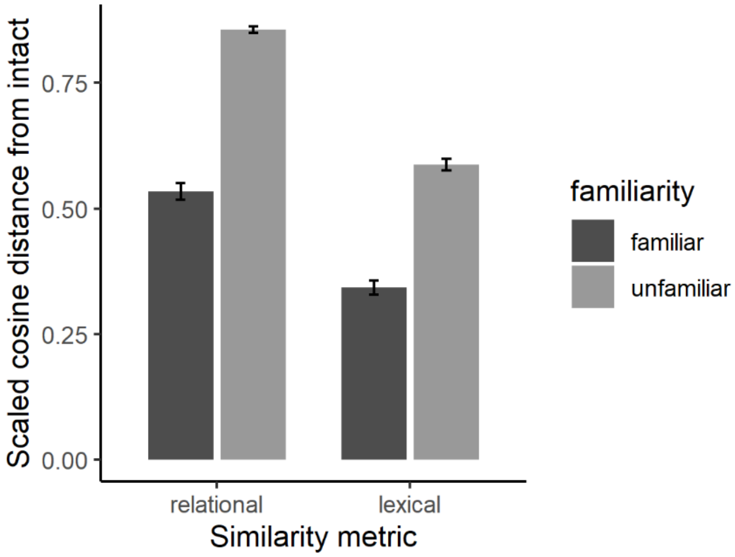


Figure 17: Mean lexical and relational cosine distances (scaled between 0 and 1) between familiar and unfamiliar word pairs and intact word pairs within each stimulus set used in Experiment 2.

Simulation results for Popov et al. (2017), Experiment 1. In order to provide a conceptual replication of the assessment of computational models I applied to my own experiments (as reported above), I used the same models to simulate human data reported by Popov et al. (2017) in their original demonstration of relational luring. Popov et al. reported human data collected for two different recognition memory tasks. The first task involved separate study and test phases and required participants to make binary 'old'/'new' judgments. The second task consisted of a more elaborate, continuous memory task, in which participants were presented with a long sequence of word pairs (> 500) and were asked to classify each stimulus into three categories based on its relation to word pairs already presented on previous trials in that sequence. Because my implementation of GCM (based on Nosofsky, 1988, 1991; Nosofsky & Zaki, 2003) produces binary responses and more naturally fits a design with separate study and test phases, I simulated the data for the first task reported by Popov et al. (2017), which was very similar to the present Experiment 2.

Popov et al.'s (2017) task consisted of three blocked study phases. In each phase, participants were instructed to commit 21 word pairs to memory. Following each study phase, participants completed a test phase in which they were presented with a different list of 21 word pairs, and were asked to provide binary responses indicating whether or not a given word pair was one of those that they had studied previously. On each test list, participants were presented with 7 old word pairs that had been shown during the prior study phase, and 14 new word pairs each consisting of individual words shown during the study phase, but that were novel in that they involved a combination of words different from any presented during the study phase. Of the 14 new word pairs, 7 were *relationally familiar* in that they were relationally similar to one of the studied word pairs (e.g., *floor:carpet* and *table:cloth* are relationally similar in that they both prominently instantiate the relation *is covered by*), and 7 were *relationally unfamiliar* in that they were not relationally similar to any of the studied word pairs. As in the present Experiment 2, the stimuli used by Popov et al. were constructed so that words were always placed in the same position in study and test pairs. Popov et al. demonstrated reliable relational luring on this task based on participant response times: Participants took longer to correctly classify new relationally familiar than new relational unfamiliar word pairs. The frequency with which participants misrecognized new pairs was numerically greater for relationally familiar than relationally unfamiliar recombinations, although this difference was not statistically reliable. (Importantly, the comparable pattern was reliable in the present Experiment 2.) I aimed to reproduce this trend based on models in the GCM framework, using the two similarity metrics, relational and lexical.

Using the same model-fitting procedure as Experiments 1 and 2, I found the maximum log-likelihood fit of the best parameters for each model, using item-level human data. Just as in Experiment 2, since individual word pairs were used in each condition, I treated word pair-condition combinations as unique items. Figure 18 presents false-alarm rates for model-generated $P(old|i)$ predictions using the best-fitting model parameters, and human data for familiar and unfamiliar word pairs. Again, using each similarity metric, GCM predicts the relational luring effect observed in the human data, as well as the higher hit rates for intact word pairs than false-alarm rates for recombined lures (human: M_{Hit} = .75, SD_{Hit} = .18, M_{FA} = .18, SD_{FA} = .13; relational: M_{Hit} = .72, M_{FA} = .17; lexical: M_{Hit} = .72, M_{FA} = .19). Similar to Experiment 2, but to a lesser extent, the luring effect generated using the relational similarity metric was closer in magnitude to the human effect than that generated using the lexical similarity metric. As was the case for Experiment 2, Popov et al. (2017) used materials that afforded more experimental control over the key manipulation of relational familiarity than those used in Experiment 1. The relational model's advantage in producing a more human-like luring effect in the present simulations thus strongly supports the importance of relation representations in accounting for human recognition memory.

An analysis of Popov et al.'s stimulus triplets (i.e., *intact*, *relationally familiar*, and *relationally unfamiliar* word pairs drawn from the same stimulus set) produced the same pattern of results as the corresponding analysis of Experiment 2's materials: Both the lexical and relational models yielded greater distances between intact and unfamiliar word pairs than between intact and familiar word pairs. The lexical model's ability to reproduce relational luring again stemmed from its partial success in capturing aspects of word meaning that track relational roles.

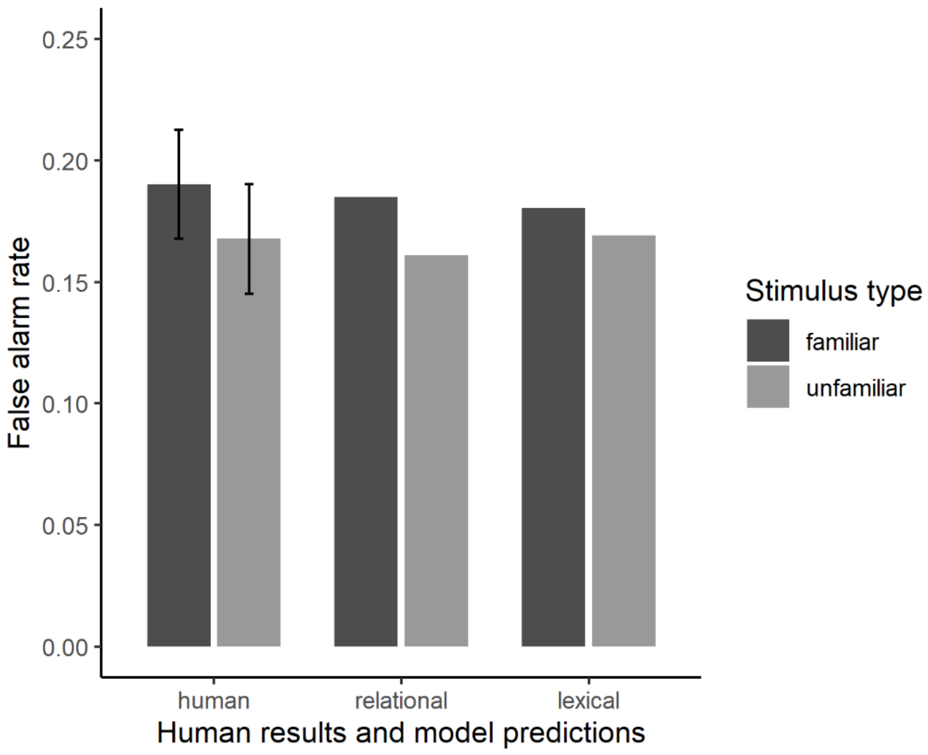


Figure 18: Human false-alarm rates from Popov et al. (2017), Experiment 1, and model predictions on the recognition memory task, broken down according to stimulus type.
Error bars reflect ±1 SEM.

In the same manner as described for the robustness analyses applied to data from my own experiments, I computed log model evidence for all 63 test items. Log model evidence was greater using relational similarity (E_{log} = -2.961 x 10^3) than lexical similarity (E_{log} = -3.614 x 10^3). I then examined the space of parameters for which relational and lexical similarity yielded relational luring within GCM for the data from Popov et al. (2017). Replicating the pattern of luring-specific model evidence for my own data in Experiments 1 and 2, I found that for the relational luring effect observed in Popov et al.'s (2017) data (M_{luring} = .0225, SD$_{luring}$ = .1078), relational similarity yielded greater model evidence (E_{luring} = 3.643) than did lexical similarity (E_{luring} = 3.631) across a wide range of the parameter space. Figure 19 depicts the luring effects produced by each similarity metric. As with Experiment 2, while the relational model showed only a slight advantage over the lexical model in luring-specific model evidence, it showed a substantial advantage over the lexical model in log model evidence. Thus for three datasets, relational similarity consistently produced a better account of the human data than lexical similarity across a wide range of model parameters.

Note the magnitudes of the fitted parameter values varied (even between Experiment 2 and the study by Popov et al., despite their use of very similar materials). These variations presumably are due to methodological differences, such as different encoding tasks (relatedness judgments in Experiment 2 vs. deliberate study in Popov et al.), number of task blocks (1 in Experiment 2 vs. 3 in Popov et al.), and task language (English in Experiment 2 vs. Bulgarian in Popov et al.).

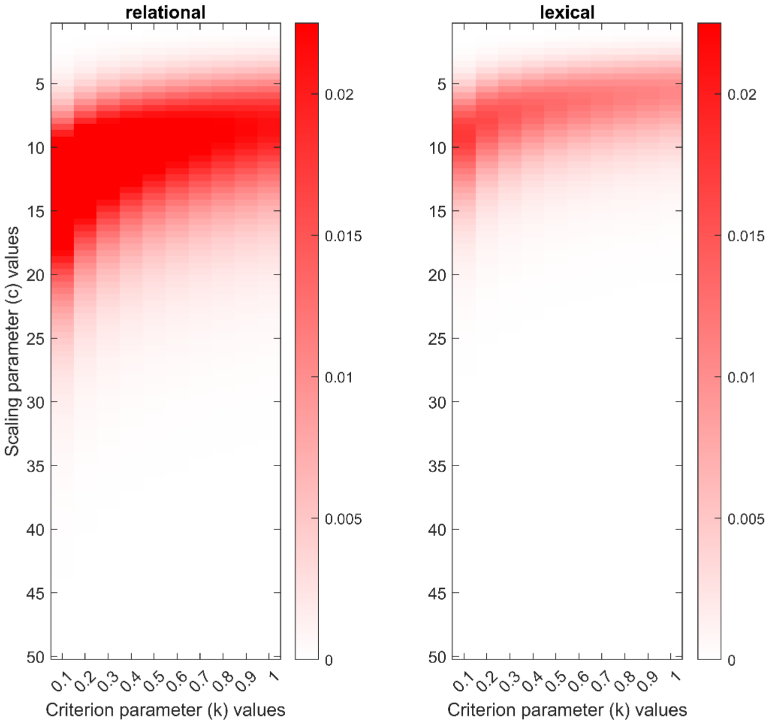


Figure 19: Simulation of model-predicted relational luring effect as a function of model parameters for Popov et al. (2017), Experiment 1.
Red color of cells indicates magnitude of model-predicted luring effect. The highest intensity of red corresponds to the magnitude of the luring effect in human data.

General Discussion

Summary

I report two experiments and simulations designed to compare alternative representations of word-pair similarity as predictors of both human analogical reasoning and recognition memory. I compared two computational models (both grounded in semantic vectors for individual words created by Word2vec; Mikolov et al., 2013) for defining the similarity between word pairs. One model was based on explicit relations between words, the other on lexical overlap between word meanings. The model based on explicit relations (BART; Lu et al., 2019) clearly provided the best account of human performance on an analogy task, in accord with previous work (e.g., Chiang et al., 2021; Ichien et al., 2021).

In my test of recognition memory, I replicated the phenomenon of relational luring reported by Popov et al. (2017): greater false recognition of word pairs formed by recombining studied words to form a novel instantiation of a familiar relation, as compared to recombinations that form an unfamiliar (i.e., unstudied) relation. I obtained the same basic pattern of false alarms using two different encoding tasks: judging whether a discernible semantic relation holds between two words in the relatedness task (Experiments 1 and 2), or judging whether two word pairs constitute a valid analogy in a verbal analogy task (Experiment 1). The fact that relation recognition yielded as much luring as an explicit analogy task is a surprising finding, as it seemed plausible that the former task would require less detailed processing of the relation. It is possible that participants paid close attention to the relation during both tasks because they expected a later memory test (as was also the case in the study by Popov et al., 2017). Alternatively, it may be that even relatively superficial relation processing is sufficient to produce the luring phenomenon. Future work will be needed to clearly disentangle the relative contributions of different encoding tasks to false recognition memory based on relations.

To assess the basis for relational luring using computational modeling, I tested the two similarity measures within a common theoretical framework provided by the Generalized Context Model (GCM; Nosofsky, 1988, 1991; Nosofsky & Zaki, 2003), a well-established instance-based model of item recognition. These computational analyses, which were applied to both experiments reported here as well as an experiment from Popov et al. (2017), yielded a nuanced interpretation. Relational similarity proved to be more accurate than lexical similarity in clustering word pairs instantiating different categories of semantic relations, but lexical similarity also was somewhat predictive (Figure 11). For all three datasets, when each model variant was fit using the optimal choice of values for the two parameters specified in GCM, the human pattern of relational luring could be predicted equally accurately using either relational or lexical similarity. Strikingly, my modeling results indicate that explicitly representing relations is not *necessary* for explaining relational luring.

However, I also performed additional analyses to assess the *robustness* of each similarity measure to variations in GCM's two model parameters: scaling parameter c and criterion parameter k. I first examined the space of parameters in the GCM model that predict item-level deviation between model predictions and human responses (using all data); and also the parameter space that specifically predicts the human luring effect. I computed the log model evidence to provide a quantitative comparison of the robustness to predicting all human data with each similarity metric. In addition, I computed luring-specific model evidence to quantitatively compare each similarity metric's ability to predict the human-generated luring effect. Both types of analyses were performed for data from Experiments 1-2 in the present paper and for Experiment 1 reported by Popov et al. (2017). For both analyses, across all three datasets, model evidence was greater for the relational similarity metric than for the lexical metric. In particular, the relational measure predicted the pattern of human data across a range of higher values of the GCM parameter *c*, which is typically interpreted as an index of sensitivity to differences among the instances stored in memory. Given the substantial procedural differences among the datasets that I modeled, the comparable findings from these analyses are particularly striking.

The greater robustness of the relational measure is consistent with the fact that this measure differentiated the abstract relation categories more accurately than did the lexical measure. In an explicit verbal analogy task in the *A:B::C:D* format, validity depends on the precise similarity of the *A:B* and *C:D* relations. Only relational similarity provides adequate precision to reliably compute validity. But in the recognition memory task, the instance-based GCM effectively computes similarity of any test pair to the entire pool of studied pairs. The GCM framework implies that the probability of incorrectly accepting a relational lure depends on its perceived similarity to an aggregate of all studied instances of that relation. If an agent is generally insensitive to subtle distinctions among individual word pairs, a coarse measure based on lexical similarity will suffice to yield greater false alarms to familiar than unfamiliar test pairs. But if the agent is instead highly sensitive to semantic distinctions among word pairs, only the more precise measure provided by relational similarity will predict a difference.

Conclusion

I conclude that by the preponderance of evidence (in particular, the greater robustness of the GCM model based on relational similarity), it is more probable that recognition memory for word pairs (like analogical reasoning) is based on explicit representations of relations between words, rather than on direct lexical similarity of individual words that form pairs. However, even if this (tentative) conclusion proves to be correct, it would not imply that lexical similarity is irrelevant to recognition. In fact, a basic requirement for obtaining relation-based false alarms is that the lure must be composed of words that were in fact shown in the study phase (in different combinations). That is, few false alarms would be expected if a test pair instantiated a familiar relation, but was composed of unstudied words. Moreover, even complex analogical reasoning by humans appears to be guided by lexical similarity of words *in addition* to similarity of explicit relations between words (Lu et al., 2022). It appears that a complete account of both reasoning and episodic memory will require integration of multiple types of similarity.

Having examined that the extent that relation processing impacts explicit comparison in Chapter 1, a process emphasizing human reasoning, and recognition in Chapter 2, a process emphasizing human memory, I move on in Chapter 3 to examine a process that integrates reasoning and memory: generative analogical inference.

Chapter 3: Relations and inference

Introduction

Human reasoners are remarkably sensitive to structural similarities. For example, despite the superficial differences between generational wealth accumulation and blood clotting, a brief elaboration of each reveals a clear analogy. In the first case, initial financial success allows a family to pass on wealth to the subsequent generation, which then grants that new generation access to social resources enabling its own financial success, affording further wealth to pass onto future generations. In the second case, an initial injury attracts blood platelets to cling to the injured site. Upon recognizing even this hint of a shared relational structure across these two processes, a reasoner can more easily map entities playing corresponding roles, such as wealth and blood platelets. Crucially, the reasoner could also infer that the presence of blood platelets would then attract yet more blood platelets to the injured site.

Computational models reproducing this ability to reason by analogy have been developed both in cognitive science (Falkenhainer et al., 1989; Hummel & Holyoak, 1997, 2003; Lu et al., 2022) and in artificial intelligence (P. W. Battaglia et al., 2018; Santoro et al., 2017; Shanahan et al., 2019). Models of analogical reasoning within cognitive science typically include explicit representations of relations, such that a relation is distinct from, but bound to the entities it relates. This property supports the recognition of structural similarity by enabling a direct comparison of the relations constituting each analog. Crucially, explicit relation representations can also constrain the *generation* of predictions about a target analog based on the source. Indeed, the generative capacity afforded by relation representations is the core of analogical inference, which human reasoners can exploit in everyday problem solving (Gick & Holyoak, 1980, 1983), technological innovation (Kittur et al., 2019), and scientific discovery (Gentner, 2002; Holyoak & Thagard, 1994b; Nersessian, 1992).

Here I introduce a new computational model of analogical inference. Like existing inference models (Burstein, 1983; Carbonell, 1983, 1993; Falkenhainer et al., 1989; Halford et al., 1994; Hofstadter & Mitchell, 1994; Holyoak & Thagard, 1989; Hummel & Holyoak, 2003; Keane & Brayshaw, 1988; Kokinov, 1994), the present model can reproduce inferences from pre-specified relations (as demonstrated in Simulations 1a and 1b). Unlike existing models, this model can also reproduce inferences from analogs for which relational structure is unspecified in the input (as demonstrated in Simulations 2-4). This model, BART-Gen, operates on explicit relation representations generated by BART (*Bayesian Analogy with Relational Transformations*) (D. Chen et al., 2017; Lu et al., 2012, 2019), a model of relation learning that acquires representations of relations from unstructured vector representations of individual word meanings. Many previous analogy models have relied on representations that are hand-coded by the modeler, and thus bypass the problem of relation acquisition altogether (Chalmers et al., 1992). In contrast, BART deals directly with the problem of learning relations from non-relational inputs, taking as inputs embeddings for individual words produced by machine-learning algorithms.

BART's relation representations have been used to predict human judgments of relational similarity among word pairs (Ichien et al., 2022), to support human-like analogical reasoning on simple four-term verbal problems (e.g., *artificial : natural :: friend : enemy*) (Lu et al., 2019), and to predict patterns of similarity in neural responses to relations during analogical reasoning (Chiang et al., 2021). When used as input to a mapping model, BART also can support analogical mapping in problems requiring finding correspondences between multiple entities across complex relational systems (e.g., mapping the solar system to atomic structure) (Lu et al., 2022).

Analogical inference

Analogical inference enables a reasoner to elaborate on their understanding of some *target* domain by exploiting an analogy between it and a better-understood *source* analog. In the case of analogical problem solving, source analogs may permit elaboration of a target problem that reveals its solution (Gick & Holyoak, 1980, 1983). Cognitive scientists have studied this ability as a component process of the capacity for analogical reasoning, which also involves in *retrieving* of one or more relevant source analogs given a target, *mapping* systematic correspondences between components of the source and target, and *schema induction* to form a more abstract representation capturing commonalities shared by the source and target (Holyoak et al., 1994).

Schema-governed categorization (SGC). One mechanism for generative inference relies on schema-governed categorization (SGC), recognizing some situation as a candidate instance of some schema-governed or relational category the members of which all instantiate some common relational structure (e.g., *positive-feedback loop* or *convergence*; Gick & Holyoak, 1983; Markman & Stilwell, 2001). In cases where the target situation lacks some properties or relations that are present in the schema that it's hypothesized to exemplify, a reasoner can use their schema to fill in those missing properties, roles, or relations to extend their knowledge of the target.

Ultimately, SGC depends both on a reasoner having a schema that potentially applies to the target and some understanding of the structure governing the target that can constrain their retrieval of that schema. By highlighting the common structure across instances, comparison plays an important role both in the acquisition of schemas or relational concepts and in the recognition of exemplars of such concepts (Christie & Gentner, 2010; Doumas et al., 2008; Gick & Holyoak, 1983; Hummel & Holyoak, 2003; K. J. Kurtz et al., 2013; Markman & Gentner, 1993b; Namy & Gentner, 2002).

Gick and Holyoak (1983) asked participants to solve Duncker's radiation problem, in which a doctor is tasked with using radiation to kill a patient's malignant stomach tumor despite the harm radiation poses to the patient's healthy tissue (Duncker, 1945). But before doing so, the researchers prompted an initial act of comparison by having participants first describe the similarities between a pair of apparently dissimilar stories but that were both analogous to each other and to the scenario described in the radiation problem. Importantly, both stories instantiated a basic convergence structure, in which some pool of resources (e.g., fire-retardant foam) was divided into smaller pools that converged on some central location (e.g., small firehoses converging to spray foam on a central fire) in order avoid some obstacle (e.g., lack of firehoses large enough to put out the fire individually). This act of comparison increased participants' solution rates on the radiation problem, suggesting that it provided participants a clue for solving the radiation problem (i.e., dividing the radiation into smaller rays, which could be fired separately to converge on the stomach tumor), and they could have only used this clue if they grasped the relational structure common to the problem and the two stories that they had previously compared. Moreover, independently-rated success in articulating that convergence structure in their similarity descriptions further predicted participant solution rates. This latter result suggests that the extent to which participants induced and applied a *convergence* schema predicted their problem solving success via SGC: Specifically, to consider *dividing* the radiation into smaller rays and to fire them so that they *converge* on the stomach tumor.

Copy-with-substitution-and-generation (CWSG). Notably, inference via SGC necessitates that a participant has some crystallized schema to exploit, but not all analogical inference relies on such complex representations. In the following, I describe an alternative that relies on a direct comparison between analogs without positing some mediating abstract schema.

Most existing models of analogical inference implement a simple pattern-completion mechanism, “copy with substitution and generation” (CWSG) (Burstein, 1983; Carbonell, 1983, 1993; Falkenhainer et al., 1989; Halford et al., 1994; Hofstadter & Mitchell, 1994; Holyoak et al., 1994; Holyoak & Thagard, 1989; Hummel & Holyoak, 2003; Keane & Brayshaw, 1988; Kokinov, 1994). Whereas SGC, discussed above, relies on a top-down categorization process, CWSG provides a bottom-up route to analogical inference that depends on a partial mapping between source and target analogs. The shared relational structure revealed in previously-recognized correspondences between the source and target, together with unmapped elements in the source, jointly constrain the generation of novel elements in the target. The main distinction between CWSG and SCG, then, resides in the *origin* of the particular constraints that each place on inference. Whereas with SGC these constraints consist of properties or relations from an abstract schema (perhaps retrieved from semantic memory) that are unattributed to the target prior to inference, with CWSG they consist of relations governing a source analog (perhaps retrieved from episodic memory) that are unmapped to the target prior to inference.

Consider Gick and Holyoak's (1980) analogy between a source story of a general attempting to overthrow an authoritarian dictator with an army of troops despite land mines preventing access to the dictator, and a target, Duncker’s radiation problem mentioned above (Duncker, 1945). A reasoner might readily map correspondences from the general to the doctor, the dictator to the tumor, and both the troops and land mines to the radiation. Notably, the ray problem omits mention of any event that directly corresponds to the general’s solution of *dividing* his army into small groups that are each light enough to avoid setting off the land mines and that are thus able to *converge* on the castle from different directions. While this solution is as yet unmapped to the target, it can serve to constrain an analogous solution in the target, in which the doctor *divides* his radiation into less-intense rays that are each weak enough to pass through the patient’s healthy tissue without damaging it and are thus able to *converge* on the patient’s tumor from different directions. Specifically, a reasoner integrates the unmapped elements of the source into the relational structure revealed during mapping, identifying relations between the unmapped elements and those with direct correspondences in the target (e.g., *divide* and *converge*) and thereby producing an elaborated relational structure governing the source (e.g., the general’s solution). Under CWSG, these newly-identified relations are then copied over from the source as partially-filled relations in the target: Mapped elements in the source serving as relata in these newly-identified relations are substituted with their direct correspondences in the target (e.g., the doctor’s potential solution to *divide* his radiation into multiple weak rays that *converge* on the patient’s tumor). Finally, novel elements in the target are proposed to complete these partially-filled relations (e.g., multiple weak radiation rays that are jointly powerful enough to kill the tumor). While CWSG and SGC constitute distinct routes to generative inference (Gick & Holyoak, 1983; Minervino et al., 2023), the present study focuses on CWSG, and particularly the status of relations in analogical inference.

Relations in CWSG. Because relations provide such an important constraint on CWSG, the success of models of analogical inference depends heavily on the nature of their relation representations. Since most existing models of this process operate on relation representations that have been hand-coded by the modeler, it is difficult to determine whether any explanatory success of these models to their inference process, or to their highly tailored relation representations that may reflect modelers’ idiosyncratic and potentially erroneous assumptions about the nature of human relation representation (Forbus et al., 2017).

Moreover, these models require that the specific relations constituting the relational structures governing source and target analogs must be pre-specified in the input to their hypothesized instantiation of analogical inference. This requirement places practical limits on these models, restricting their applicability to large-scale inference over arbitrary domains, where direct coding may be prohibitively labor-intensive. More importantly, this requirement also limits their theoretical scope, rendering them unable to explain situations in which a human reasoner infers, for example, that the rind of a watermelon is analogous to a cigarette butt without first providing some discrete relation concept that is instantiated both by *rind* and *watermelon* and by *butt* and *cigarette* (something like *disposable-part-of*).

One source of this requirement for pre-specified relation concepts is the use of symbol-argument-argument (SAA) notation to model relations as multi-place predicates, often expressed as verbs (e.g., *chase*(*X*,*Y*)) or prepositions (e.g., *above*(*X*,*Y*)) in natural language (Burstein, 1983; Carbonell, 1983, 1993; Falkenhainer et al., 1989; Halford et al., 1994; Hofstadter & Mitchell, 1994; Holyoak et al., 1994; Holyoak & Thagard, 1989; Keane & Brayshaw, 1988; Kokinov, 1994). While this approach to relation representation captures the sense in which relations are explicitly structured (i.e., dynamically bound to their relata), it cannot capture the semantic richness of relation concepts (Doumas & Hummel, 2004). Specifically, SAA notation cannot capture graded similarity among relation concepts; within SAA, relations are either identical to each other or not; e.g., *love* is no more similar to *like* than is *kill* (but see Forbus et al., 2017; Silliman & Kurtz, 2019, for an alternative perspective).

An alternative to SAA notation is to represent the semantic content of relations either as distributions over the lexical features that characterize their relata (Doumas et al., 2008; Hummel & Holyoak, 2003), or within a separate representational space constituted by distinctively relational features (Chaffin & Herrmann, 1987; Lu et al., 2019). The latter approach to relation representation potentially captures gradations in relational similarity that characterize human relational knowledge (Chaffin & Herrmann, 1984; Ichien et al., 2022; Perfetti, 1967; Popov et al., 2020; Winston et al., 1987). More pertinently, representing relations within a multi-dimensional feature spaces enables representation of relations that have no prior lexical entry. However, this approach does not directly explain how an inference model is able to handle the lack of relations pre-specified in the input.

Eduction of relations

In the present paper I introduce a computational model capable of analogical inference regardless of whether the relations it operates over are pre-specified in its input (as required by existing models) or are unspecified (unlike existing models). Like its predecessors, the present model implements CWSG, but it differs from previous proposals in the way that relation representation is treated. The present model departs from SAA notation, representing relations as distributed vectors within a representational space, where abstract relations constitute individual dimensions (Lu et al., 2019, 2022). This model emphasizes Charles Spearman's observation that analogical reasoning often depends on what he termed the *eduction of relations* (Spearman, 1923): mentally "filling in the blanks" in the problem as posed, by retrieving or computing relations between constituent elements.

For verbal problems, Spearman's concept of "relation" refers to the semantic relation between lexical concepts. Semantic relations are more than mere associations; e.g., *hot* : *cold* :: *love* : *adore* consists of two word pairs that are each strongly associated via a salient relation, but the problem does not form a valid analogy because the *A:B* and *C:D* relations mismatch. At the same time, semantic relations are not necessarily represented as "predicates" as typically incorporated into analogy models as verbs or prepositions (e.g., *chase*(*X*,*Y*)) or prepositions (e.g., *above*(*X*,*Y*)). In contrast, one can represent the proposition *dogs chase cats* by identifying semantic relations for three pairs of content words: e.g., *dogs* : *chase, chase* : *cats*, *dogs* : *cats*. Rather than denoting relations themselves, verbs, like nouns, denote concepts that enter into pairwise semantic relations.

That the present model can operate on input in which relations are not directly stated raises an important question about analogical inference: What do relations contribute? In assuming that relations governing analogical inference are pre-specified, existing models of analogical inference make the logically prior assumption that humanlike inference requires explicit representations of relations. In principle, analogical inference could be performed without explicit relation representations, and this proposal has been explored to a small extent with limited success (Leech et al., 2008; Mikolov, Sutskever, et al., 2013; Peterson et al., 2020; Rumelhart & Abrahamson, 1973). In the present chapter, I compare our model, which operates on explicit relation representations, with control models lacking such representations. In a series of simulations, I systematically show that relations contribute to analogical inference by enabling generalization across semantic domains: leveraging analogies across a source and target is robust to variations in the similarity or degree of association between the source and target (Doumas & Hummel, 2005; Gentner, 1983; Holyoak, 2012).

Overview of the chapter

In this chapter, I introduce a new model of analogical inference that operates on relation representations acquired by an existing model of relation acquisition and representation, Bayesian Analogy with Relational Transformation (BART) (Lu et al., 2012, 2019). I describe relation representation in BART before detailing inference in our new model, BART-Gen. I move on to detail four simulations that demonstrate BART-Gen's ability to perform inference with pre-specified relations in simple phrases (e.g., *A robin is a kind of ?*; Simulation 1) and with unspecified relations in four-term analogies (e.g., *blindness : sight :: poverty : ?*; Simulation 2), six-term analogies (e.g., *weapon : gun : rifle :: clothing : sweater : ?*; Simulation 3), and extended analogies akin to those used in early studies of analogical problem solving (Gick & Holyoak, 1980, 1983) (e.g., *solar system : planet : mass : gravity : sun :: atom : electron : charge : electromagnetism : ?;* Simulation 4).

Across four simulations I compare this inference model, which operates on explicit relation representations, with control models that lack such representations. In order to evaluate models within these simulations, I collected human-generated responses in open-ended naturalistic experiments and compare models in their ability to produce frequently-generated human responses. The contribution of this work is twofold: 1) I introduce a novel model of analogical inference capable of operating on input both when relations are pre-specified and when they are unspecified, and 2) I systematically compare our inference model with control models lacking relation representations to clarify what relations contribute to the inference process.

Computational modeling

Relation representation in BART

BART[1] learns explicit representations of the semantic relations between word pairs from unstructured vector representations of individual word meanings (Lu et al., 2012, 2019). In the present simulations, BART's input consists of concatenated pairs of word vectors from Word2vec[2] (Mikolov, Sutskever, et al., 2013) and uses supervised learning with positive and negative

[1]https://cvl.psych.ucla.edu/wp-content/uploads/sites/162/2021/04/BART2code.zip

[2] https://code.google.com/archive/p/word2vec/

examples to acquire each relation representation individually. For example, a vector formed by concatenating the individual vectors for *old* and *young* would constitute a positive example for the relation *X is the opposite of Y*, and might also serve as a negative example of the relation *X is a synonym of Y*. After learning, BART computes a relation vector consisting of the posterior probability that a word pair instantiates each of the learned relations.

As shown in Figure 1, the BART model uses a three-stage process to learn a broad range of semantic relations. In its first stage, BART uses difference-ranking operations to partially align relationally important features. The model generates a ranked feature vector based on the difference values between the raw feature vectors of two entities, but ordering those values according to their magnitude. Augmenting the raw semantic features with ranked features addresses the issue that across instances different semantic dimensions may be relevant to a relation. This first stage culminates in the generation of a 1200-dimension augmented feature vector for each word pair, consisting of the concatenation of raw and ranked feature vectors for each word in the pair.

In the second stage, BART uses logistic regression with elastic net regularization to select a subset of important feature dimensions across word pairs f_s. In the third stage, BART uses Bayesian logistic regression with f_s to estimate weight distributions w for representing a particular relation R by applying Bayes rule as:

$$P(w|f_s, R) \propto \; P(R|f_s, w)P(w). \tag{9}$$

The first term is the likelihood defined by a logistic function on w and f_s (selected in the second stage), $\frac{1}{1+e^{-w^T \mathrm{f_s}}}$. The second term is the prior distribution of w, defined as a multivariate normal distribution, $N(\mu_0, \Sigma_0)$, with a mean vector $\mu_0 = (\beta, -\beta)$, consisting of the β values of weights estimated in the second stage of logistic regression.

BART was trained by combining two datasets of human-generated word pairs, each chosen as an example of a specific semantic relation. The first dataset (Jurgens et al., 2012) consists of at least 20 word pairs (e.g., *bird:robin*) instantiating each of 79 semantic relations (e.g., *X is a type of Y*) taken from a taxonomy proposed by Bejar et al. (1991), which includes 10 major relation categories (e.g., *class inclusion*). The second dataset consists of at least 10 word pairs instantiating each of 56 additional semantic relations (Popov et al., 2017). Across both datasets, BART acquired 135 semantic relations via supervised learning. Since BART's learned weights w can be expressed as two separate halves (i.e., those associated with the first relational role, w_1, and those associated with the second relational role, w_2), BART can automatically generate representations of the converse of each learned relation by swapping the relation weights associated with each individual relational role. Thus, upon learning a representation of *X is a type of Y*, BART can also learn a representation of its converse, *Y is a superordinate of X*—the same relation but with the roles flipped. This operation effectively doubles BART's pool of learned relations from 135 to 270 in total.

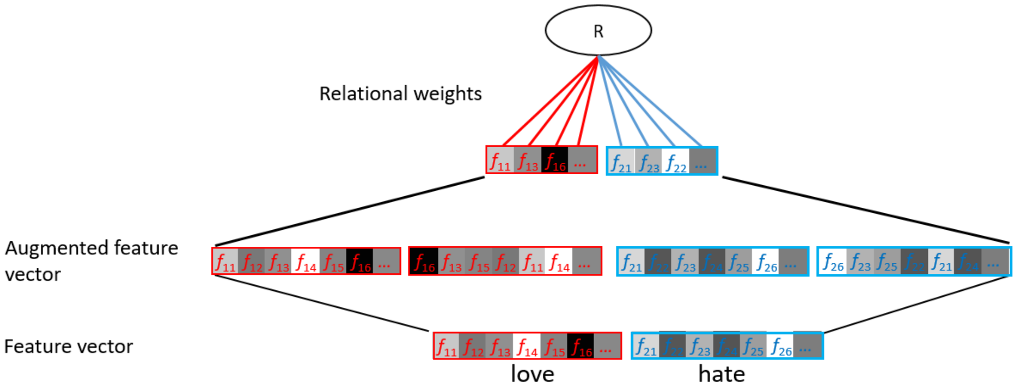

R
Relational weights
f11 f13 f16 ...
f21 f23 f22 ...
Augmented feature vector
f11 f12 f13 f14 f15 f16 ...
f16 f13 f15 f12 f11 f14 ...
f21 f22 f23 f24 f25 f26 ...
f26 f23 f25 f22 f21 f24 ...
Feature vector
f11 f12 f13 f14 f15 f16 ...
f21 f22 f23 f24 f25 f26 ...
love
hate

Figure 20: An illustration of three-stage model of BART for learning a relation from word pairs.

After learning weight distributions associated with selected features across word pairs in its training set f_L,R_L, BART can estimate how likely any novel pair of words A and B instantiates a learned relation R_i, $P(R_i|f_A, f_B)$ by marginalizing the weight distribution for that relation:

$$P(R_i|f_A, f_B) = \int P(R_i|f_A, f_B, w)P(w|f_L, R_L)dw. \qquad (10)$$

Hence, given any pair of words A:B, BART can perform this operation for each of its learned relations and then generate a relation vector R_{AB}, in which the value of each element is a posterior probability reflecting how good an example A and B are of that particular relation, as shown in Figure 21. For example, given that *old* and *young* constitute a good example of the relation *X is the opposite of Y* but a poor example of the relation *X causes Y*, $R_{old:young}$ would have a high value for the dimension corresponding to the first relation, but a low value for the dimension corresponding to the second dimension. Ichien et al. (2021) added a power transformation to these relation vectors, raising each relation dimension to a power of 5, and found that adding this transformation ("winners take most") improves the model's ability to capture human judgments of relational similarity. Accordingly, I incorporated the same power transformation in the present simulations.

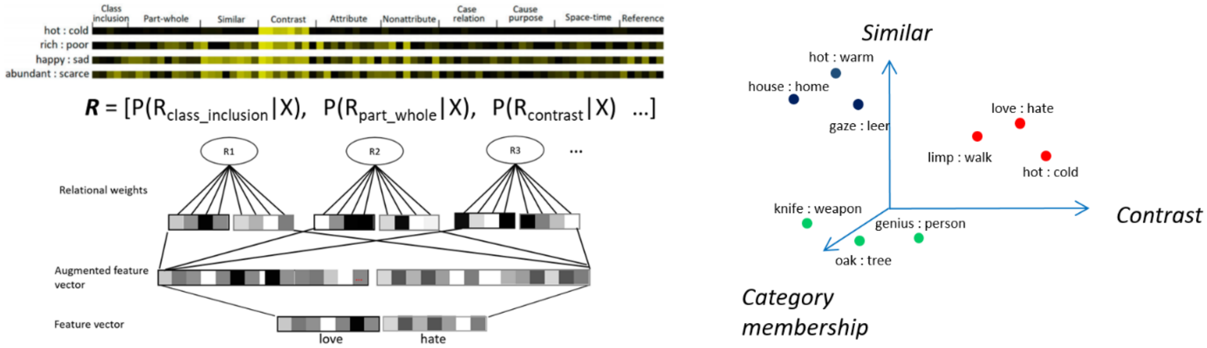


Figure 21: An illustration of using relation vector to capture distributed representations of semantic relations in BART.

Generative inference in BART-Gen

BART-Gen uses the relation representations acquired by BART to perform analogical inference of generating individual entities using relation information. I first detail its algorithm for inference using pre-specified relations, and then describe the extended algorithm for inference using unspecified relations.

Inference with a specified relation. Recall that the second stage of BART's learning algorithm uses logistic regression with elastic net regularization to select a subset of informative feature dimensions of a word pair, f_s. Given the individual words combined within that word pair, these selected feature dimensions can be separated into those corresponding to a query word, f_{s_1}, and those corresponding to the other word, f_{s_2}. Given a first word and a relation, and the hypothesis that a relation R holds between this word and the other word, BART-Gen generates a probability distribution of the second word f_{s_2}, using the following inference:

$$P\left(f_{s_2} | R = 1, f_{s_1}\right) \propto P\left(R = 1 \middle| f_{s_1}, f_{s_2}\right) P\left(f_{s_2} \middle| f_{s_1}\right). \tag{11}$$

The likelihood term, $P\left(R = 1 \middle| f_{s_1}, f_{s_2}\right)$, is the probability that R holds for the generated word with the feature vector of f_{s_2} and the query word with the feature vector of f_{s_1}. As with Equation 9, the likelihood term $P\left(R = 1 \middle| f_{s_1}, f_{s_2}\right)$ is defined using a logistic function:

$$P\left(R = 1 \middle| f_{s_1}, f_{s_2}, w_1, w_2\right) = \frac{1}{1+e^{-w_1^T f_{s_1} - w_2^T f_{s_2}}}. \tag{12}$$

In Equation 12, the mean vectors of weight distribution, w learned in BART, are written as two separate components: those associated with the first word's relational role, w_1, and those associated with the second word's relational role, w_2. Note that I only used the mean values of the weights, and did not include the variability in the distribution. Correspondingly, the selected feature dimensions of a given word pair f_s are rewritten as those corresponding to the first word, f_{s_1}, and the second word, f_{s_2}.

The prior term, $P(f_{s_2}|f_{s_1})$, follows a multivariate normal distribution with the mean as the feature vector of the first word f_{s_1}, which is defined as:

$$P\left(f_{s_2}\middle|f_{s_1}\right) = N(f_{s_1}, \sigma^2 I). \tag{13}$$

BART-Gen uses the semantic embedding of the first word as a prior for generating D, in that the means of the prior $P(f_{s_2}|f_{s_1})$ are the feature values of the first word, reflecting the assumption that the two words are semantically associated. The prior term also assumes equal variance σ^2 for semantic features of the second word. σ^2 is a parameter that controls the degree to which the generated word is semantically associated with the query word in the prior. Larger values of σ^2 correspond to a weaker degree of prior semantic association in the inference.

To compute the inference in Equation 12, I adopted a variational method for Bayesian parameter estimation (Jaakkola & Jordan, 2000), and used the following updating rules for the mean **μ** and covariance matrix **V** of the feature distribution for the generated target word, as well as the variational parameter ξ:

$$\begin{gathered}\boldsymbol{V}^{-1} = \frac{\boldsymbol{I}}{\sigma^2} + 2\lambda(\xi)\boldsymbol{w}_2\boldsymbol{w}_2^{\boldsymbol{T}}, \\ \boldsymbol{\mu} = \boldsymbol{V}\left(\frac{\boldsymbol{I}}{\sigma^2} f_{s_2} + \frac{\boldsymbol{w}_2}{2} - 2k\lambda(\xi)\boldsymbol{w}_2\right), \\ \xi^2 = \boldsymbol{w}_2^{\boldsymbol{T}}(\boldsymbol{V} + \boldsymbol{\mu}\boldsymbol{\mu}^{\boldsymbol{T}})\boldsymbol{w}_2,\end{gathered} \tag{14}$$

where $\lambda(\xi) = \frac{tanh\left(\frac{1}{2}(\xi+k)\right)}{4(\xi+k)}$ and $k = \boldsymbol{w}_1^{\boldsymbol{T}} f_{s_1}$.

Implementational details. To determine the value of σ^2 in the prior term, I rely on an assessment of the sparsity of the query word's semantic neighborhood (i.e., representational space populated by its nearest neighbors), where larger values of σ^2 are used for problems where the space around the query word is sparsely populated with other words (i.e., where there are relatively few words around the query word), and smaller values of σ^2 are used for problems where the query word is densely populated. *Ceteris paribus,* this computation increases the degree of semantic association between the query word and a to-be-generated target word if the query word has several close associates and decreases that semantic association if the query word does not have many close associates. I compute this sparsity value (i.e., the variance of the query word's semantic neighborhood) as the average of squared Euclidean distance between *C* and its 100 nearest neighbor words, normalized by the dimensionality $dim_f$of the semantic space to capture the variability for each dimension (e.g., $dim_f = 300$ for the word2vec vectors I use in the present simulations):

$$\sigma^2 = \frac{\sum_{k=1}^{100} d(C,k)^2}{100 dim_f{}^2}. \tag{15}$$

The BART-Gen inference balances the likelihood guided by relation representation and the prior guided by semantic similarity to the query word, so as to generate maximum a posteriori (MAP) estimates of feature values for the generated target words on selected dimensions, $\hat{f}_{s_2}$.

Note that f_{s_2} is only a subset of all feature dimensions along which the generated target word is represented, f_2. In order to generate semantic embedding for the generated target word along the feature dimensions that were *not* selected by BART's learning algorithm, BART-Gen simply copies over the corresponding feature values from the query word, f_{ns_1}. Hence, by combining the generated feature values for selected dimensions and copying feature values for unselected feature dimensions, BART-Gen specifies a complete prediction for f_2 for a specific query word and a relation:

$$\hat{f}_2 = \langle f_{ns_1}, \hat{f}_{s_2} \rangle. \tag{16}$$

We evaluate this algorithm for inference with pre-specified relations in Simulation 1.

Inference with unspecified relations. Solving a generative analogy problem, *A:B :: C:?*, requires generating a *D* word such that the word pair formed by *C* and generated *D* instantiate the same relations as the source word pair consisting of *A* and *B*. BART-Gen generates the target *D* word using a maximum posterior estimate given the embeddings of words *A*, *B* and *C*,

$$\widehat{f}_D = argmax\, P(f_D \mid f_C, f_A, f_B). \tag{17}$$

To solve this task, BART-Gen first needs to represent the relation holding between *A* and *B*. To do this, BART-Gen applies Equation 10 to word pair *AB* to infer relations instantiated by this word pair by estimating a relation vector R_{AB}. The model then uses a delta function to transfer the relations from AB and the query word C to generate the D word, as shown in the equation below:

$$P(f_D \mid f_C, f_A, f_B) = \int P(f_D|r, f_C)\delta(r - R_{AB})P(R_{AB}|f_A, f_B)\,dr. \tag{18}$$

The first term in Equation 18 $P(f_D|r, f_C)$ can be computed using Equation 3 for inference with a specific relation; the second term is $\delta(r - R_{AB})$, a delta function showing the same relation used in word pair AB and word pair CD; the third term $P(R_{AB}|f_A, f_B)$ can be calculated using Equation 10 to infer the relation vector for the AB word pair.

BART-Gen relies on a distributed vector representation of the relation holding between a pair of concepts A and B, R_{AB}, populated by posterior probabilities corresponding to a distinct relation learned by BART (see Equation 10). BART-Gen then forms a transient, explicit representation of that relation and then applies it to generate a prediction of *D*, given the constraint that *C* and *D* instantiate the same relation.

Implementational details. BART-Gen applies a simple filtering mechanism to R_{AB}, which sets relation probability in the bottom 25th quantile to be 0. This mechanism helps reduce noise in

BART-Gen's relation representation R_{AB} by enabling it to ignore any relation dimensions that are not highly expressed in word pair *AB*.

For four-term *A:B :: C:D* problems, the inference is a combined result from two constituent analogies respectively based on the relations shared between the *AB* pair and *CD* pair, and between the *AC* pair and *BD* pair. To infer the target word *D*, I repeat this process twice by applying the relation holding between *A* and *B* that is partially filled by *C* to generate an embedding for the missing *D*, and the relation holding between *A* and *C* that is partially filled by *B* to generate another embedding for the missing *D*. These two embeddings are then combined in a weighted sum,

$$\widehat{f}_D = \alpha * \widehat{f}_{D|R_{AB}} + (1 - \alpha) * \widehat{f}_{D|R_{AC}}, \quad (18)$$

The weight α is determined by their degree of semantic association, as computed by the relative ratio of cosine similarity:

$$\alpha = \frac{sim(f_A, f_B)}{sim(f_A, f_B) + sim(f_A, f_C)}. \qquad (19)$$

This approach, which I evaluate in Simulation 2, enables BART-Gen to exploit any relations holding *across* analogs, especially in semantically near analogies, where *A* and *C* are highly associated, and thus some meaningful relation between source and target analogs is more likely to exist. For example, in the near analogy *blindness:sight :: deafness:?*, there is a fairly clear *lack-of* relation holding between *blindness* and *sight*, and there also some meaningful relation between *blindness* and *deafness* that can also contribute to the generation of *hearing*.

For analogies of greater complexity (e.g., six-term analogies *A:B:C :: D:E:?*), I repeat the process instantiated in Equations 16 and 17 for each four-term analogy that includes the to-be-generated term (e.g., *A:C :: D:?* and *B:C :: E:?*). This process results in multiple unique embeddings (i.e., one embedding for each four-term analogy); similar to Equation 18, I combine these embeddings in a weighted sum,

$$\widehat{f}_{Gen} = \sum_{i=1}^{n} \alpha_i \widehat{f}_{Gen|R_{AB_{ik}}}, \quad (20)$$

where f_{Gen} is the embedding representing the to-be-generated term, k is the term that is directly analogous to the to-be-generated term (e.g., *C* in *A:B:C :: D:E:?*), and n is the number of four-term analogies that include this to-be-generated term (e.g., 2 for six-term analogies). This weighted sum is scaled according to the degree of semantic association between the source terms in each analogy, using a variant of Equation 19:

$$\alpha_i = \frac{sim(f_i, f_k)}{\sum_{j=1}^{n} sim(f_j, fk)}. \quad (21)$$

This approach, which I evaluate in Simulation 3, enables BART-Gen to incorporate the entire source analog in its proposal for the to-be-generated term.

Control Model 1: Bidirectional Encoder Representations from Transformers (BERT). For comparison with BART-Gen for Simulations 1 and 2, we derived generative inferences from a major natural language processing (NLP) model, *Bidirectional Encoder Representations from Transformers* (BERT; Devlin et al., 2019). BERT (no relation to BART!) is a prominent example of a transformer architecture. Like other similar NLP models, BERT is trained on a masked-language modeling task, in which it predicts masked words in sentences drawn from huge text corpora. Given an incomplete sentence such as “A [MASK] is a type of bird.”, BERT is trained to predict words that would complete that sentence with the highest probability. Importantly, BERT and similar models routinely solve generation tasks without any explicit relation representations, instead relying solely on the statistics of word usage in their training corpora.

Control model 2: Word2vec parallelogram model of analogy. As another control model in Simulations 2 and 3, I implement Rumelhart and Abrahamson's (1973) parallelogram model of analogy. I implement this model using Word2vec word embeddings, which has shown some success in solving simple verbal analogy problems (Mikolov et al., 2013; but see Peterson et al., 2020, for simulations and extended discussion of its limitations). This model operates over lexical representations only, and does not represent relations explicitly in the sense that they constitute representations that are separable from their relata. Instead, relations are treated as the generic difference vector between distributed representations of lexical items.

The clearest way to characterize this model is to instantiate it in a four-term *A:B :: C:?* generative analogy, which requires generating a *D* word such that the difference vector between f_C and generated $\widehat{f}_D$ instantiate matches that between f_A and f_B:

$$\widehat{f}_D = f_C - f_A + f_B. \qquad (22)$$

We compare this model to BART-Gen as a non-relational control for four-term analogy problems in Simulation 2.

For analogies of greater complexity, I implement a Equation 13 and variant of Equation 20, following the same basic scheme as BART-Gen. Similar to what I describe above, I repeat the process instantiated in Equation 22 for each of *n* unique four-term analogies that include the to-be-generated term (e.g., *A:C :: D:?* and *B:C :: E:?*), and this process results in multiple unique embeddings (i.e., one embedding for each four-term analogy),

$$\widehat{f}_{Gen} = \sum_{i=1}^{n} \alpha_i \widehat{f}_{Gen|f_i - f_k}. \qquad (23)$$

Here, f_{Gen} is the embedding representing the to-be-generated term, *k* is the term that is directly analogous to the to-be-generated term (e.g., *C* in *A:B:C :: D:E:?*), and *n* is the number of four-term analogies that include this to-be-generated term (e.g., 2 for six-term analogies). This weighted sum is scaled according to the degree of semantic association between the source terms in each analogy, using Equation 21. I compare this extension of the parallelogram model to BART-Gen in Simulation 3.

Simulation 1: Inference using pre-specified relations

In the first simulations, I test BART-Gen's ability to reason from pre-specified relations. I operationalize this capacity as generating a word *D* (e.g., *bird*) that best instantiates a known relation *R* (e.g., *is a type of*) with a query word C (e.g., *robin*). I restrict my analyses to those relations for which BART has learned an explicit representation, comparing the performance of BART-Gen with that of BERT.

Simulation 1a: Jurgens et al. (2012)

I evaluated model performance in their ability to produce human-like responses to partially-filled relations, formatted as sentence-completion problems (e.g., "A robin is a type of ____."). In order to construct these problems, I used the dataset of human-generated word pairs used to train BART (Jurgens et al., 2012), thus ensuring that BART-Gen had an explicit representation of each relation mentioned in these problems. Each of these word pairs were generated as an example of some semantic relation (e.g., *robin:bird* exemplifies the relation *X is a type of Y*), and I combined each word pair with its semantic relation to produce sentence-long statements (e.g., "A robin is a type of bird."). Each statement was then used to generate two problems, one omitting the first word of its word pair (e.g., "A ____ is a type of bird.") and the other omitting the second word (e.g., "A robin is a type of ____."). I refer to the remaining word of each word pair in a given problem as the "query word" (e.g., "bird" for the first problem above and "robin" for the second problem above).

I collected human responses to sentence-completion problems generated from a selection of 16 statements, each consisting of a different relation and a word pair that was highly typical of the relation. These relations were evenly divided among four relation categories from Bejar et al. (1991): *class inclusion*, *part-whole*, *case relation*, and *cause-purpose*. Since each statement was used to generate two problems (differing in which word was omitted), I acquired responses to 32 problems in total. I separated these problems into two 16-problem lists, counterbalanced and presented in randomized orders across participants. Each list consisted of a single problem generated from each statement. Study procedure and analyses were pre-registered on AsPredicted (#84748).

Participants. Participants were 100 MTurk workers (M_{age} = 39.06, SD_{age} = 9.19; 45 female, 55 male) who completed our tasks online for payment of $2. The study was approved by the Institutional Review Board at UCLA. Participants had a minimum education level of a U.S. high school graduate, and were sampled from the following English-speaking countries: Australia, Canada, Ireland, New Zealand, South Africa, the United Kingdom, and the United States. We excluded data from 2 participants who reported having trouble paying attention while completing the study, as well as 2 other participants who provided nonsensical responses. Since each participant completed 16 out of the total 32 problems, roughly 50 participants provided responses for each problem.

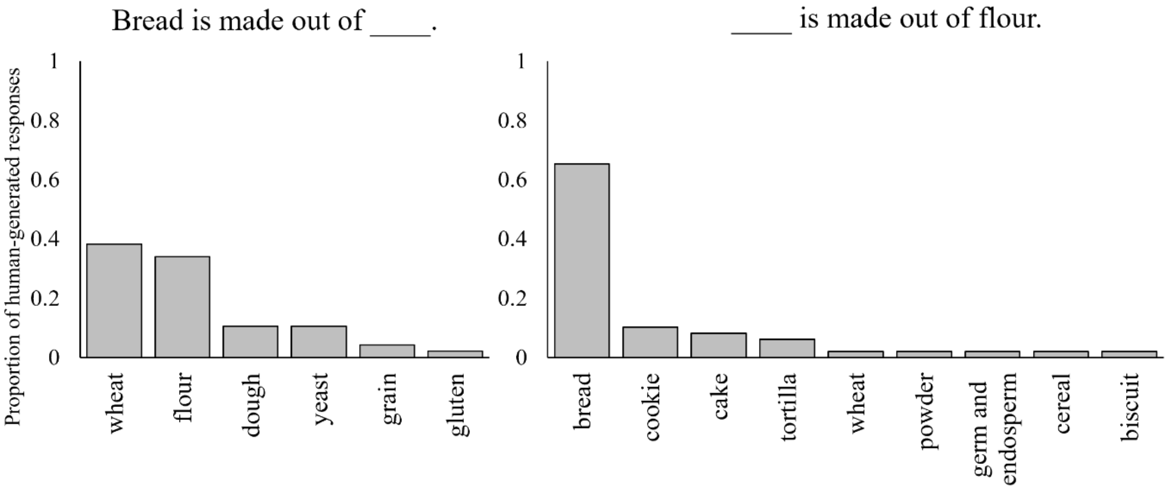


Figure 22: Proportion of human-generated responses to two sentence completion problems, constructed from the same statement. These statements are based on the word pair bread:flour and the relation X is made out of Y.

Results and discussion. Across problems, participants generated a variety of responses, which were largely sensible. Figure 22 shows the proportions of human-generated responses for two sentence completion problems constructed out of the same statement. The most frequent human responses matched the ‘correct’ response included in the Jurgens et al. (2012) norms for 24 out of the 32 problems. In the following, I will detail how I evaluated BART-Gen and BERT’s performance on these problems.

For each problem, BART-Gen generated a Word2vec word embedding, given a query word and a relation (e.g., *robin, X is a kind of Y*). In order to assess the model, I generated a ranking of words, ordered according to the extent to which each word was semantically associated (i.e., from smaller cosine distance to larger cosine distance) with BART-Gen’s generated embedding. This ranking ordered a pool of ~1,500 words constructed out of all words that more than one person generated in response to analogy problems in the present dataset, as well as in the dataset used in Simulations 2a and 2b. Requiring that more than one person generated a given completion is a standard approach to pre-processing generation data, which helps control the quality of responses (Nelson et al., 2004; Peterson et al., 2020). I adopted a similar approach to evaluate BERT, but instead of ranking responses according to their semantic association with some model-generated embedding, I plugged each word of the ~1,500-word response pool in as a completion for each problem and ranked responses according to the model-generated logistic probabilities associated with each response. These values reflect the probability that the language model would generate it as a completion in the masked-language modeling task on which it was trained.

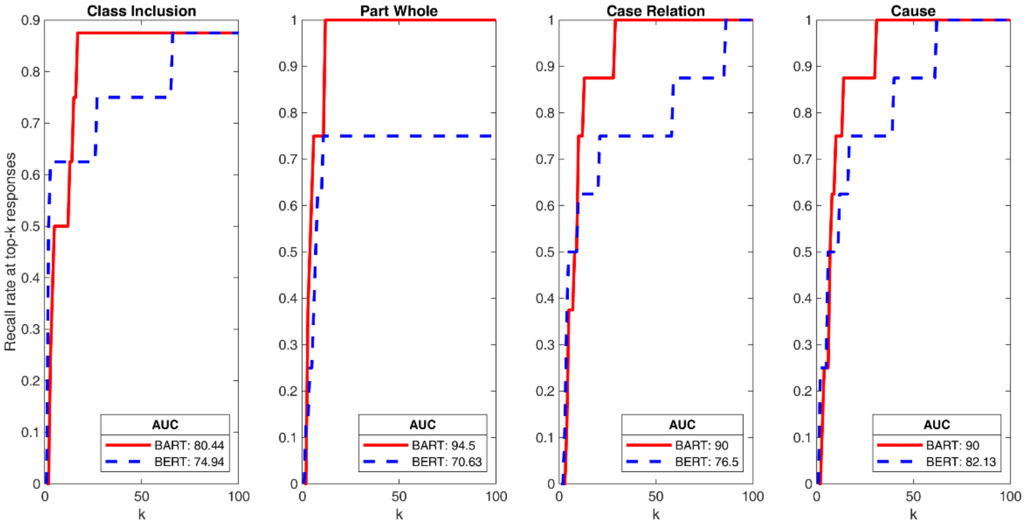


Figure 23: Results from Simulation 1a with generative relation problems (e.g., robin is a type of ?), showing median ranks for the most frequent human-generated response, among all human-generated responses across the task (lower ranks indicate better performance).

Given a ranking for each problem across BART-Gen and BERT, I computed the proportion of problems for which at least one of the most frequent human-generated responses (i.e., those generated by at least 10% of human participants for a given problem) was ranked lower than *k* among all human-generated responses, for *k* = [1,100]. I used these proportions to generate the recall accuracy curve depicted in Figure 24, which plots the proportions mentioned above, as a function of *k*. I quantify model performance as the area under the recall accuracy curve (AUC), where higher values indicate better performance. As shown in Figure 23, BART-Gen (C*lass Inclusion* AUC = 82.43%; *Part Whole* AUC = 94.50%; *Case Relation* AUC = 90.00%; and *Cause* AUC = 90.00 %) outperformed BERT (C*lass Inclusion* AUC = 74.94%; *Part Whole* AUC =

70.63%; *Case Relation* AUC = 76.50%; and *Cause* AUC = 82.13%) across all relation categories. That BART-Gen outperforms BERT, a non-relational large language model, on the masked-language modeling task on which BERT was trained, indicates that BART-Gen has considerable promise as a model of human generative inference. This result also speaks in favor of BART-Gen's usage of explicit relation representations over BERT's non-relational, associative approach to generative inference. In the following simulation, I expanded the test dataset by useing the entire Jurgens norms dataset to further probe the contribution of relation representations in BART-Gen to its success at reproducing humanlike generative inference.

Simulation 1b: Contribution of relations in Jurgens et al. (2012)

The Jurgens norms dataset consists of over 3,000 word pairs instantiating one of 79 semantic relations, organized in to 10 relation types, according to taxonomy developed by Bejar et al. (1991). I used the same procedure for generating problems in Simulation 1A to produce the input to BART-Gen. I took the 20 most typical word pairs for each of those 79 semantic relations, yielding 1,580 statements in total. Each statement yielded two relation completion problems, which omitted either the first word in its word pair (e.g., *bird*) or the second word (e.g., *robin*), yielding 3,160 of these problems with which to evaluate BART-Gen. As mentioned above, solving each of these problems involved generating the omitted word based on the given, query word and the pre-specified relation.

For all problems, BART-Gen produced an embedding to complete the partially-filled relation it was provided in its input. In order to characterize BART-Gen's contribution to accurate inference (as defined in Jurgens et al. (2012)), I computed the semantic distance between the nearest word to the model-generated embedding and the word embedding for the correct answer (e.g., *robin*), and I compared that to the semantic distance between the word embedding for the query word (e.g., *bird*) and that for the correct answer. Unsurprisingly, the model-generated word was much closer to the correct answer than was the query word, and Figure 22 shows this difference (i.e., BART-Gen's contribution index) for all 79 semantic relations, broken down according to 10 relation types given by Bejar et al. (1991).

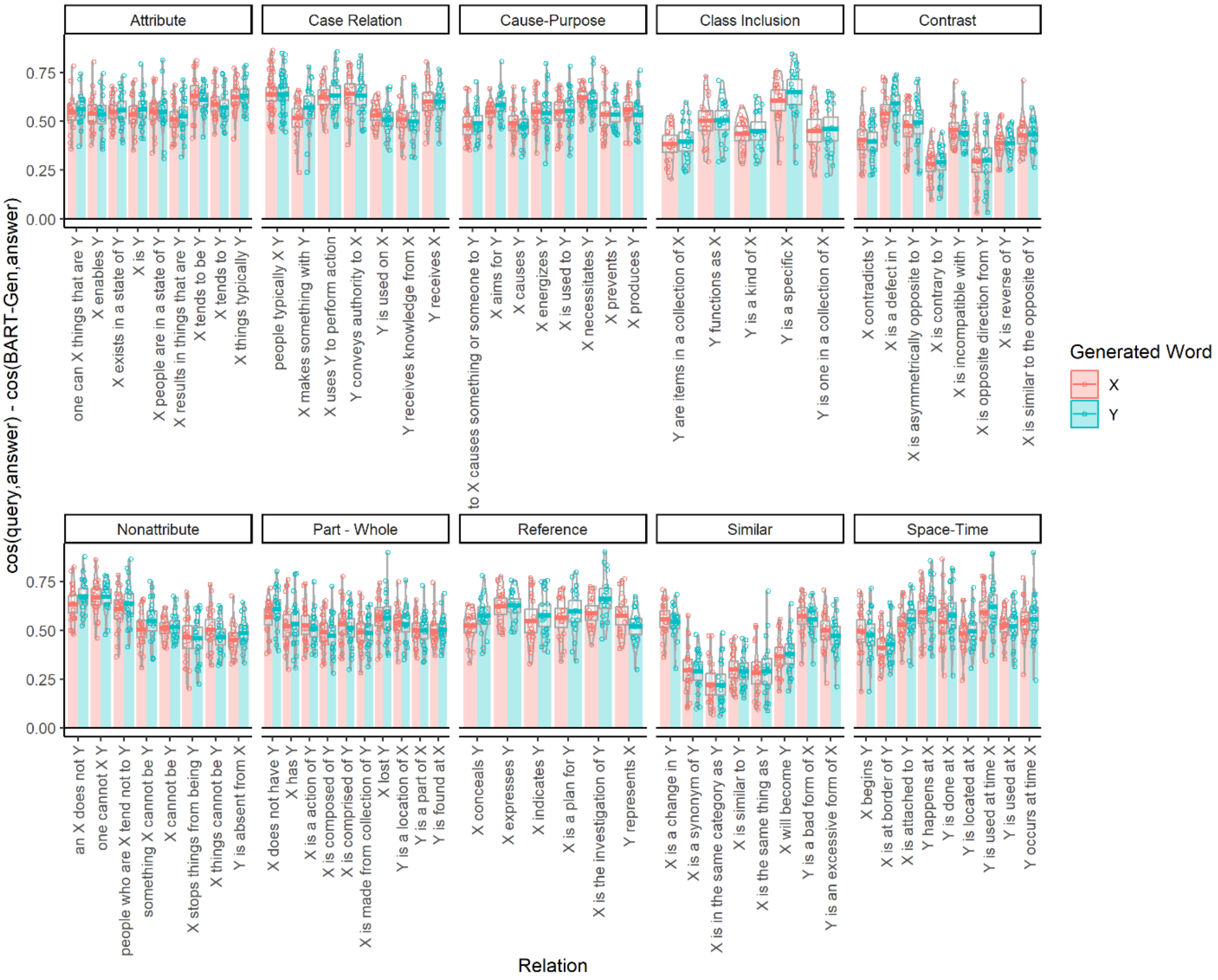


Figure 24: BART-Gen's contribution index to accurate inference in Jurgens et al. (2012).
This contribution was computed as the difference between the semantic distance between the nearest word to the model-generated embedding and the word embedding for the correct answer (e.g., *robin*), and the semantic distance between the word embedding for the query word (e.g., *bird*) and that for the correct answer. Results are broken down according to relation (X-axis) and relation type (each panel), as defined in Bejar et al. (1991). Red bars reflect generation of X within each relation description, and blue bars reflect generation Y.

Notably, BART-Gen's contribution was more modest for relations between relata that are already highly associated (i.e., *class inclusion*, *contrast*, and *similar*). For these relations, semantic association already provides a reasoner with a good basis for inference, and there is less of an opportunity for explicit relations to improve inference. This apparent trend is consistent with the conjecture that inference over explicit relations promotes generalization beyond semantic association. If this conjecture is true, BART-Gen's contribution to inference should be more pronounced for relations linking relata that are not already highly associated with one another.

This first set of simulations provides an initial test of BART-Gen's ability to perform generative inference. Specifically, I showed that BART-Gen was much more successful at reproducing human-like completions of relation-based sentences than was a large-language model BERT that was trained on that very task. Next, I compared the BART-Gen to a baseline rooted simply in semantic association with the query word, as opposed to the explicit relation representations used by BART-Gen. Overall, BART-Gen generated completions that were much closer than the baseline to human-generated completions from Jurgens et al. (2012) dataset, and as the degree to which BART-Gen improved performance over the baseline model was particularly pronounced when the query word was semantically distant from the word to be generated. These simulations thus suggest not only that BART-Gen shows considerable promise as a model of generative inference, but it also supports the broader theoretical claim that explicit relation representations contribute to inference by enabling a reasoner to go beyond the restrictions of imposed by mere association.

In the next set of simulations, I test BART-Gen's ability to perform generative inference with unspecified relations. In contrast to existing models of analogical inference, BART-Gen is unique in the ability to first educe the relations holding within a source analog, in order to constrain inference. I begin with simulations of four-term analogy problems in Simulations 2a and 2b and then move on to more complex analogy problems in Simulation 3.

Stimulation 2: Inference with unspecified relations in four-term analogies

In the next pair of simulations, I shift focus from inference based on pre-specified relations to solving analogy problems based on unspecified relations. I operationalize this inference as the ability to generate a word *D* (e.g., *bird*) that, when linked to a given word *C* (e.g., *robin*), is most analogous to another pair of words *A* (e.g., *sedan*) and *B* (e.g., *car*).

Table 7: Examples of Jurgens analogy problems. A:B pairs are listed in the leftmost column, and C-terms are listed in the middle column. Most frequent human responses, along with the percentage of participants generating that response are in the rightmost column.

A:B	*C*	Most frequent response
famine:plentitude	*novice*	*expert* (55.17%)
joke:laughter	*exercise*	*sweat* (40.62%)
pardon:sin	*brush*	*hair* (36.00%)
simmer:boil	*giggle*	*laugh* (64.52%)

Simulation 2a: Jurgens et al. (2012) analogies

In Simulation 2a, I compare BART-Gen and BERT, as well as the Word2vec parallelogram model on analogy problems generated from the same Jurgens et al. (2012) dataset that I used to evaluate the model in Simulation 1. Recall that this dataset consists of over 3,000 word pairs instantiating one of 79 semantic relations. I use a dataset of 588 four-term analogy problems and human responses to these problems reported in Peterson et al. (2020), and examples of these problems are shown in Table 7. Each problem (e.g., *famine:plentitude :: exercise:?*) was constructed by combining two word-pairs that were each generated as examples of the same semantic relation in Jurgens et al. (2012) (e.g., *famine:plentitude* and *exercise:fitness* both exemplify *X causes Y*) and then dropping the second term from the second word pair (e.g., *fitness*). These problems only use a subset of the word pairs constituting the Jurgens et al. (2012) dataset.

Both BART-Gen and the Word2vec parallelogram model generate an embedding that constitutes their response to a given completion problem, whereas BERT generates a logistic probability associated with a given completion. I adopt the same approach to model evaluation in

Simulation 2a and 2b as I did in Simulation 1a. I ranked model-generated responses, using semantic association between model-generated embeddings for BART-Gen and the Word2vec parallelogram model, and using logistic probabilities for BERT. Importantly, for BERT, I used two different text input styles. The first input style instantiated analogy problems in a natural language sentence, much like the input using in Simulation 1a and in its masked-language modeling training task (e.g., “Famine is related to plentitude, just as novice is related to [MASK].”). The second input style adopted traditional four-term analogy notation (e.g. “famine : plentitude :: novice : [MASK]”). I refer to BERT performance from these input styles as “BERT sentence” and “BERT analogy”, respectively.

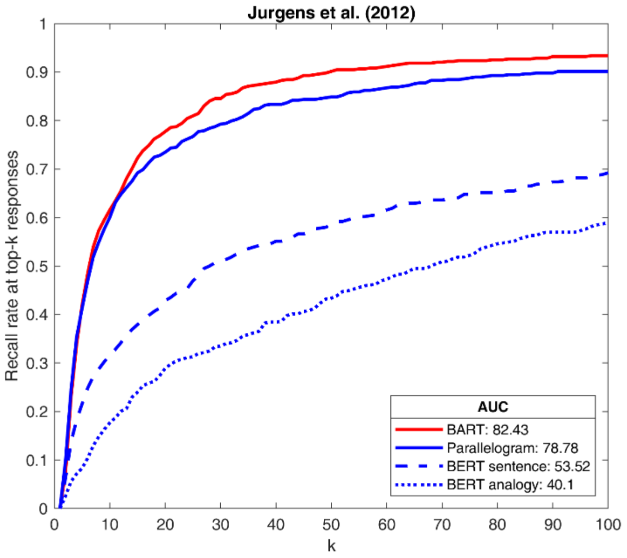


Figure 25: Recall rate as a function of top-k ranked model responses including the most frequent human responses. Recall rate is defined as the proportion of problems where the most frequent human responses were among top-*k* ranked responses for BART-Gen (red), Word2vec parallelogram model (solid blue), BERT sentence (dashed blue), and BERT analogy (dotted blue) for k = [1,100]. Area under the curve (AUC) quantifies model performance.

Using these rankings, I constructed a recall accuracy curve for each model, which are depicted in Figure 26. These curves plot the proportion of at least one of the frequent human responses (i.e., those generated by at least 10% of human participants) ranked among the top-*k* responses, as a function of *k*. While BART-Gen (AUC = 82.43%) did outperform the non-relational Word2vec parallelogram model (AUC = 78.78%), it did so only slightly. On the other hand, BART-Gen far outperformed both versions of the non-relational language model BERT (BERT sentence AUC = 53.52% and BERT analogy AUC = 40.10%). These results demonstrate the strength of BART-Gen's approach to generative inference, even in the absence of pre-specified relations. However, that it only slightly outperformed the non-relational Word2vec parallelogram model motivates further examination of the advantages that BART-Gen's explicitly relational inference present over non-relational approaches. In the following, I test the hypothesis that relations promote effective inference by enabling generalization beyond what is already permitted by mere association. In order to do so, I examine whether semantic distance between analogs contributed to the extent to which BART-Gen outperformed control models.

Table 8: Examples of Green analogy problems.
Each A:B pair listed in the leftmost column was used in both a near and a far analogy problems. Whereas near analogies had a C-term (second column from the left) that was highly associated with the A-term, far analogies had a C-term (fourth column) that was less associated with the A-term. Most frequent responses, along with the percentage of participants generating that response are in the third and fifth columns for near and far analogies respectively.

	Near analogy		Far analogy	
A:B	*C*	Most frequent response	*C*	Most frequent response
answer:riddle	*solution*	*problem* (73.33%)	*key*	*lock* (56.84%)
blindness:sight	*deafness*	*hearing* (90.32%)	*poverty*	*wealth* (48.28%)
eraser:pencil	*whiteout*	*pen* (56.67%)	*amnesia*	*memory* (58.06%)
nose:scent	*tongue*	*taste* (87.88%)	*antenna*	*signal* (50.00%)

Simulation 2b: Green et al. (2012) analogies

In Simulation 2b, I compare model performance on another four-term analogy problem dataset that explicitly manipulates semantic distance between source and target analogs (Green et al., 2010, 2012). This dataset consists of 80 four-term analogy problems developed by Green et al. (2010) and was adapted for generative analogical inference by Green et al. (2012). Half of these problems consist of *near* analogies, in which the *A*-terms are semantically associated with the *C* terms (e.g., ***answer**:riddle :: **solution**:?*). The other half consists of *far* analogies in which the corresponding terms are semantically distant (e.g., ***answer**:riddle :: **key**:?*), see examples in Table 8. Because this dataset holds constant source analogs across *near* and *far* analogies, it provides a clean test of the effect of semantic distance on model performance, while controlling for any variability due to idiosyncrasies in source analogs. In general, human reasoners have greater difficulty solving far than near problems (Green et al., 2010, 2012). Importantly, this set of problems is based on very specific relations that BART had not acquired during training; hence this dataset constitutes a strong test of generalization for BART's relation representations, as well as a natural basis for evaluating BART-Gen's algorithm for generating relational inferences from any analog. To assess model performance for generative inference, I used human responses to these problems reported in (Peterson et al., 2020), i.e., those generated by at least 10% of human participants for a given problem. The measure is the recall accuracy, as the proportion of problems for which at least one of the most frequent human-generated responses (i.e., those generated by at least 10% of human participants for a given problem) was ranked lower than *k* among, for *k* = [1,100].

Results and discussion. As in Simulation 2a, I evaluated model performance by computing the AUC for the recall accuracy curves constructed out of the proportion of problems for which top-*k* ranked model responses contained one of the most frequent human responses. I assessed the recall accuracy curves for near problems and far problems separately, and these are shown in Figure 27. BART-Gen did not outperform all non-relational models on near analogy problems (BART-Gen AUC = 87.64%; Word2vec parallelogram AUC = 83.43; BERT sentence AUC = 92.91%; BERT analogy AUC = 54.18%). Notably, BERT sentence was more successful than BART-Gen on these problems, and this result highlights the efficacy of a non-relational approach to inference involving the generalization between semantically associated analogs. In contrast, BART-Gen outperformed all non-relational models on far analogy problems (BART-Gen AUC = 74.05%; Word2vec parallelogram AUC = 65.07%; BERT sentence AUC = 49.15%; BERT analogy AUC = 38.86%). This latter result highlights that relation processing, as instantiated in BART-Gen, promotes generalization across analogs that are not already highly associated with one another.

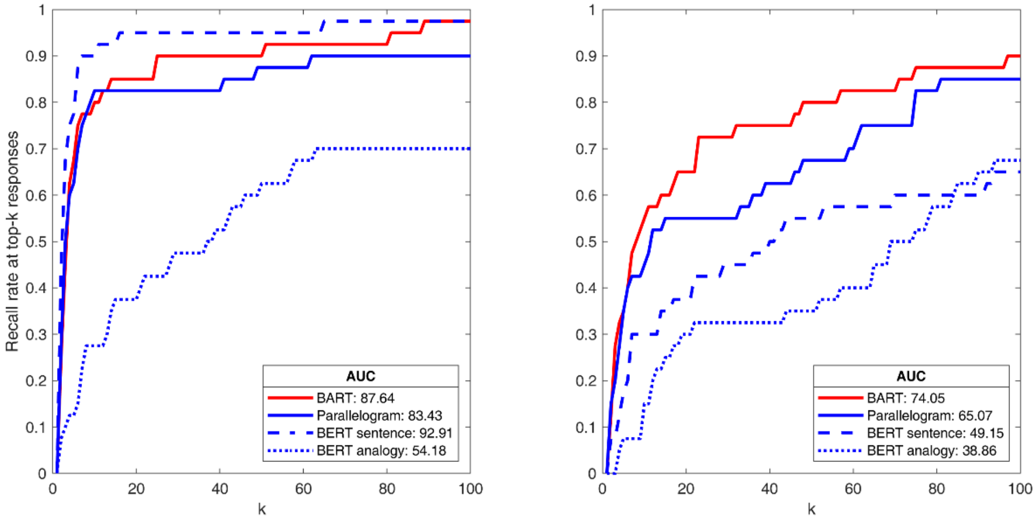


Figure 26: Recall accuracy plotting the proportion of problems where the most frequent human responses were among top-k ranked responses for BART-Gen (red), Word2vec parallelogram model (solid blue), BERT sentence (dashed blue), and BERT analogy (dotted blue) for k = [1,100].
Area under the curve quantifies model performance. Curves are plotted separately for near analogy problems (left panel) and far analogy problems (right panel).

Together, the results of Simulations 2a and 2b for four-term analogy problems provide an initial test of BART-Gen's ability to perform generative inference without specific relations being specified in its input, thus necessitating the eduction of relations constituting the source analog. I showed that BART-Gen was much more successful at reproducing human-like completions of four-term analogy problem than non-relational control models, and more specifically, I showed that BART-Gen's advantage over non-relational models consists in its ability to perform inference across semantically distant analogs. These simulations support BART-Gen's promise as a model of generative inference with unspecified relations. Moreover, with Simulation 1b, Simulation 2b supports the broader theoretical claim that explicit relation representations contribute to inference by enabling a reasoner to go beyond the restrictions imposed by mere association. In a final simulation, Simulation 3, I generalize BART-Gen's ability to perform generative inference with unspecified relations to more naturalistic and complex analogy problems.

Simulation 3: Inference with unspecified relations in extended analogies

Thus far, my test of BART-Gen has used highly-simplified verbal problems. However, empirical demonstrations and computational models of analogical inference typically use experimental materials involving much more elaborate problems than four-term analogies (Burstein, 1983; Carbonell, 1983, 1993; Falkenhainer et al., 1989; Gick & Holyoak, 1980, 1983; Halford et al., 1994; Hofstadter & Mitchell, 1994; Holyoak et al., 1994; Holyoak & Thagard, 1989; Hummel & Holyoak, 2003; Keane & Brayshaw, 1988; Kokinov, 1994; Minervino et al., 2023). In the next set of simulations, I evaluate BART-Gen and examine the impact of its explicit relation representations on similarly elaborate problems, consisting of semantically distant analogies between systems of concepts (Turney, 2008).

In a final set of simulations, I compare BART-Gen and the Word2vec parallelogram model in their ability to generate completions of more elaborate analogies between systems of concepts, which are more akin to the complex experimental materials used by cognitive scientists studying analogical reasoning. Simulation 3 uses stimuli from Turney (2008), which consist of twenty long-form analogies, originally used to compare a model of analogical mapping with human reasoners. Each of these analogies consists of two sets of 5-8 analogous terms, respectively constituting source and target analogs. Ten problems were drawn from scientific analogies described in

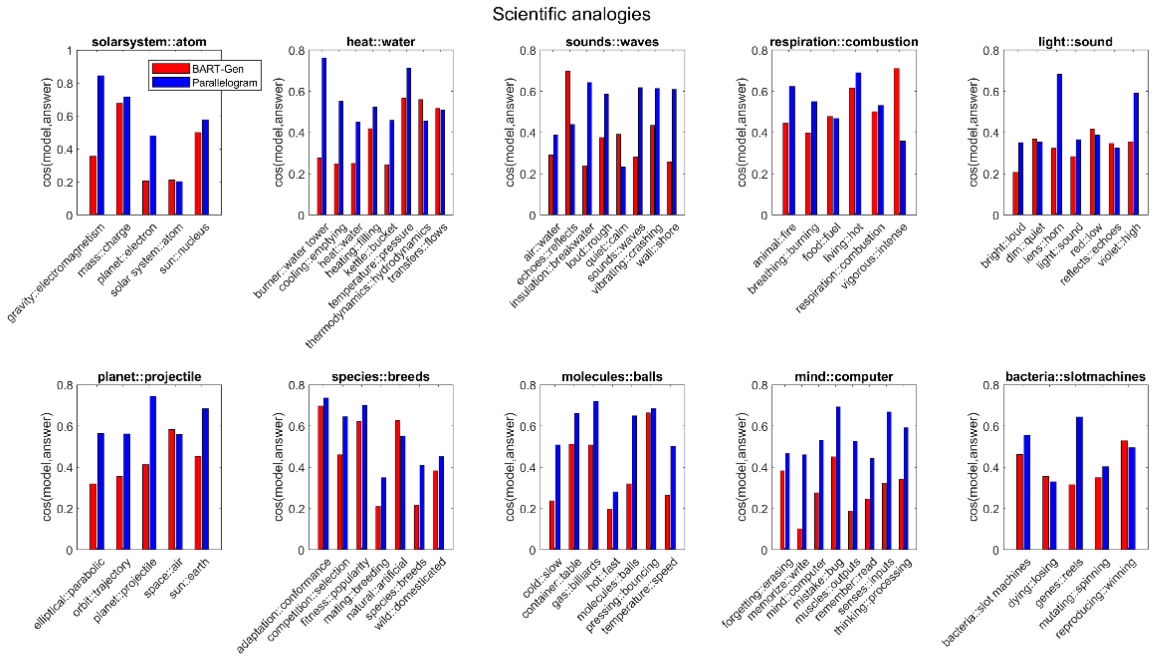
Holyoak and Thagard (1994), and the other ten were drawn from metaphors described in Lakoff and Johnson (2003), and examples of each problem type are shown in Table 9.

Table 9: Examples of Turney extended analogies.

	Source		***Target***	
Type	**Domain**	**Terms**	**Domain**	**Terms**
Science	Solar system	*solar system* *sun* *planet* *mass* *gravity*	Rutherford-Bohr model of the Atom	*atom* *nucleus* *electron* *charge* *electromagnetism*
	Mind	*mind* *thinking* *forgetting* *memorize* *remember* *memory* *muscles* *senses* *mistake*	Computer	*computer* *processing* *erasing* *write* *read* *memory* *outputs* *inputs* *bug*
Metaphor	Seeds	*seeds* *planted* *fruitful* *fruit* *grow* *wither* *blossom*	Ideas	*ideas* *inspired* *productive* *product* *develop* *fail* *succeed*
	Grounds for a building	*foundations* *buildings* *supporting* *solid* *weak* *crack*	Reason for a theory	*reasons* *theories* *confirming* *rational* *dubious* *flaw*

To construct generative analogy problems from these analogies, I iteratively omitted each term from the target and had each model generate an embedding to fill in that omitted term. Thus, an analogy consisting of five terms constituting each of the source and target yielded five unique problems, one omitting each of the terms and for which each model generated an embedding for the omitted term.

In order to compare model performance, I adopted a similar approach as in Simulation 1b, where I computed the distance between each model's generated embedding and the correct answer as stipulated in Turney's (2008) materials. Across all problems, a Wilcoxon rank-sum test showed that BART-Gen's model-generated embedding was closer to the correct answer than that generated by the Word2vec parallelogram model (*W* = 26186; *p* > .001). Figure 28 shows the cosine distance between each model-generated embedding and the correct answer, for each individual problem.



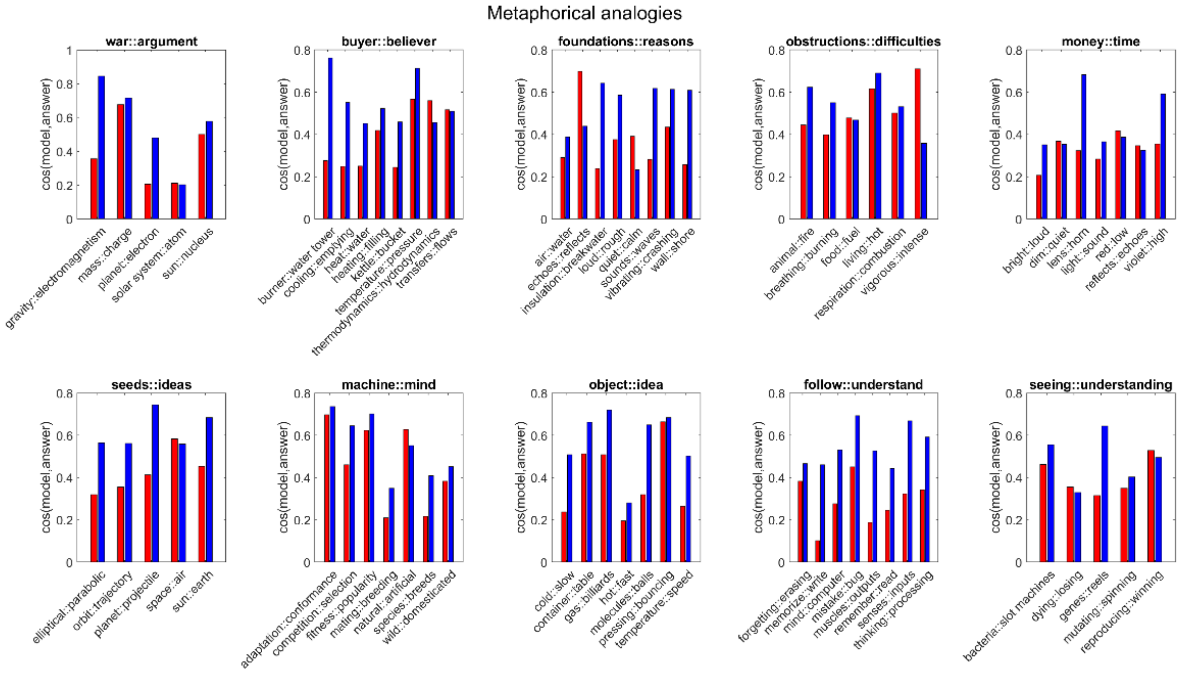


Figure 27: Distance between model-generated embeddings and the correct answer for each analogy problem, as stipulated in Turney's (2008) materials for BART-Gen (red) and Word2vec parallelogram (blue). Top panel shows performance on scientific analogies from Holyoak and Thagard (1994) and bottom panel on metaphorical analogies from Lakoff and Johnson (2003). Lower distance indicates better performance.

Next, I tested the hypothesis that relation processing benefits inference by increasing a model's ability to generalize across semantically distant analogs. Recall that the results from Simulation 1b and 2b lent support to this hypothesis in sentence-completion problems and four-term analogy problems, respectively. In order to assess this claim for complex analogy problems, I computed the difference in the distance from BART-Gen's generated embedding to the correct answer and that from the Word2vec parallelogram model's generated embedding to the correct answer. Positive values thus reflect the extent to which BART-Gen approximated the correct answer better than the Word2vec parallelogram model. I then computed the spearman rank correlation between this difference value and the semantic distance between the correct answer and its corresponding term in the source (e.g., for the generative analogy problem in which the task was to generate the term *charge*, given the source domain *solar* system and the target domain *atom*, I computed the distance between *charge* and its corresponding term in the source *mass*).

As shown in Figure 29, BART-Gen outperformed the parallelogram model on most problems (as indicated by the positive difference of cosine distances for most data points), and the extent to which it did correlated well with the semantic distance between the source terms and correct answers ($\rho = .38$, $p < .001$). This result provides further evidence that BART-Gen's explicit relation representations promote far generalization across analogs.

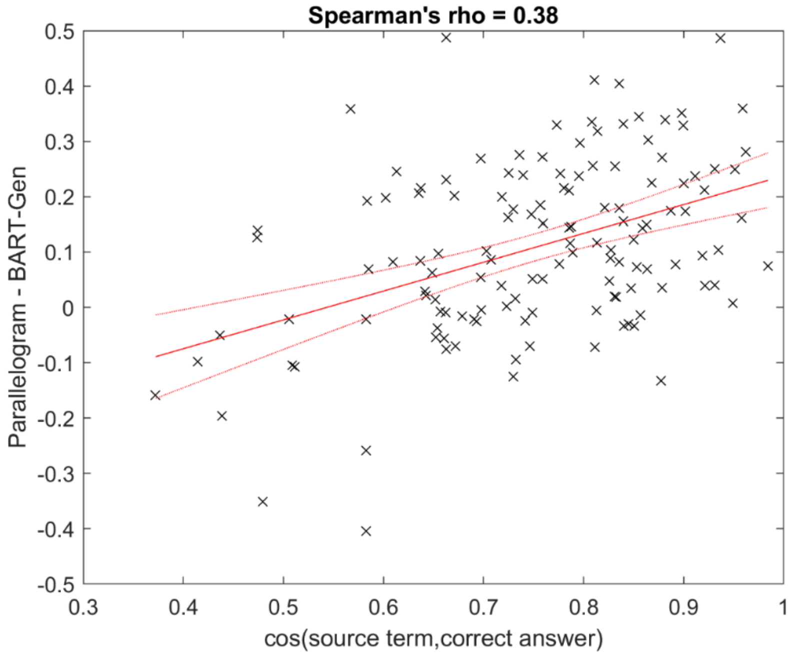


Figure 28: Difference in the semantic distance between BART-Gen's generated embedding and the correct answer and the distance between the parallelogram model's generated embedding and the correct answer, as a function of the semantic distance between source and terms.

General Discussion and Conclusion

In the present chapter, I introduced BART-Gen, a new model capable of two related forms of generative inference: reasoning from pre-specified relations, and reasoning from unspecified relations. In the first form, a reasoner completes a partially-specified instance of a stated relation (e.g., *robin is a type of* ____), and in the second, a reasoner first educes the relation holding among some source analog, copies that over to corresponding elements in a target analog, and generates a completion of the unfilled relation (e.g., *sedan:car :: robin:____*).

Taken together, the results from all simulations support the explanatory power of BART-Gen as a model of human generative inference. Notably, the model can operate with relations pre-specified in its input (Simulations 1a and 1b), as existing models can, but it can also operate even in the absence of such pre-specified input (Simulations 2a, 2b, and 3), unlike existing models. Finally, by comparing BART-Gen's performance with non-relational models, BERT (Simulations 1a, 2a, and 2b) and the Word2vec parallelogram model (Simulations 2a, 2b, and 3), I showed that BART-Gen's advantage over these models tended to be most prominent when relata within (Simulation 1b) and across analogs (Simulation 2b and 3) were semantically distant. These results support the hypothesis that relations promote generalization beyond the restrictions posed by mere association. Across three simulations comparing BART-Gen, which operates on explicit representations of relations learned from non-relational inputs (word embeddings produced by Word2vec), with non-relational baseline models, we showed that BART-Gen's relation representations helped the model to generalize across semantically distance analogs.

General Discussion

In this dissertation, I adopted an approach to examining the presence of relation representation that takes into consideration the cognitive demands of relation processing and the availability of cognitively cheaper non-relational alternatives. In my first chapter I showed that, in contrast to previous claims that relations subserve all comparison judgments (Gentner & Markman, 1994; Markman, 1996; Markman & Gentner, 1993b, 1993a), the processing demand involved in making comparisons assessing difference discourage the use of relations, whereas relation processing is preserved in comparisons assessing similarity. In my second chapter, I went on to show that while non-relational lexical representations could, in principle, explain a phenomenon, previously attributed to the usage of relation representations in recognition memory (Popov et al., 2017), such representations are more likely to generate such a phenomenon than are lexical representations. Finally, in my third chapter, I introduced a computational model of generative inference, a cognitive process that integrates human reasoning and memory (respectively emphasized separated in Chapters 1 and 2), and showed with a series of simulations that the explicit representation of relations promoted inference across semantically distant analogs.

Open issues

While human relation processing has been extensively studied, it is a broad area of research spanning many subdisciplines of cognitive science. The findings from each of my chapters open new issues and raise further questions about human relation processing. In Chapter 1, I argued for a representational asymmetry between the relations *same* and *different*, that *same* is represented accordingly but *different* is represented as *not-same* making the latter more relationally complex and thus more demanding to process (Andrews & Halford, 2002; Halford et al., 1998). This hypothesis explains two phenomena that I demonstrated in Chapter 1: That people have more difficulty assessing relational difference than relational similarity, and that people tend to incorporate more relational information when judging similarity than when judging difference. If same and different maintain this asymmetric relation, how might other, oppositely-defined relations that often participate in everyday human thinking, like, for instance, *truth* and *falsity*. Such an asymmetric relation between *truth* and *falsity* (i.e., that *true* is represented accordingly but that *false* is represented as *not-true*) might predict, perhaps, that sentence verification judgments might tend to be more accurate or faster than sentence falsification judgments. The existence of more such asymmetries in relation processing further lend credence to the impact of one property of a language of thought (Fodor, 1979; Quilty-Dunn et al., 2022): the usage of representations in human cognition that are sensitive to logical operators (e.g., *not*, which instantiates *negation*).

Moving from the realm of reasoning to that of memory, although Chapter 2 ultimately lent credence to the claim that relation representations impact episodic memory, it also raised an important methodological point. That a model operating only on non-relational lexical representations could reproduce a phenomenon that was assumed to necessitate relation processing raises the possibility that other putatively relational phenomena may also be reproduced in a computational model operating over non-relational representations. For instance, one source of evidence for the usage of relation representations in human cognition is the phenomenon known as ‘relational priming’ (Estes & Jones, 2006; Popov & Hristova, 2015; Spellman et al., 2001). If some conceptual content is represented, then processing instances of that content should prime subsequent events in which that same or similar content is processed. Spellman and colleagues (2001) were the first to show that processing an instance of some relation (e.g., *bear:cave*) do indeed prime subsequent events in which that same or similar relations (e.g., *bird:nest*) are processed. However, to my knowledge, this relational priming effect has not be reproduced in a computational model. It thus remains to be seen if this effect could, in principle, be reproduced without positing relation representations through, perhaps, lexical representations that encode typical relational roles (e.g., *the superordinate category* in the *X is a kind of Y* relation) among their semantic features.

The simulations in Chapter 3 are similar to the results in Chapter 2 in that although they ultimately provided further evidence for relation processing, the strong performance of non-relational models reveals the surprisingly large expressive capacity of purely associative representations that lack any explicit relational structure. For example, analogy problems between semantically associated analogs might well be solvable without any explicit relation processing. This would explain why generating solutions to analogy problems featuring semantically distant but not semantically associated analogs induces relation processing in subsequent tasks (Vendetti et al., 2014). Generating completions only of semantically distant but not proximal analogies is likely to induce relation processing, since the latter could be solved using computationally cheaper approaches involving non-relational representations. Moreover, Chapter 1 showed that people modulate their relation processing based on the complexity of the reasoning task: Since *difference* judgments are more complex than *similarity* judgments, they elicit *less* relation processing. How do people actively do this? Given that some tasks are solvable through processes that do not incorporate relation processing (e.g., comparisons, semantically near analogy problems), what explains whether a human reasoner come to adopt either kind of approach in a given situation?

Beyond these questions raised by each individual chapter of my dissertation, they collectively prompt further questions: With the exception of the visual comparison task in Chapter 1, the studies performed in this dissertation make exclusive use of verbal materials and operationalize relation processing as processing semantic relations holding between lexical concepts. Do some of the same phenomena continue to hold when relation processing is instead instantiated with visuospatial relations? For instance, do representations of visually-presented relational structures also impact episodic memory such that human reasoners might feel false familiarity of a novel instance of a previously processed relation?

Another issue, raised only indirectly throughout my dissertation concerns relation learning. All of the modeling work in this dissertation centered around BART, a model of relation learning and acquisition (Lu et al., 2012, 2019). The field of relation learning is divided on the mechanisms that constitute relation learning. An earlier view instantiated in another computational model of relation learning, *Discovery of Relations by Analogy* (DORA; Doumas et al., 2008; see also Chen et al., 2019 for another such model), proposes that relation learning presupposes the ability to recognize analogies between exemplars of a given relation. According to this view, relation learning consists in a process of structural alignment and *intersection discovery* in which relationally similar analogs are compared, and an abstract representation learned from the comparison consists in the shared relational structure across analogs (Forbus et al., 2017; Gentner, 1983; Hummel & Holyoak, 2003). This view correctly predicts that learning probabilistic relational categories is very difficult (Kittur et al., 2004, 2006) and only possible if the relational structure is somehow re-represented so as to make the category deterministic at that re-represented level of abstraction (Jung & Hummel, 2015b, 2015a).

However, this view also predicts that supervised learning of relational concepts should always benefit from comparing exemplars of the same relational category rather than those of different relational categories, since only within-category comparisons enable this *intersection discovery* learning process in a clear way. However, Corral and colleagues (2018) showed that human learners were more effective and quicker to learn a number of relational categories governed by semantic relations, visuospatial relations, even functional causal models from *between-category* comparisons, which violates the prediction made by an *intersection discovery* learning process. This observation led Corral et al. (2018) to propose an alternative approach to human relation induction that relies on feature-based processing.

Past work on the effects of exemplar sequencing during supervised learning has clarified the benefits of between-category comparisons on feature-based category learning. Kornell and Bjork (2008) asked participants to learn individual artists' painting styles from labeled exemplars and showed that participants learned much more effectively when exemplar sequencing was 'interleaved', with consecutively-presented exemplars belonging to different categories, than when they were 'massed', with consecutively-presented exemplars belonging to the same category. Whereas the interleaved sequencing encourages between-category comparisons, the massed sequencing encourages within-category comparisons (Goldstone, 1996). Interestingly, the literature on exemplar sequencing has produced mixed results, with interleaved sequencing working better some times and with massed sequencing working better at other times (e.g., Carpenter & Mueller, 2013; Kornell & Bjork, 2008; K. H. Kurtz & Hovland, 1956; Vlach et al., 2008), and category similarity structure appears to be a major determinant of the relative efficacy of one sequencing type over the other (Brunmair & Richter, 2019): Interleaved sequencing is more likely to benefit learning when exemplars belonging to different categories are highly similar to each other because the between-category comparisons that it emphasizes provide more opportunities to discriminate between categories. On the other hand, massed sequencing is more effective for learning categories whose exemplars are fairly dissimilar because the within-category comparisons that it encourages provide more opportunities to learn what unifies to-be-learned categories (Carvalho & Goldstone, 2015, 2017). Does the feature-based route to relational category learning follow-suit? Does relational similarity interact with comparison type in the same way that featural similarity does for feature-based category learning?

BART, as well as a number of other computational models of relation induction, instantiates and adds precision to Corral and colleagues' (2018) proposal for a feature-based approach relational category learning (Davidson & Lake, 2021; Geiger et al., 2023; Lu et al., 2012, 2019; Shanahan et al., 2019; Thibodeau et al., 2013). Instead of relying on analogical comparison, this approach consists of simpler statistical learning mechanisms performed over feature-based representations of relata (e.g., an unstructured feature vector) rather than analogical mapping performed on structured representations. For example, instead of representing a relation as a structured binding between two roles X and Y, as in *X is larger than Y*, one could represent the same content as a global feature of a set of relata, represented as a unified entity Z, as in *Z is lopsided*. Here, representations of relations are recoverable either as a unitary feature or as a distribution of features, and relations are (or are reducible) to features that are formally equivalent to other stimulus properties like size, shape, or color. If human learners adopt both of these approaches, what are their relative merits in the representations they produce? When do human reasoners use one over the other? There are certainly more unresolved issues in human relation processing and representation than those outlined above. As this is an active area of research featuring continual development of computational models and behavioral and neural experiments that test the assumptions of these models, future research will likely shed light on these issues, while also revealing further issues to investigate.

Across all three chapters, the conclusion of my dissertation favors relation representations as a core resource in human cognition. However, the contribution of my dissertation is not restricted to this conclusion, but extends to an instantiation of an approach to studying human cognition that emphasizes not only the expressive advantages of sophisticated cognitive processing (Fodor, 1979; Quilty-Dunn et al., 2022), but also the limitations of human cognition (Griffiths, 2020).

References

Ambridge, B. (2020). Against stored abstractions: A radical exemplar model of language acquisition. *First Language, 40*, 509–559. https://doi.org/10.1177/0142723719869731

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*, 409–429. https://doi.org/10.1037/0033-295X.98.3.409

Andrews, G., & Halford, G. S. (2002). A cognitive complexity metric applied to cognitive development. *Cognitive Psychology*, *45*(2), 153–219. https://doi.org/10.1016/s0010-0285(02)00002-6

Armstrong, D. M. (1978). *A Theory of Universals: Volume 2: Universals and Scientific Realism*. CUP Archive.

Arthur, W., Tubre, T. C., Paul, D. S., & Sanchez-Ku, M. L. (1999). College-sample psychometric and normative data on a short form of the Raven Advanced Progressive Matrices Test. *Journal of Psychoeducational Assessment, 17*(4), 354–361. https://doi.org/10.1177/073428299901700405

Bapst, V., Sanchez-Gonzalez, A., Doersch, C., Stachenfeld, K. L., Kohli, P., Battaglia, P. W., & Hamrick, J. B. (2019). *Structured agents for physical construction* (arXiv:1904.03177; Version 2). arXiv. https://doi.org/10.48550/arXiv.1904.03177

Bassok, M., & Medin, D. L. (1997). Birds of a feather flock together: Similarity judgments with semantically rich stimuli. *Journal of Memory and Language*, *36*(3), 311–336. https://doi.org/10.1006/jmla.1996.2492

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67,* 1–48. https://doi.org/10.18637/jss.v067.i01

Battaglia, P., Pascanu, R., Lai, M., & Rezende, D. J. (2016). *Interaction networks for learning about objects, relations and physics*. 9.

Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., Gulcehre, C., Song, F., Ballard, A., Gilmer, J., Dahl, G., Vaswani, A., Allen, K., Nash, C., Langston, V., … Pascanu, R. (2018). Relational inductive biases, deep learning, and graph networks. *ArXiv:1806.01261 [Cs, Stat]*. http://arxiv.org/abs/1806.01261

Bejar, I. I., Chaffin, R., & Embretson, S. (1991). *Cognitive and Psychometric Analysis of Analogical Problem Solving*. Springer US. https://doi.org/10.1007/978-1-4613-9690-1

Bhatia, S., & Aka, A. (2022). Cognitive modeling with representations from large-scale digital data. *Current Directions in Psychological Science*, *31*(3), 207–214. https://doi.org/10.1177/09637214211068113

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, *94*(2), 115–147. https://doi.org/10.1037/0033-295X.94.2.115

Brunmair, M., & Richter, T. (2019). Similarity matters: A meta-analysis of interleaved learning and its moderators. *Psychological Bulletin*, *145*(11), 1029–1052. https://doi.org/10.1037/bul0000209

Brysbaert, M., Mandera, P., McCormick, S. F., & Keuleers, E. (2019). Word prevalence norms for 62,000 English lemmas. *Behavior Research Methods*, *51*(2), 467–479. https://doi.org/10.3758/s13428-018-1077-9

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911. https://doi.org/10.3758/s13428-013-0403-5

Bunge, S. A., Wendelken, C., Badre, D., & Wagner, A. D. (2005). Analogical reasoning and prefrontal cortex: Evidence for separable retrieval and integration mechanisms. *Cerebral Cortex (New York, N.Y.: 1991)*, *15*(3), 239–249. https://doi.org/10.1093/cercor/bhh126

Burstein, M. (1983, August 22). *A Model of Learning by Incremental Analogical Reasoning and Debugging*. AAAI Conference on Artificial Intelligence. https://www.semanticscholar.org/paper/A-Model-of-Learning-by-Incremental-Analogical-and-Burstein/989f715c3c851c90a897f0dd06f25def1caab1c6

Carbonell, J. G. (1983). Learning by analogy: Formulating and generalizing plans from past experience. In R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.), *Machine Learning* (pp. 137–161). Morgan Kaufmann. https://doi.org/10.1016/B978-0-08-051054-5.50009-1

Carbonell, J. G. (1993). Derivational analogy: A theory of reconstructive problem solving and expertise acquisition. In *Readings in knowledge acquisition and learning: Automating the construction and improvement of expert systems* (pp. 727–738). Morgan Kaufmann Publishers Inc.

Carey, S. (2011). *The Origin of Concepts* (1. iss. Oxford Univ. paperback). Oxford Univ. Press.

Carpenter, P. A., & Just, M. A. (1975). Sentence comprehension: A psycholinguistic processing model of verification. *Psychological Review*, *82*, 45–73. https://doi.org/10.1037/h0076248

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review*, *97*(3), 404–431. https://doi.org/10.1037/0033-295X.97.3.404

Carpenter, S. K., & Mueller, F. E. (2013). The effects of interleaving versus blocking on foreign language pronunciation learning. *Memory & Cognition*, *41*(5), 671–682. https://doi.org/10.3758/s13421-012-0291-4

Carvalho, P. F., & Goldstone, R. L. (2015). The benefits of interleaved and blocked study: Different tasks benefit from different schedules of study. *Psychonomic Bulletin & Review*, *22*(1), 281–288. https://doi.org/10.3758/s13423-014-0676-4

Carvalho, P. F., & Goldstone, R. L. (2017). The sequence of study changes what information is attended to, encoded, and remembered during category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(11), 1699–1719. https://doi.org/10.1037/xlm0000406

Chaffin, R., & Herrmann, D. J. (1984). The similarity and diversity of semantic relations. *Memory & Cognition*, *12*(2), 134–141. https://doi.org/10.3758/BF03198427

Chaffin, R., & Herrmann, D. J. (1987). Relation element theory: A new account of the representation and processing of semantic relations. In *Memory and learning: The Ebbinghaus Centennial Conference* (pp. 221–245). Lawrence Erlbaum Associates, Inc.

Challis, B. H., & Sidhu, R. (1993). Dissociative effect of massed repetition on implicit and explicit measures of memory. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *19*(1), 115–127. https://doi.org/10.1037//0278-7393.19.1.115

Chalmers, D. J., French, R. M., & Hofstadter, D. R. (1992). High-level perception, representation, and analogy: A critique of artificial intelligence methodology. *Journal of Experimental & Theoretical Artificial Intelligence*, *4*(3), 185–211. https://doi.org/10.1080/09528139208953747

Chen, D., Lu, H., & Holyoak, K. J. (2017). Generative Inferences Based on Learned Relations. *Cognitive Science*, *41*, 1062–1092. https://doi.org/10.1111/cogs.12455

Chen, K., Rabkina, I., McLure, M. D., & Forbus, K. D. (2019). Human-Like Sketch Object Recognition via Analogical Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), Article 01. https://doi.org/10.1609/aaai.v33i01.33011336

Chiang, J. N., Peng, Y., Lu, H., Holyoak, K. J., & Monti, M. M. (2021). Distributed Code for Semantic Relations Predicts Neural Similarity during Analogical Reasoning. *Journal of Cognitive Neuroscience*, *33*(3), 377–389. https://doi.org/10.1162/jocn_a_01620

Christie, S., & Gentner, D. (2010). Where Hypotheses Come From: Learning New Relations by Structural Alignment. *Journal of Cognition and Development*, *11*(3), 356–373. https://doi.org/10.1080/15248371003700015

Clark, H. H. (1976). Semantics and Comprehension. In *Semantics and Comprehension*. Mouton. https://doi.org/10.1515/9783110871029

Clark, H. H., & Chase, W. G. (1972). On the process of comparing sentences against pictures. *Cognitive Psychology*, *3*(3), 472–517. https://doi.org/10.1016/0010-0285(72)90019-9

Corral, D., Kurtz, K. J., & Jones, M. (2018). Learning relational concepts from within- versus between-category comparisons. *Journal of Experimental Psychology. General*, *147*(11), 1571–1596. https://doi.org/10.1037/xge0000517

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*(1), 87–114. https://doi.org/10.1017/S0140525X01003922

Davidson, G., & Lake, B. M. (2021). *Examining Infant Relation Categorization Through Deep Neural Networks* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/esvuw

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Diedenhofen, B., & Musch, J. (2015). cocor: A comprehensive solution for the statistical comparison of correlations. *PloS One*, *10*(3), e0121945. https://doi.org/10.1371/journal.pone.0121945

Doumas, L. A. A., & Hummel, J. E. (2004). A fundamental limitation of symbol-argument-argument notation as a model of human relational representations. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *26*(26).

Doumas, L. A. A., & Hummel, J. E. (2005). Approaches to Modeling Human Mental Representations: What Works, What Doesn't, and Why. In *The Cambridge handbook of thinking and reasoning* (pp. 73–91). Cambridge University Press.

Doumas, L. A. A., Hummel, J. E., & Sandhofer, C. M. (2008). A theory of the discovery and predication of relational concepts. *Psychological Review*, *115*(1), 1–43. https://doi.org/10.1037/0033-295X.115.1.1

Duncker, K. (1945). On problem-solving. *Psychological Monographs*, *58*, i–113. https://doi.org/10.1037/h0093599

Dunn, O. J., & Clark, V. (1969). Correlation Coefficients Measured on the Same Individuals. *Journal of the American Statistical Association*, *64*(325), 366–377. https://doi.org/10.2307/2283746

Edwards, B. J., Williams, J. J., Gentner, D., & Lombrozo, T. (2019). Explanation recruits comparison in a category-learning task. *Cognition*, *185*, 21–38. https://doi.org/10.1016/j.cognition.2018.12.011

Estes, Z., & Hasson, U. (2004). The importance of being nonalignable: A critical test of the structural alignment theory of similarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30,* 1082–1092. https://doi.org/10.1037/0278-7393.30.5.1082

Estes, Z., & Jones, L. L. (2006). Priming via relational similarity: A copper horse is faster when seen through a glass eye. *Journal of Memory and Language*, *55*(1), 89–101. https://doi.org/10.1016/j.jml.2006.01.004

Falkenhainer, B., Forbus, K. D., & Gentner, D. (1989). The structure-mapping engine: Algorithm and examples. *Artificial Intelligence*, *41*(1), 1–63. https://doi.org/10.1016/0004-3702(89)90077-5

Fodor, J. A. (1979). *The Language of thought* (1st paperback printing). Harvard Univ. Press.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1–2), 3–71. https://doi.org/10.1016/0010-0277(88)90031-5

Forbus, K. D., Ferguson, R. W., Lovett, A., & Gentner, D. (2017). Extending SME to Handle Large-Scale Cognitive Modeling. *Cognitive Science*, *41*(5), 1152–1201. https://doi.org/10.1111/cogs.12377

Friel, N., & Wyse, J. (2012). Estimating the evidence – a review. *Statistica Neerlandica*, *66*(3), 288–308. https://doi.org/10.1111/j.1467-9574.2011.00515.x

Geiger, A., Carstensen, A., Frank, M. C., & Potts, C. (2023). Relational reasoning and generalization using nonsymbolic neural networks. *Psychological Review*, *130*(2), 308.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, *7*(2), 155–170. https://doi.org/10.1207/s15516709cog0702_3

Gentner, D. (2002). Analogy in Scientific Discovery: The Case of Johannes Kepler. In L. Magnani & N. J. Nersessian (Eds.), *Model-Based Reasoning: Science, Technology, Values* (pp. 21–39). Springer US. https://doi.org/10.1007/978-1-4615-0605-8_2

Gentner, D., & Forbus, K. D. (2011). Computational models of analogy. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(3), 266–276. https://doi.org/10.1002/wcs.105

Gentner, D., & Kurtz, K. J. (2005). Relational categories. In W. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman, & P. Wolff (Eds.), *Categorization inside and outside the laboratory: Essays in honor of Douglas L. Medin.* (pp. 151–175). American Psychological Association. https://doi.org/10.1037/11156-009

Gentner, D., & Markman, A. B. (1994). Structural alignment in comparison: No difference without similarity. *Psychological Science*, *5*(3), 152–158. https://doi.org/10.1111/j.1467-9280.1994.tb00652.x

Gentner, D., Rattermann, M. J., & Forbus, K. D. (1993). The Roles of Similarity in Transfer: Separating Retrievability From Inferential Soundness. *Cognitive Psychology*, *25*(4), 524–575. https://doi.org/10.1006/cogp.1993.1013

Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology*, *12*(3), 306–355. https://doi.org/10.1016/0010-0285(80)90013-4

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology*, *15*, 1–38. https://doi.org/10.1016/0010-0285(83)90002-6

Goldstone, R. L. (1996). Isolated and interrelated concepts. *Memory & Cognition*, *24*(5), 608–628. https://doi.org/10.3758/BF03201087

Goldstone, R. L., & Medin, D. L. (1994). Time course of comparison. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 29–50. https://doi.org/10.1037/0278-7393.20.1.29

Goldwater, M. B. (2017). Grammatical Constructions as Relational Categories. *Topics in Cognitive Science*, *9*(3), 776–799. https://doi.org/10.1111/tops.12272

Goldwater, M. B., Markman, A. B., & Stilwell, C. H. (2011). The empirical case for role-governed categories. *Cognition*, *118*(3), 359–376. https://doi.org/10.1016/j.cognition.2010.10.009

Goldwater, M. B., & Schalk, L. (2016). Relational categories as a bridge between cognitive and educational research. *Psychological Bulletin*, *142*(7), 729–757. https://doi.org/10.1037/bul0000043

Golonka, S., & Estes, Z. (2009). Thematic relations affect similarity via commonalities. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *35*(6), 1454–1464. https://doi.org/10.1037/a0017397

Green, A. E., Kraemer, D. J., Fugelsang, J. A., Gray, J. R., & Dunbar, K. N. (2012). Neural correlates of creativity in analogical reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(2), 264.

Green, A. E., Kraemer, D. J. M., Fugelsang, J. A., Gray, J. R., & Dunbar, K. N. (2010). Connecting long distance: Semantic distance in analogical reasoning modulates frontopolar cortex activity. *Cerebral Cortex*, *20*(1), 70–76. https://doi.org/10.1093/cercor/bhp081

Griffiths, T. L. (2020). Understanding Human Intelligence through Human Limitations. *Trends in Cognitive Sciences*, *24*(11), 873–883. https://doi.org/10.1016/j.tics.2020.09.001

Günther, F., Rinaldi, L., & Marelli, M. (2019). Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions. *Perspectives on Psychological Science, 14*(6), 1006–1033. https://doi.org/10.1177/1745691619861372

Hafri, A., Gleitman, L. R., Landau, B., & Trueswell, J. C. (2023). Where word and world meet: Language and vision share an abstract representation of symmetry. *Journal of Experimental Psychology. General*, *152*(2), 509–527. https://doi.org/10.1037/xge0001283

Halford, G. S., Wilson, W. H., Guo, J., Gayler, R. W., Wiles, J., & Stewart, J. E. M. (1994). Connectionist implications for processing capacity limitations in analogies. In *Analogical connections* (pp. 363–415). Ablex Publishing.

Halford, G. S., Wilson, W. H., & Phillips, S. (1998). Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. *Behavioral and Brain Sciences*, *21*(6), 803–831. https://doi.org/10.1017/S0140525X98001769

Hasson, U., & Glucksberg, S. (2006). Does understanding negation entail affirmation?: An examination of negated metaphors. *Journal of Pragmatics*, *38*(7), 1015–1032. https://doi.org/10.1016/j.pragma.2005.12.005

Hochmann, J.-R. (2021). Asymmetry in the complexity of same and different representations. *Current Opinion in Behavioral Sciences*, *37*, 133–139. https://doi.org/10.1016/j.cobeha.2020.12.003

Hochmann, J.-R., Carey, S., & Mehler, J. (2018). Infants learn a rule predicated on the relation same but fail to simultaneously learn a rule predicated on the relation different. *Cognition*, *177*, 49–57. https://doi.org/10.1016/j.cognition.2018.04.005

Hochmann, J.-R., Mody, S., & Carey, S. (2016). Infants' representations of same and different in match- and non-match-to-sample. *Cognitive Psychology*, *86*, 87–111. https://doi.org/10.1016/j.cogpsych.2016.01.005

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian Model Averaging: A Tutorial. *Statistical Science*, *14*(4), 382–401.

Hofstadter, D. R., & Mitchell, M. (1994). The Copycat project: A model of mental fluidity and analogy-making. In *Analogical connections* (pp. 31–112). Ablex Publishing.

Holyoak, K. J. (2012). Analogy and relational reasoning. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford Handbook of Thinking and Reasoning* (pp. 234–259). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199734689.001.0001

Holyoak, K. J., Novick, L. R., & Melz, E. R. (1994). Component processes in analogical transfer: Mapping, pattern completion, and adaptation. In *Analogical connections* (pp. 113–180).

Holyoak, K. J., & Thagard, P. (1989). Analogical Mapping by Constraint Satisfaction. *Cognitive Science*, *13*(3), 295–355. https://doi.org/10.1207/s15516709cog1303_1

Holyoak, K. J., & Thagard, P. (1994a). *Mental Leaps: Analogy in Creative Thought*. The MIT Press. https://doi.org/10.7551/mitpress/4549.001.0001

Holyoak, K. J., & Thagard, P. (1994b). *Mental Leaps: Analogy in Creative Thought*. The MIT Press. https://doi.org/10.7551/mitpress/4549.001.0001

Honke, G., & Kurtz, K. J. (2019). Similarity is as similarity does? A critical inquiry into the effect of thematic association on similarity. *Cognition*, *186*, 115–138. https://doi.org/10.1016/j.cognition.2019.01.016

Hoyos, C., & Gentner, D. (2017). Generating explanations via analogical comparison. *Psychonomic Bulletin & Review*, *24*(5), 1364–1374. https://doi.org/10.3758/s13423-017-1289-5

Hoyos, C., Horton, W. S., Simms, N. K., & Gentner, D. (2020). Analogical Comparison Promotes Theory-of-Mind Development. *Cognitive Science*, *44*(9), e12891. https://doi.org/10.1111/cogs.12891

Hummel, J. E. (2010). Symbolic Versus Associative Learning. *Cognitive Science*, *34*(6), 958–965. https://doi.org/10.1111/j.1551-6709.2010.01096.x

Hummel, J. E. (2011). Getting symbols out of a neural architecture. *Connection Science*, *23*(2), 109–118. https://doi.org/10.1080/09540091.2011.569880

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review*, *104*(3), 427–466. https://doi.org/10.1037/0033-295X.104.3.427

Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review*, *110*(2), 220–264. https://doi.org/10.1037/0033-295X.110.2.220

Ichien, N., Lu, H., & Holyoak, K. J. (2022). Predicting patterns of similarity among abstract semantic relations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *48*(1), 108–121. https://doi.org/10.1037/xlm0001010

Jaakkola, T. S., & Jordan, M. I. (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing*, *10*(1), 25–37. https://doi.org/10.1023/A:1008932416310

Jones, M., & Love, B. C. (2007). Beyond common features: The role of roles in determining similarity q. *Cognitive Psychology*, 36.

Jung, W., & Hummel, J. E. (2015a). Making Probabilistic Relational Categories Learnable. *Cognitive Science*, *39*(6), 1259–1291. https://doi.org/10.1111/cogs.12199

Jung, W., & Hummel, J. E. (2015b). Revisiting Wittgenstein's puzzle: Hierarchical encoding and comparison facilitate learning of probabilistic relational categories. *Frontiers in Psychology*, *6*. https://www.frontiersin.org/articles/10.3389/fpsyg.2015.00110

Jurgens, D. A., Turney, P. D., Mohammad, S. M., & Holyoak, K. J. (2012). SemEval-2012 Task 2: Measuring degrees of relational similarity. *Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM)*, 356–364.

Kalkstein, D. A., Hackel, L. M., & Trope, Y. (2020). Person-centered cognition: The presence of people in a visual scene promotes relational reasoning. *Journal of Experimental Social Psychology*, *90*, 104009. https://doi.org/10.1016/j.jesp.2020.104009

Keane, M. T., & Brayshaw, M. (1988). The incremental analogy machine: A computational model of analogy. *Proceedings of the 3rd European Conference on European Working Session on Learning*, 53–62.

Kittur, A., Holyoak, K. J., & Hummel, J. E. (2006). Using Ideal Observers in Higher-order Human Category Learning. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *28*(28). https://escholarship.org/uc/item/2hm1829m

Kittur, A., Hummel, J. E., & Holyoak, K. J. (2004). Feature-vs. relation-defined categories: Probab (alistic) ly not the same. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *26*(26).

Kittur, A., Yu, L., Hope, T., Chan, J., Lifshitz-Assaf, H., Gilon, K., Ng, F., Kraut, R. E., & Shahaf, D. (2019). Scaling up analogical innovation with crowds and AI. *Proceedings of the*

National Academy of Sciences, *116*(6), 1870–1877. https://doi.org/10.1073/pnas.1807185116

Kokinov, B. N. (1994). A hybrid model of reasoning by analogy. In *Analogical connections* (pp. 247–318). Ablex Publishing.

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the "enemy of induction"? *Psychological Science, 19*(6), 585–592. https://doi.org/10.1111/j.1467-9280.2008.02127.x

Kroger, J. K., Holyoak, K. J., & Hummel, J. E. (2004). Varieties of sameness: The impact of relational complexity on perceptual comparisons. *Cognitive Science*, 24.

Kroger, J. K., Saab, F. W., Fales, C. L., Cohen, M. A., & Holyoak, K. J. (2002). Recruitment of anterior dorsolateral prefrontal cortex in human reasoning: A parametric study of relational complexity. *Cerebral Cortex*, *12*(5), 477–485. https://doi.org/10.1093/cercor/12.5.477

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*, 22–44. https://doi.org/10.1037/0033-295X.99.1.22

Kubricht, J. R., Holyoak, K. J., & Lu, H. (2017). Intuitive Physics: Current Research and Controversies. *Trends in Cognitive Sciences*, *21*(10), 749–759. https://doi.org/10.1016/j.tics.2017.06.002

Kurtz, K. H., & Hovland, C. I. (1956). Concept learning with differing sequences of instances. *Journal of Experimental Psychology*, *51*, 239–243. https://doi.org/10.1037/h0040295

Kurtz, K. J., Boukrina, O., & Gentner, D. (2013). Comparison promotes learning and transfer of relational categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(4), 1303–1310. https://doi.org/10.1037/a0031847

Lakoff, G., & Johnson, M. (2003). *Metaphors We Live By* (W. a new Afterword, Ed.). University of Chicago Press. https://press.uchicago.edu/ucp/books/book/chicago/M/bo3637992.html

Leech, R., Mareschal, D., & Cooper, R. P. (2008). Analogy as relational priming: A developmental and computational perspective on the origins of a complex cognitive skill. *The Behavioral and Brain Sciences*, *31*(4), 357–378; discussion 378-414. https://doi.org/10.1017/S0140525X08004469

Lenth, R. V. (2023). *emmeans: Estimated Marginal Means, aka Least-Squares Means*. https://CRAN.R-project.org/package=emmeans

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A Network Model of Category Learning. *Psychological Review*, *111*(2), 309–332. https://doi.org/10.1037/0033-295X.111.2.309

Lu, H., Chen, D., & Holyoak, K. J. (2012). Bayesian analogy with relational transformations. *Psychological Review*, *119*(3), 617–648. https://doi.org/10.1037/a0028719

Lu, H., Ichien, N., & Holyoak, K. J. (2022). Probabilistic analogical mapping with semantic relation networks. *Psychological Review*, *129*(5), 1078–1103. https://doi.org/10.1037/rev0000358

Lu, H., Wu, Y. N., & Holyoak, K. J. (2019). Emergence of analogy from relation learning. *Proceedings of the National Academy of Sciences*, *116*(10), 4176–4181. https://doi.org/10.1073/pnas.1814779116

Mahr, J., & Schacter, D. L. (2023). *A language of episodic thought? (Commentary on Quilty-Dunn et al.)*. PsyArXiv. https://doi.org/10.31234/osf.io/8p3cb

Marcus, G. F. (2003). *The Algebraic Mind: Integrating Connectionism and Cognitive Science*. MIT Press.

Marcus, G. F., Vijayan, S., Bandi Rao, S., & Vishton, P. M. (1999). Rule learning by seven-month-old infants. *Science (New York, N.Y.)*, *283*(5398), 77–80. https://doi.org/10.1126/science.283.5398.77

Markman, A. B. (1996). Structural alignment in similarity and difference judgments. *Psychonomic Bulletin & Review*, *3*(2), 227–230. https://doi.org/10.3758/BF03212423

Markman, A. B., & Gentner, D. (1993a). Splitting the differences: A structural alignment view of similarity. *Journal of Memory and Language*, *32*(4), 517–535. https://doi.org/10.1006/jmla.1993.1027

Markman, A. B., & Gentner, D. (1993b). Structural alignment during similarity comparisons. *Cognitive Psychology*, *25*(4), 431–467. https://doi.org/10.1006/cogp.1993.1011

Markman, A. B., & Stilwell, C. H. (2001). Role-governed categories. *Journal of Experimental & Theoretical Artificial Intelligence*, *13*(4), 329–358. https://doi.org/10.1080/09528130110100252

Martin, A. E., & Doumas, L. A. A. (2020). Tensors and compositionality in neural systems. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1791), 20190306. https://doi.org/10.1098/rstb.2019.0306

Medin, D. L., Goldstone, R. L., & Gentner, D. (1990). Similarity involving attributes and relations: Judgments of similarity and difference are not inverses. *Psychological Science*, *1*(1), 64–69. https://doi.org/10.1111/j.1467-9280.1990.tb00069.x

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *ArXiv:1301.3781*. http://arxiv.org/abs/1301.3781

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *ArXiv:1310.4546*. http://arxiv.org/abs/1310.4546

Minervino, R. A., Margni, A., & Trench, M. (2023). Analogical inferences mediated by relational categories. *Cognitive Psychology*, *142*, 101561. https://doi.org/10.1016/j.cogpsych.2023.101561

Morrison, R. G., Doumas, L. A. A., & Richland, L. E. (2011). A computational account of children's analogical reasoning: Balancing inhibitory control in working memory and relational representation. *Developmental Science*, *14*(3), 516–529. https://doi.org/10.1111/j.1467-7687.2010.00999.x

Namy, L. L., & Gentner, D. (2002). Making a silk purse out of two sow's ears: Young children's use of comparison in category learning. *Journal of Experimental Psychology: General*, *131*(1), 5–15. https://doi.org/10.1037/0096-3445.131.1.5

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 402–407. https://doi.org/10.3758/BF03195588

Nersessian, N. (1992). How Do Scientists Think? Capturing the Dynamics of Conceptual Change in Science. In R. Giere & H. Feigl (Eds.), *Cognitive Models of Science* (pp. 3–45). University of Minnesota Press.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57. https://doi.org/10.1037/0096-3445.115.1.39

Nosofsky, R. M. (1988). Similarity, frequency, and category representations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 54–65. https://doi.org/10.1037/0278-7393.14.1.54

Nosofsky, R. M. (1991). Relation between the Rational Model and the Context Model of Categorization. *Psychological Science*, *2*(6), 416–421. https://doi.org/10.1111/j.1467-9280.1991.tb00176.x

Nosofsky, R. M., & Zaki, S. R. (2003). A hybrid-similarity exemplar model for predicting distinctiveness effects in perceptual old-new recognition. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *29*(6), 1194–1209. https://doi.org/10.1037/0278-7393.29.6.1194

Penn, D. C., Holyoak, K. J., & Povinelli, D. J. (2008). Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. *Behavioral and Brain Sciences*, *31*(02). https://doi.org/10.1017/S0140525X08003543

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. https://doi.org/10.3115/v1/D14-1162

Perfetti, C. A. (1967). A study of denotative similarity with restricted word associations. *Journal of Verbal Learning and Verbal Behavior*, *6*(5), 788–795. https://doi.org/10.1016/S0022-5371(67)80087-2

Peterson, J. C., Chen, D., & Griffiths, T. L. (2020). Parallelograms revisited: Exploring the limitations of vector space models for simple analogies. *Cognition*, *205*, 104440. https://doi.org/10.1016/j.cognition.2020.104440

Popov, V., & Hristova, P. (2015). Unintentional and efficient relational priming. *Memory & Cognition*, *43*(6), 866–878. https://doi.org/10.3758/s13421-015-0514-6

Popov, V., Hristova, P., & Anders, R. (2017). The relational luring effect: Retrieval of relational information during associative recognition. *Journal of Experimental Psychology: General*, *146*(5), 722–745. https://doi.org/10.1037/xge0000305

Popov, V., Pavlova, M., & Hristova, P. (2020). *The Internal Structure of Semantic Relations: Effects of Relational Similarity and Typicality* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/fqd4b

Quilty-Dunn, J., Porot, N., & Mandelbaum, E. (2022). The Best Game in Town: The Re-Emergence of the Language of Thought Hypothesis Across the Cognitive Sciences. *Behavioral and Brain Sciences*, 1–55. https://doi.org/10.1017/S0140525X22002849

R. Core Team. (2021). R: A language and environment for statistical computing (Version 4.0. 5). *R Foundation for Statistical Computing*.

Reder, L. M., Nhouyvanisvong, A., Schunn, C. D., Ayers, M. S., Angstadt, P., & Hiraki, K. (2000). A mechanistic account of the mirror effect for word frequency: A computational model of remember–know judgments in a continuous recognition paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(2), 294.

Rescorla, Robert. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. *Classical Conditioning, Current Research and Theory*, *2*, 64–69.

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 803–814. https://doi.org/10.1037/0278-7393.21.4.803

Rumelhart, D. E., & Abrahamson, A. A. (1973). A model for analogical reasoning. *Cognitive Psychology*, *5*(1), 1–28. https://doi.org/10.1016/0010-0285(73)90023-6

Sagi, E., Gentner, D., & Lovett, A. (2012). What difference reveals about similarity. *Cognitive Science*, *36*(6), 1019–1050. https://doi.org/10.1111/j.1551-6709.2012.01250.x

Santoro, A., Raposo, D., Barrett, D. G. T., Malinowski, M., Pascanu, R., Battaglia, P., & Lillicrap, T. (2017). A simple neural network module for relational reasoning. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4974–4983.

Schonberg, C., Marcus, G. F., & Johnson, S. P. (2018). The roles of item repetition and position in infants' abstract rule learning. *Infant Behavior and Development*, *53*, 64–80. https://doi.org/10.1016/j.infbeh.2018.08.003

Shanahan, M., Nikiforou, K., Creswell, A., Kaplanis, C., Barrett, D., & Garnelo, M. (2019). *An Explicitly Relational Neural Network Architecture* (arXiv:1905.10307; Version 1). arXiv. https://doi.org/10.48550/arXiv.1905.10307

Shanks, D. R., & Dickinson, A. (1988). Associative Accounts of Causality Judgment. In G. H. Bower (Ed.), *Psychology of Learning and Motivation* (Vol. 21, pp. 229–261). Academic Press. https://doi.org/10.1016/S0079-7421(08)60030-4

Shepard, R. N. (1987). Toward a Universal Law of Generalization for Psychological Science. *Science*, *237*(4820), 1317–1323. https://doi.org/10.1126/science.3629243

Sherman, M. A. (1976). Adjectival negation and the comprehension of multiply negated sentences. *Journal of Verbal Learning and Verbal Behavior*, *15*(2), 143–157. https://doi.org/10.1016/0022-5371(76)90015-3

Silliman, D. C., & Kurtz, K. J. (2019). Evidence of analogical re-representation from a change detection task. *Cognition*, *190*, 128–136. https://doi.org/10.1016/j.cognition.2019.04.031

Simmons, S., & Estes, Z. (2008). Individual differences in the perception of similarity and difference. *Cognition*, *108*(3), 781–795. https://doi.org/10.1016/j.cognition.2008.07.003

Spearman, C. (1923). *The nature of "intelligence" and the principles of cognition*. Macmillan.

Spellman, B. A., Holyoak, K. J., & Morrison, R. G. (2001). Analogical priming via semantic relations. *Memory & Cognition*, *29*(3), 383–393. https://doi.org/10.3758/BF03196389

Thibodeau, P. H., Flusberg, S. J., Glick, J. J., & Sternberg, D. A. (2013). An emergent approach to analogical inference. *Connection Science*, *25*(1), 27–53. https://doi.org/10.1080/09540091.2013.821458

Turney, P. D. (2008). The Latent Relation Mapping Engine: Algorithm and Experiments. *Journal of Artificial Intelligence Research*, *33*, 615–655. https://doi.org/10.1613/jair.2693

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*, 327–352. https://doi.org/10.1037/0033-295X.84.4.327

Utsumi, A. (2020). Exploring What Is Encoded in Distributional Word Vectors: A Neurobiologically Motivated Analysis. *Cognitive Science*, *44*(6), e12844. https://doi.org/10.1111/cogs.12844

Vendetti, M. S., Wu, A., & Holyoak, K. J. (2014). Far-Out Thinking: Generating Solutions to Distant Analogies Promotes Relational Thinking. *Psychological Science*, *25*(4), 928–933. https://doi.org/10.1177/0956797613518079

Vlach, H. A., Sandhofer, C. M., & Kornell, N. (2008). The spacing effect in children's memory and category induction. *Cognition*, *109*(1), 163–167. https://doi.org/10.1016/j.cognition.2008.07.013

Waltz, J. A., Lau, A., Grewal, S. K., & Holyoak, K. J. (2000). The role of working memory in analogical mapping. *Memory & Cognition*, *28*(7), 1205–1212. https://doi.org/10.3758/bf03211821

Weinberger, A. B., Gallagher, N. M., Colaizzi, G., Liu, N., Parrott, N., Fearon, E., Shaikh, N., & Green, A. E. (2022). Analogical mapping across sensory modalities and evidence for a general analogy factor. *Cognition*, *223,* 105029. https://doi.org/10.1016/j.cognition.2022.105029

Winston, M. E., Chaffin, R., & Herrmann, D. (1987). A Taxonomy of Part-Whole Relations. *Cognitive Science*, *11*(4), 417–444. https://doi.org/10.1207/s15516709cog1104_2