Cognitive Psychology 141 (2023) 101550

Contents lists available at ScienceDirect

Cognitive Psychology

journal homepage: www.elsevier.com/locate/cogpsych

Relational and lexical similarity in analogical reasoning and recognition memory: Behavioral evidence and computational evaluation

Nicholas Ichien [a,*], Katherine L. Alfred [b], Sophia Baia [c], David J.M. Kraemer [d], Keith J. Holyoak [a,e], Silvia A. Bunge [c,f], Hongjing Lu [a,g]

[a] Department of Psychology, University of California, Los Angeles, United States of America
[b] Department of Psychological and Brain Sciences, Dartmouth College, United States of America
[c] Department of Psychology, University of California, Berkeley, United States of America
[d] Department of Education, Dartmouth College, United States of America
[e] Brain Research Institute, University of California, Los Angeles, United States of America
[f] Helen Wills Neuroscience Institute, University of California, Berkeley, United States of America
[g] Department of Statistics, University of California, Los Angeles, United States of America

ARTICLE INFO

Keywords:
Similarity
Analogy
Recognition memory
False memory

ABSTRACT

We examined the role of different types of similarity in both analogical reasoning and recognition memory. On recognition tasks, people more often falsely report having seen a recombined word pair (e.g., *flower: garden*) if it instantiates the same semantic relation (e.g., *is a part of*) as a studied word pair (e.g., *house: town*). This phenomenon, termed *relational luring*, has been interpreted as evidence that explicit relation representations—known to play a central role in analogical reasoning—also impact episodic memory. We replicate and extend previous studies, showing that relation-based false alarms in recognition memory occur after participants encode word pairs either by making relatedness judgments about individual words presented sequentially, or by evaluating analogies between pairs of word pairs. To test alternative explanations of relational luring, we implemented an established model of recognition memory, the Generalized Context Model (GCM). Within this basic framework, we compared representations of word pairs based on similarities derived either from explicit relations or from lexical semantics (i.e., individual word meanings). In two experiments on recognition memory, best-fitting values of GCM parameters enabled *both* similarity models (even the model based solely on lexical semantics) to predict relational luring with comparable accuracy. However, the model based on explicit relations proved more robust to parameter variations than that based on lexical similarity. We found this same pattern of modeling results when applying GCM to an independent set of data reported by Popov, Hristova, and Anders (2017). In accord with previous work, we also found that explicit relation representations are necessary for modeling analogical reasoning. Our findings support the possibility that explicit relations, which are central to analogical reasoning, also play an important role in episodic memory.

* Corresponding author at: Department of Psychology, University of California, Los Angeles, 405 Hilgard Ave., Los Angeles, CA 90095-1563, United States of America.
E-mail address: ichien@g.ucla.edu (N. Ichien).

https://doi.org/10.1016/j.cogpsych.2023.101550
Received 2 June 2022; Received in revised form 12 January 2023; Accepted 17 January 2023
Available online 30 January 2023

1. Introduction

Human reasoning depends on the ability to represent the world not only in terms of individual concepts, such as *beagle* and *dog*, but also in terms of the *relations* between concepts, such as the fact that a beagle *is a kind of* dog. Computational models of human analogical reasoning have incorporated explicit representations of relations, which enable a relation to link multiple pairs of concepts while remaining distinct from any particular pair of concepts (e.g., Falkenhainer, Forbus, & Gentner, 1989; Hummel & Holyoak, 1997). For example, the relation *is a kind of* can also link *spear* and *weapon*, and an indefinite number of other concept pairs, while maintaining its separate identity.

1.1. Relational luring in recognition memory

If relations have explicit representations used in reasoning tasks, 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.

1.2. 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, Chang, Lee, & Toutanova, 2019; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013; Pennington, Socher, & Manning, 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, Rinaldi, & Marelli, 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, Markman, & Stilwell, 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 we 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, we 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*; Lu, Chen, & Holyoak, 2012; Lu, Wu, & Holyoak, 2019; Lu, Ichien, & Holyoak, 2022), or can be implicit in the raw word embeddings (Mikolov et al., 2013; Pennington et al., 2014).

2. Experiment 1

We 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, we 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, we compare a measure of *relational* similarity between explicit relation representations with a measure of *lexical* similarity between individual word meanings. Based on previous findings, we 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/).

2.1. Method

2.1.1. 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 our 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).

2.1.2. 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 our 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. Fig. 1 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.

2.1.3. 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

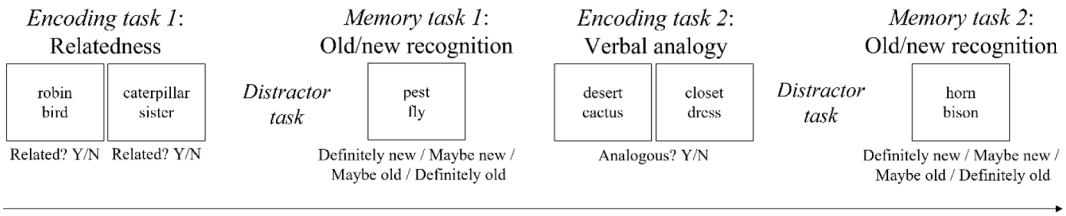

Fig. 1. Task structure. Participants completed six tasks, divided into two blocks (columns) of three tasks each. Task order was fixed. The two blocks of tasks were the same except for the encoding task, with assignment of specific word pairs counterbalanced across the two sets.

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 *concreteness* and *prevalence* norms. 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. We used concreteness norms presented by Brysbaert, Warriner, and Kuperman (2014), which were collected as ratings on a 5-point scale from 1 (abstract) to 5 (concrete). Word pair stimuli were eliminated from our 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. We used prevalence ratings presented by Brysbaert, Mandera, McCormick, and Keuleers (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 our study if either of the two words in a pair had a prevalence rating lower than 2. In our analyses, we also included word-pair concreteness and prevelance as covariates, along with word-pair length and frequency based on norms derived from American film and television show subtitles (Brysbaert and New, 2009).

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.

2.1.4. 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.

A total of 108 word pairs were used for the memory tasks, with each word pair drawn from one of four types (see Table 1). 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

Table 1
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	✔			

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 our 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, we 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).

2.2. Results

2.2.1. Encoding tasks

Overall, participants performed well on both of the encoding tasks: relatedness task, $M_{Acc} = 0.94$, $SD_{Acc} = 0.03$; verbal analogy task, $M_{Acc} = 0.76$, $SD_{Acc} = 0.12$. Fig. 2 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 ($M_{D'} = 2.80$; $SD_{D'} = 0.66$). 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.

2.2.2. Recognition memory

Participants showed good overall performance in recognizing studied word pairs following both encoding tasks (relatedness: $M_{Acc} = 0.81$, $SD_{Acc} = 0.12$; verbal analogy: $M_{Acc} = 0.80$, $SD_{Acc} = 0.13$). They correctly recognized old word pairs with responses of either "Maybe Old" or "Definitely Old", exhibiting a high hit rate (relatedness: $M_{Hit} = 0.90$, $SD_{Hit} = 0.12$; verbal analogy: $M_{Hit} = 0.86$, $SD_{Hit} =$

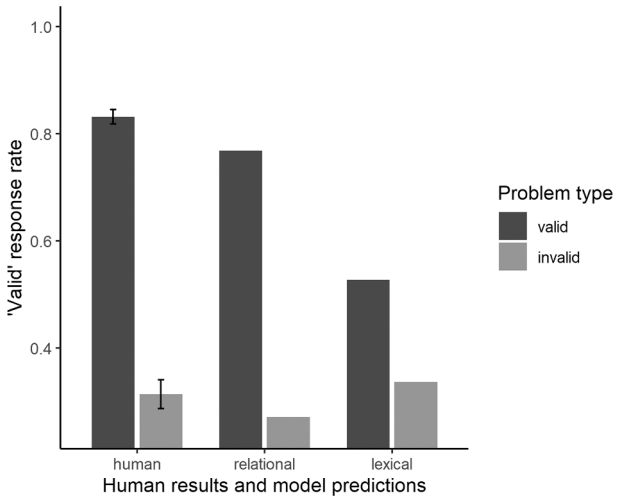

Fig. 2. Human and model-predicted (i.e., relational and lexical) 'valid' responses on the verbal analogy task in Experiment 1. Darker bars represent hits on valid analogies, and lighter bars represent false alarms on invalid analogies. Error bars reflect ±1 standard error of the mean for human responses.

0.14). However, they also sometimes misrecognized recombined word pairs (familiar, unfamiliar, or unrelated), exhibiting a substantial false alarm rate (relatedness: $M_{FA} = 0.24$, $SD_{FA} = 0.16$; verbal analogy: $M_{FA} = 0.24$, $SD_{FA} = 0.17$). Fig. 3 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: $M_{FA} = 0.33$, $SD_{FA} = 0.19$; verbal analogy: $M_{FA} = 0.30$, $SD_{FA} = 0.19$) than relationally unfamiliar word pairs (relatedness: $M_{FA} = 0.21$, $SD_{FA} = 0.20$; verbal analogy: $M_{FA} = 0.22$, $SD_{FA} = 0.22$), and for unfamiliar than unrelated pairs (relatedness: $M_{FA} = 0.09$, $SD_{FA} = 16$; verbal analogy: $M_{FA} = 0.11$, $SD_{FA} = 0.19$). The higher false alarm rate for familiar than unfamiliar pairs is consistent with the relational luring phenomenon reported by Popov et al. (2017).

To statistically test whether Experiment 1 replicated the relational luring effect, while controlling for other potential covariates, we analyzed false alarm data using logistic mixed-effects models. We used the *glmer* function from version 1.1.26 of the LME4 package (Bates, Maechler, Bolker, & Walker, 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, we 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.

We 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*. We 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 our 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), we 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, we 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, we 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, we 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. We 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 our study varied word positions between the encoding tasks and the memory test task. In our 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, we 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). We 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

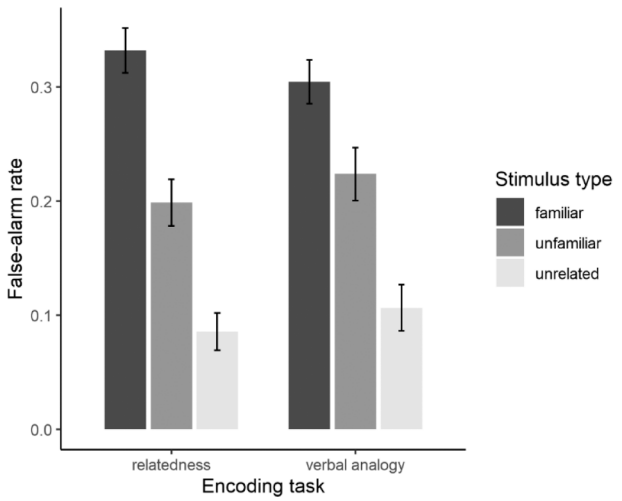

Fig. 3. Human false-alarm rates on the recognition memory task in Experiment 1, broken down by relatedness and verbal analogy encoding task and by familiar, unfamiliar, and unrelated stimulus types. Error bars reflect ± SEM.

model prediction error. Inspecting fit parameters, we 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, $\beta = 0.95$, $z = 3.09$, $p = .002$.

3. Experiment 2

Although Experiment 1 demonstrated relational luring, we were 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 we 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, we 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.

3.1. Method

3.1.1. 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 our 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).

3.1.2. Procedure

Because relatedness judgments and solving verbal analogies both proved sufficient to induce relational luring in Experiment 1, we 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. Fig. 4 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.

3.1.3. 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 & Hristova, 2015). All stimuli were based on a pool of 84 semantically-related word pairs. To create the present stimulus set, we edited Popov et al.'s translated stimuli in a few ways. We 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*).

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 2 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

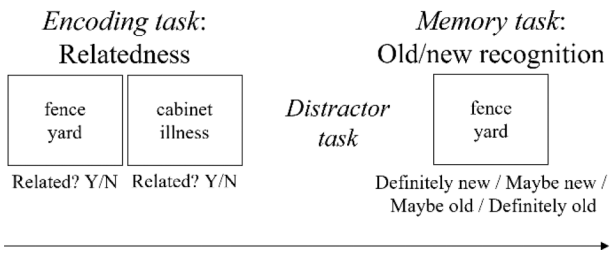

Fig. 4. Task structure for *Experiment 2*. Participants completed three tasks in a fixed order.

Table 2
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*	

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 3 shows the encoding-task and memory-task stimuli generated from a single stimulus set for two distinct participants.

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*. (We 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, we 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. We 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.

3.2. Results

3.2.1. Encoding task

Overall, participants performed well on the encoding task: $M_{Acc} = 0.90$, $SD_{Acc} = 0.06$, with a low rate of false positive judgments ($M_{FA} = 0.09$, $SD_{FA} = 0.06$).

3.2.2. Memory task

Overall, participants performed well on the memory task: $M_{Acc} = 0.76$, $SD_{Acc} = 0.113$, with a moderately high false-alarm rate ($M_{FA} = 0.16$, $SD_{FA} = 0.14$). We also found that false alarms were more frequent for relationally familiar word pairs ($M_{FA} = 0.17$, $SD_{FA} = 0.16$) than relationally unfamiliar word pairs ($M_{FA} = 0.14$, $SD_{FA} = 0.13$). As in Experiment 1, we fit logistic mixed-effects models to the human false-alarm data. We 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*. We found that omitting the stimulus type term reliably increased model prediction error, $\Delta AIC = 8$, $\chi^2(1) = 9.97$, $p = .002$, indicating that

Table 3
Stimuli in Experiment 2 generated from the set presented in Table 2, adapted from Popov et al. (2017).

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

participants were more likely to false alarm on relationally familiar than relationally unfamiliar word pairs (see Fig. 5). Hence, despite various methodological differences between the present study and the experiment reported by Popov et al. (2017), we 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 (0.03) is smaller than that demonstrated in Experiment 1 (0.11). 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), we 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.

4. Tests of computational models

4.1. Measures of word-pair similarity

To predict performance on both the analogy task and the recognition memory task, we 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. We implemented specific versions of both possibilities, all rooted in 300-dimensional word embeddings created by Word2vec.

As shown in Fig. 6 top panel, to compute relational similarity we 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, Lu, & Holyoak, 2022), as shown in Fig. 6 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:

$$d_{Rel_{ij}} = cos\left(R_{i,}, R_j\right). \tag{1}$$

As shown in Fig. 6 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. We 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*. We 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}:

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

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

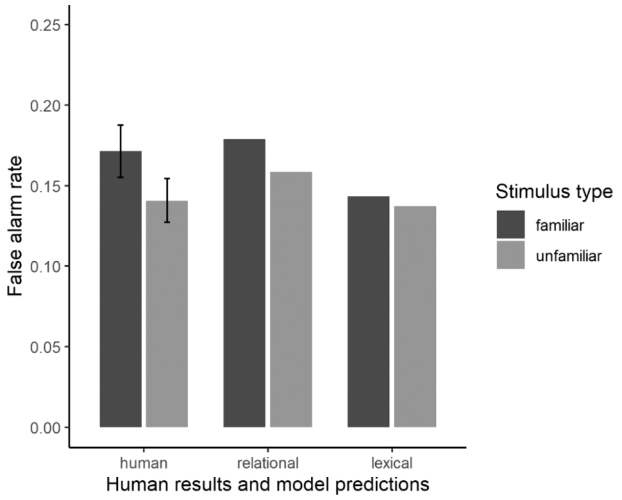

Fig. 5. Human false-alarm rates and model predictions on the recognition memory task in Experiment 2, broken down according to stimulus type (relationally familiar and relationally unfamiliar word pairs). Error bars reflect ±1 SEM.

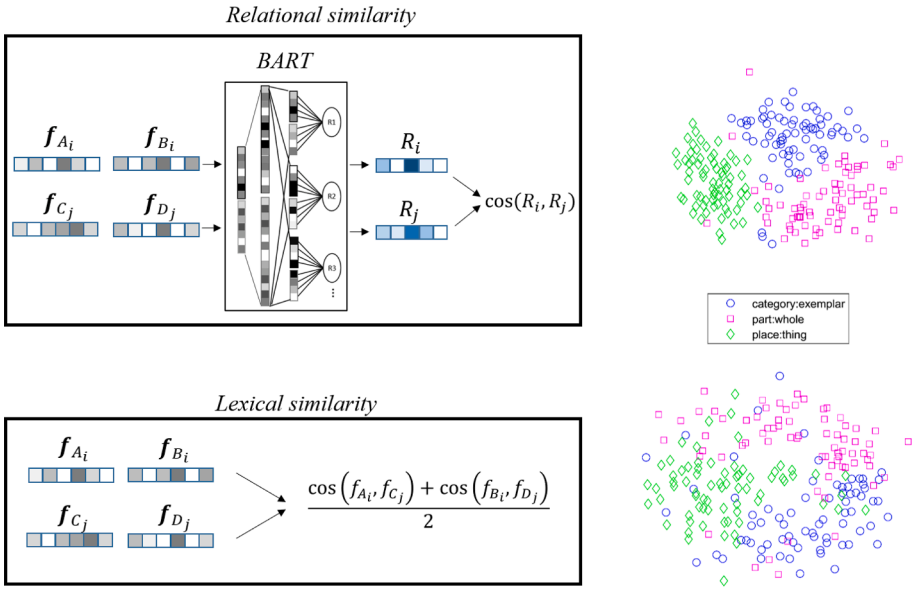

Fig. 6. An illustration of relation similarity model (left top panel) and lexical similarity model (left bottom panel), and the resulting 2-D plot of similarity space derived using each (right panel). The scatter plots of similarity spaces are derived from 216 word-pair stimuli instantiating *category: 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). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

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*), Fig. 6 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).

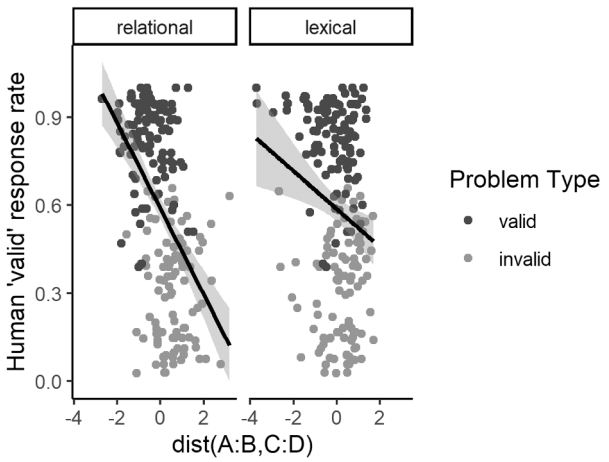

Fig. 7. Human item-level 'valid' response rates on verbal analogy problems in Experiment 1, plotted against z-scored distance (dissimilarity) metrics predicted by the relational model (left) and by the lexical model (right). Each point represents a single analogy problem, and point shade reflects whether a problem features a valid analogy (dark grey) or an invalid analogy (light grey). The scatter plots were overlaid with a fitted regression line.

4.2. 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, Peng, Lu, Holyoak, & Monti, 2021) responses to analogy problems. In order to predict yes/no decisions about analogy problems, we 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, we used relational and lexical similarity metrics. Based on prior modeling of verbal analogical reasoning (Chiang et al., 2021; Lu et al., 2019) and of explicit judgments of relation similarity (Ichien et al., 2022), we predicted that the model based on relational similarity would best predict human judgments on the explicit analogy task.

Fig. 2 (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 (0.75), nearly matching human proportion correct (0.76). The alternative model based on lexical (non-relational) similarity performed poorly (0.59 correct).

An item-level analysis corroborated these results. We 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 ($r = 0.47$) than was lexical similarity ($r = 0.21$; $z = 3.68$, $p < .001$). Fig. 7 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).

4.3. Modeling recognition memory

To provide a formal account of relational luring, we adapted an established model of recognition memory, the Generalized Context Model (GCM; Nosofsky, 1986, 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, Medin, & Gureckis, 2004). If a version of GCM is able to account for relation-based false alarms, we 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, we 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

$$P(old|i) = \frac{F_i}{F_i + k}, \tag{3}$$

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

$$F_i = \sum_{j \in J} s_{ij}. \tag{4}$$

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,

$$s_{ij} = e^{-cd_{ij}}, \tag{5}$$

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 our 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, we scaled all model-generated distance values between 0 and 1. As our representations are high-dimensional, we adopt cosine distance to compute d_{ij}, rather than the Minkowski power formula typically used in previous work (e.g., Nosofsky, 1986, 1988, 1991; Nosofsky & Zaki, 2003).

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

4.3.1. Simulation results for Experiment 1

First, we modeled human recognition memory performance for Experiment 1. Because we found no reliable differences in either false alarm rates or overall accuracy across the two encoding tasks, we 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". We 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. We used a binomial distribution as the likelihood function to fit the scaling parameter c and criterion parameter k that maximized the log-likelihood. Table 4 summarizes fit model parameters, maximum log-likelihood, and RMSD and spearman correlations between fit model predictions and item-level human data. Fig. 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 Fig. 8 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: $M_{Hit} = 0.88$, $SD_{Hit} = 0.10$, $M_{FA} = 24$, $SD_{FA} = 0.15$; relational: $M_{Hit} = 0.79$, $M_{FA} = 0.26$; lexical: $M_{Hit} = 0.79$, $M_{FA} = 0.28$).

Next, we 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, we 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, we computed the log model evidence (Friel & Wyse, 2012; Hoeting, Madigan, & Raftery, 1999) by averaging the log likelihood that each model predicts the proportion of our 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; $k = [0.1,1]$ with a stepsize of 0.1). We 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 we used for model fitting. As shown in Table 5, we found that the log model evidence for the relational similarity metric was $E_{log} = -1.324 \times 10^4$, substantially greater than that for lexical similarity, $E_{log} = -1.569 \times 10^4$. This analysis provides converging evidence that the relational model provides a more robust account of the human data than does the lexical model.

We also examined the range of parameters in models that generate the effect of relational luring. In this analysis we focused on model judgments for two types of test pairs, relationally familiar and relationally unfamiliar word pairs. We 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 Fig. 8, 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 Fig. 9 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).

To provide a quantitative comparison of the robustness with which model predicts relational luring using each similarity metric, we 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 we used to compute log model evidence ($c = [0,50]$ with a stepsize of 0.5; $k = [0.1,1]$ 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 $SD_{luring} = 0.1240$, which was the observed standard deviation 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 5, we 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 Fig. 6, 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

Table 4
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	0.20	0.20	-5.013×10^3	-5.063×10^3	0.163	0.169	0.794	0.764
Exp. 2	15.5	10.5	0.30	0.20	-2.836×10^3	-2.768×10^3	0.159	0.149	0.658	0.665
Popov et al. Exp. 1	11.5	8.0	0.40	0.40	-1.288×10^3	-1.271×10^3	0.199	0.192	0.669	0.644

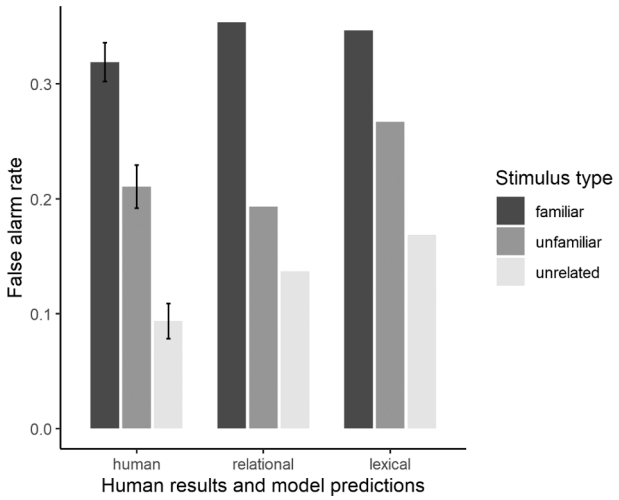

Fig. 8. Human false-alarm rates and model predictions on the recognition memory task in Experiment 1, broken down according to familiar, unfamiliar and unrelated stimulus types. Error bars reflect ±1 SEM.

Table 5
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 = [0.1,1]).

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

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.

Fig. 10 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 Fig. 6) 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.

4.3.2. Simulation results for Experiment 2

Using the same model-fitting procedure as for Experiment 1, we optimized GCM parameters with the maximum log-likelihood fit to the item-level human data for each similarity metric, using a binomial likelihood function. Fig. 5 (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: $M_{Hit} = 0.82$, $SD_{Hit} = 0.22$, $M_{FA} = 0.16$, $SD_{FA} = 0.14$; relational: $M_{Hit} = 0.78$, $M_{FA} = 0.17$; lexical: $M_{Hit} = 0.84$, $M_{FA} = 0.17$).

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

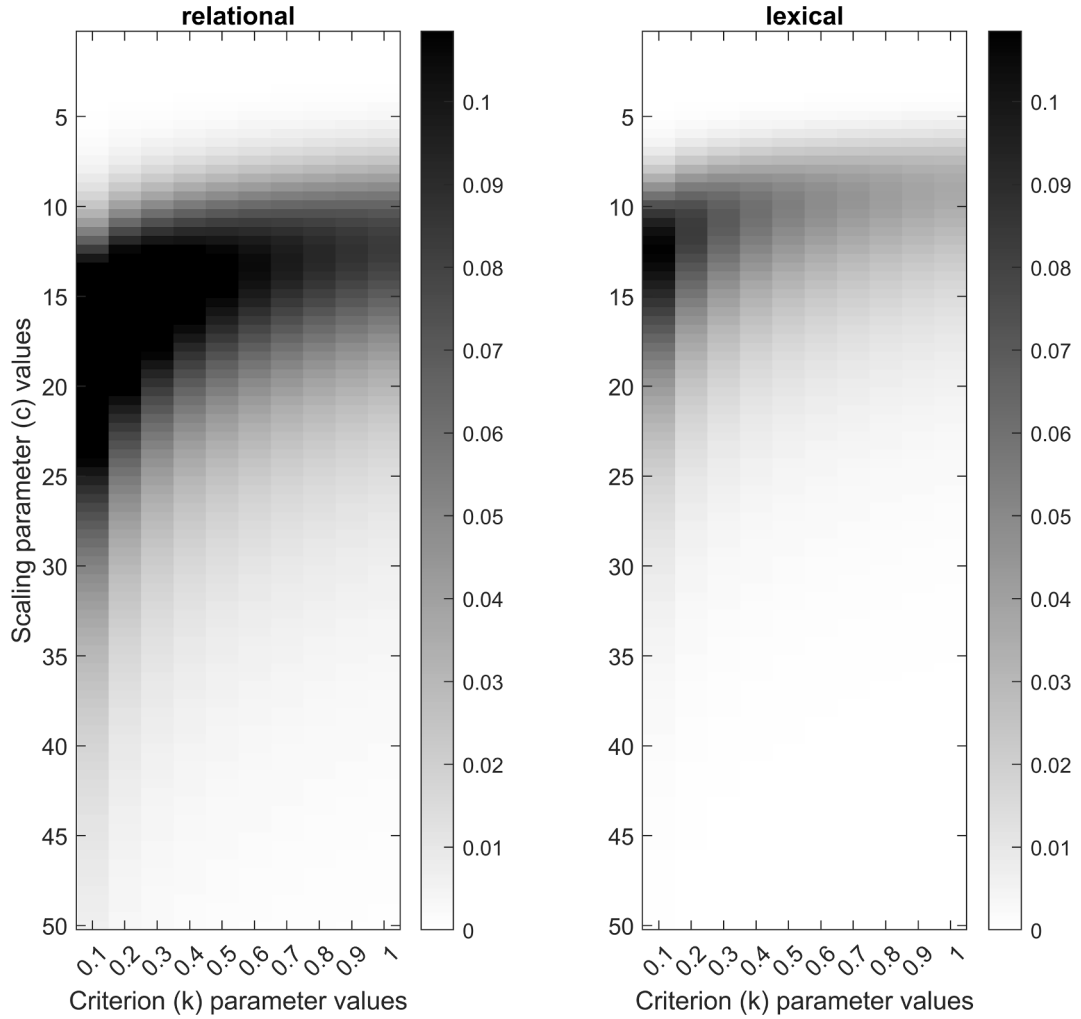

Fig. 9. Simulation of model-predicted relational luring effect in Experiment 1 as a function of model parameters. Each cell represents a combination of values for GCM's scaling parameter *c* (*y*-axis) and its criterion parameter *k* (*x*-axis), respectively. Given the pair of parameter values for each cell, cell shading represents the model-predicted difference of false alarm rates between familiar word pairs and unfamiliar word pairs (i.e., relational luring effect). Darker cells indicate a greater magnitude of model-predicted luring effect. Black cells correspond to the magnitude of the luring effect observed in human data.

generated within the lexical similarity metric. Moreover, as shown in Fig. 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, we 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 5, we found greater model evidence for the relational model ($E_{log} = -6.738 \times 10^3$) than for the lexical model ($E_{log} = -8.131 \times 10^3$). As for Experiment 1, we 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 5 shows, we found that the model evidence for the luring effect observed in our human data ($M_{luring} = 0.0306$, $SD_{luring} = 0.1174$) was greater for the relational model ($E_{luring} = 3.310$) than for the lexical model ($E_{luring} = 3.291$). We 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, we maintain that relational similarity

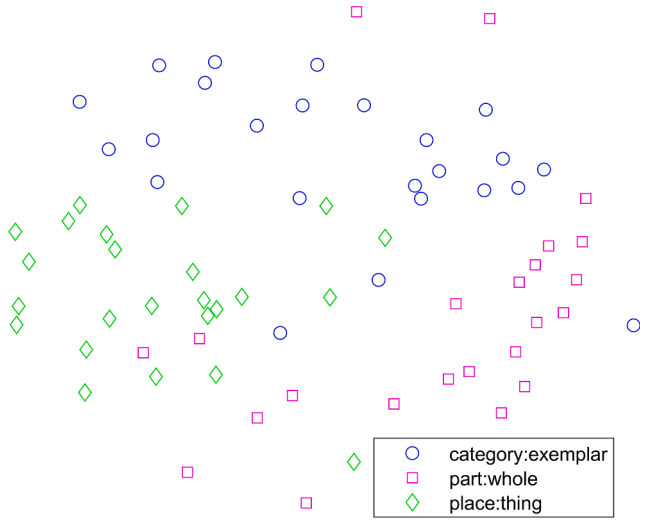

Fig. 10. Multidimensional scaling based on for lexical similarity among individual first words in pairs used in the memory task for Experiment 1. Both colors and shapes redundantly indicate word-pair relations (category:exemplar, part:whole, and place:thing). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

more robustly accounts for human data across a wide range of parameter values.

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, we 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 3 for two examples of *intact*, *relationally familiar*, and *relationally unfamiliar* triplets generated from the same stimulus set.) We computed the relational and lexical distances between each relationally familiar and each relationally unfamiliar word pair and its corresponding intact word pair. Fig. 12 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.

4.3.3. Simulation results for Popov et al. (2017), Experiment 1

In order to provide a conceptual replication of the assessment of computational models we applied to our own experiments (as reported above), we 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 our implementation of GCM (based on Nosofsky, 1986, 1988, 1991; Nosofsky & Zaki, 2003) produces binary responses and more naturally fits a design with separate study and test phases, we 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

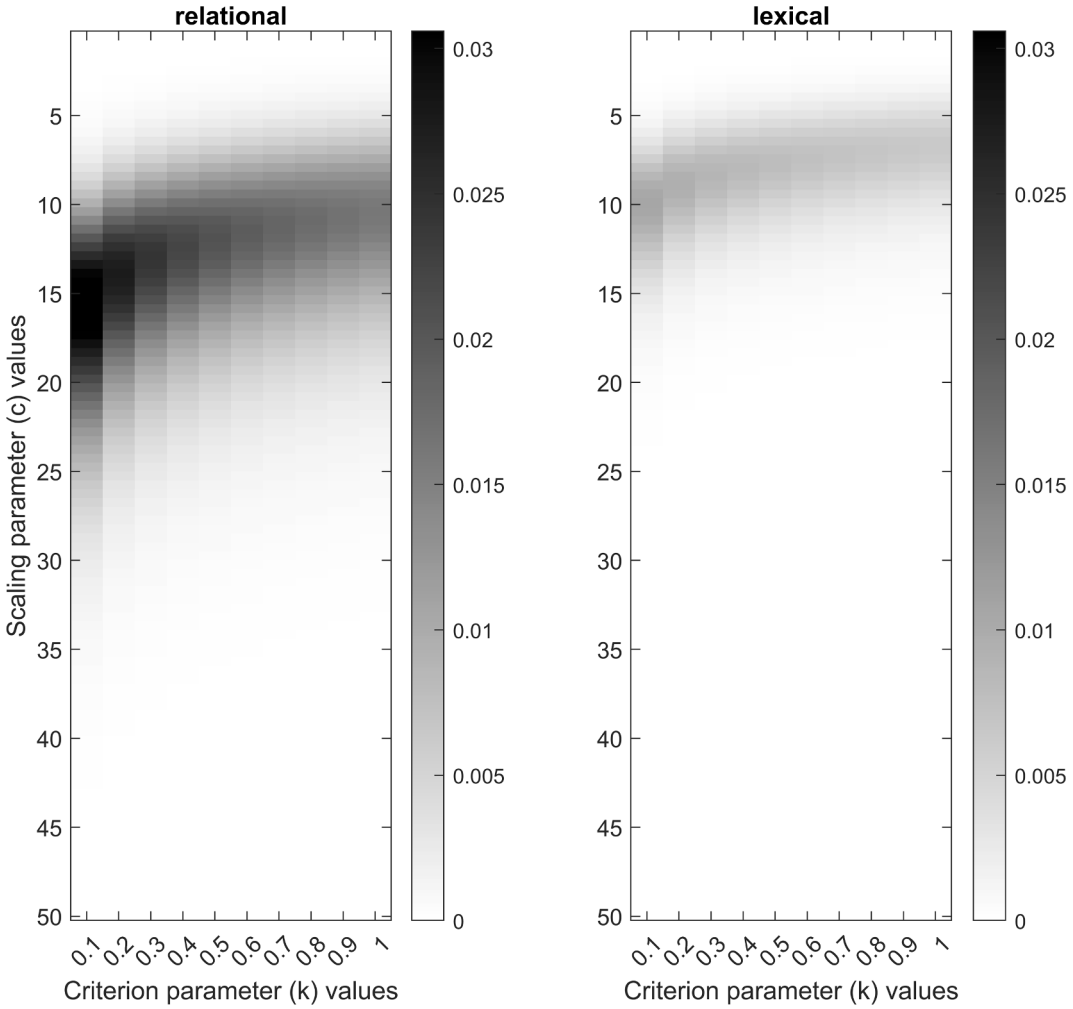

Fig. 11. Simulation of model-predicted relational luring effect in Experiment 2 as a function of model parameters (scaling parameter *c* and its criterion parameter *k*. Shading of cells indicates magnitude of model-predicted luring effect. Black cells correspond to the magnitude of the luring effect in human data.

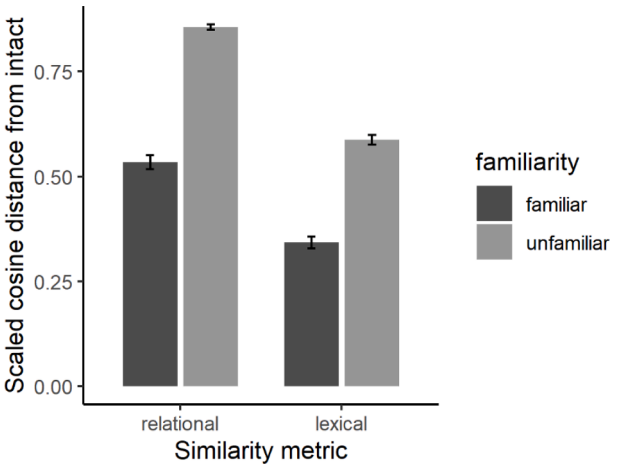

Fig. 12. Mean lexical and relational cosine distances (scaled between 0 and 1) between familiar and unfamiliar word pairs and intact word pairs within each stimulus set used in Experiment 2.

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.) We 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, we 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, we treated word pair-condition combinations as unique items. Fig. 13 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: $M_{Hit} = 0.75$, $SD_{Hit} = 0.18$, $M_{FA} = 0.18$, $SD_{FA} = 0.13$; relational: $M_{Hit} = 0.72$, $M_{FA} = 0.17$; lexical: $M_{Hit} = 0.72$, $M_{FA} = 0.19$). 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.

In the same manner as described for the robustness analyses applied to data from our own experiments, we computed log model evidence for all 63 test items. Log model evidence was greater using relational similarity ($E_{log} = -2.961 \times 10^3$) than lexical similarity ($E_{log} = -3.614 \times 10^3$). We 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 our own data in Experiments 1 and 2, we found that for the relational luring effect observed in Popov et al.'s (2017) data ($M_{luring} = 0.0225$, $SD_{luring} = 0.1078$), 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. Fig. 14 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.).

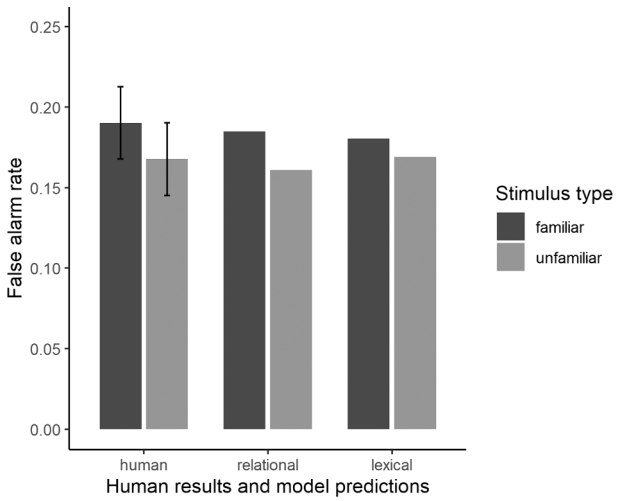

Fig. 13. Human false-alarm rates from Popov et al. (2017), Experiment 1, and model predictions on the recognition memory task, broken down according to stimulus type. Error bars reflect ±1 SEM.

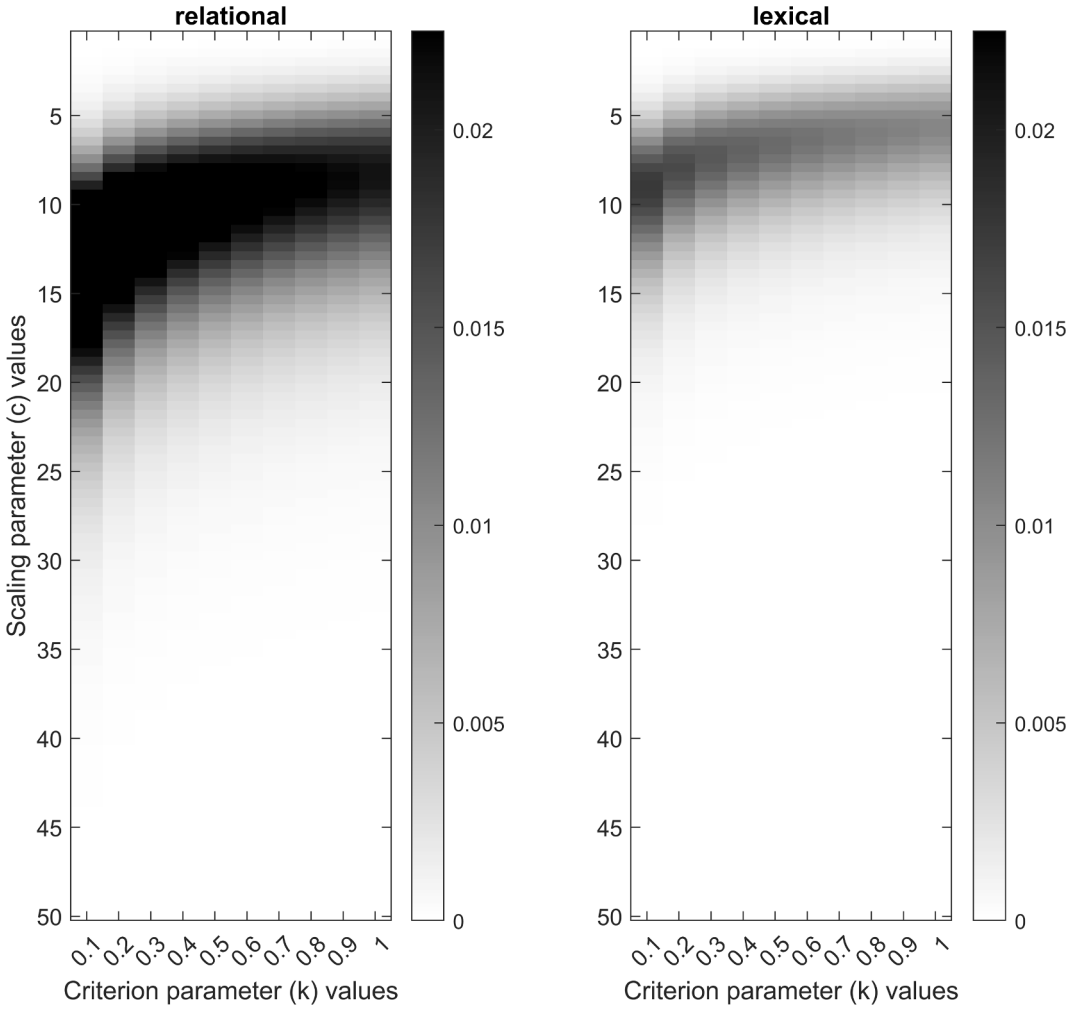

Fig. 14. Simulation of model-predicted relational luring effect as a function of model parameters (scaling parameter *c* and its criterion parameter *k*) for Popov et al. (2017), Experiment 1. Shading of cells indicates magnitude of model-predicted luring effect. Black cells correspond to the magnitude of the luring effect in human data.

5. General discussion

5.1. Summary

We report two experiments and simulations designed to compare alternative representations of word-pair similarity as predictors of both human analogical reasoning and recognition memory. We 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, Lu and Holyoak, 2022).

In our test of recognition memory, we 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. We 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, we tested the two similarity measures within a common theoretical framework provided by the Generalized Context Model (GCM; Nosofsky, 1986, 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 (Fig. 6). 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, our modeling results indicate that explicitly representing relations is not *necessary* for explaining relational luring.

However, we 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. We 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. We computed the log model evidence to provide a quantitative comparison of the robustness to predicting all human data with each similarity metric. In addition, we 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 we 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.

5.2. Conclusion

We 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.

Declaration of Competing Interest

The authors declare no competing interests.

Data availability

Data from the human experiment and code for the simulations are available at https://osf.io/vmn4z/.

Acknowledgements

Preparation of this paper 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. We thank Ven Popov for providing raw data reported in Popov et al. (2017). A preliminary report of part of this research was presented at the 44th Annual Conference of the Cognitive Science Society (online, July 2022).

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cogpsych.2023.101550.

References

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review, 98*, 409–429.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*, 1–48.

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

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.

Brysbaert, M., & New, B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*, 977–990.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods, 46*, 904–911.

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.

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.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers of language understanding. In *, Vol. 1. Proceedings of the 2019 conference for the North American chapter of the Association for Computational Linguistics: Human language technologies* (pp. 4171–4186).

Diedenhofen, B., & Musch, J. (2015). Cocor: A comprehensive solution for the statistical comparison of correlations. *PLoS One, 10*(4), Article e0121945.

Dunn, O. J., & Clark, V. A. (1969). Correlation coefficients measured on the same individuals. *Journal of the American Statistical Association, 64*, 366–377.

Falkenhainer, B., Forbus, K. D., & Gentner, D. (1989). The structure-mapping engine: Algorithm and examples. *Artificial Intelligence, 41*, 1–63.

Friel, N., & Wyse, J. (2012). Estimating the evidence – A review. *Statistica Neerlandica, 66*(3), 288–308.

Goldwater, M. B., Markman, A. B., & Stilwell, C. H. (2011). The empirical case for role-governed categories. *Cognition, 118*, 359–376.

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.

Hoeting, J., Madigan, D., & Raftery, A. E. (1999). Bayesian model averaging: A tutorial. *Statistical Science, 14*(3), 382–401.

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

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.

Jones, M., & Love, B. C. (2007). Beyond common features: The role of roles in determining similarity. *Cognitive Psychology, 55*(3), 196–231.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99*, 22–44.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review, 111*, 309–332.

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*. https://doi.org/10.1037/rev0000358. Advance online publication.

Lu, H., Wu, Y. N., & Holyoak, K. J. (2019). Emergence of analogy from relation learning. *Proceedings of the National Academy of Sciences, USA, 116*(10), 4176–4181.

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

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems, 26*, 3111–3119.

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

Nosofsky, R. M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*(1), 54–65.

Nosofsky, R. M. (1991). Tests of an exemplar model for relating perceptual classification and recognition memory. *Journal of Experimental Psychology: Human Perception and Performance, 17*(1), 3–27.

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.

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. *Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Popov, V., & Hristova, P. (2015). Unintentional and efficient relational priming. *Memory & Cognition, 43*(6), 866–878.

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.

R Core Team. (2021). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at https://www.R-project.org/.

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.

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*(4), 803–814.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science, 237*(4820), 1317–1323.

Utsumi, A. (2020). Exploring what is encoded in distributional word vectors: A neurobiologically motivated analysis. *Cognitive Science, 44*, Article e12844. https://doi.org/10.1111/cogs.12844