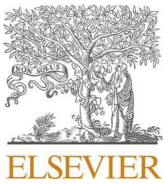
Cognitive Psychology 151 (2024) 101661



Contents lists available at ScienceDirect

Cognitive Psychology

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

Cognitive complexity explains processing asymmetry in judgments of similarity versus difference

Nicholas Ichien [a,*], Nyusha Lin [b], Keith J. Holyoak [b], Hongjing Lu [b,c]

[a] Department of Psychology, University of Pennsylvania, United States
[b] Department of Psychology, University of California, Los Angeles, United States
[c] Department of Statistics, University of California, Los Angeles, United States

ARTICLE INFO

Keywords:
Similarity
Difference
Negation
Relations
Features

ABSTRACT

Human judgments of similarity and difference are sometimes asymmetrical, with the former being more sensitive than the latter to relational overlap, but the theoretical basis for this asymmetry remains unclear. We test an explanation based on the type of information used to make these judgments (relations versus features) and the comparison process itself (similarity versus difference). We propose that asymmetries arise from two aspects of cognitive complexity that impact judgments of similarity and difference: processing relations between entities is more cognitively demanding than processing features of individual entities, and comparisons assessing difference are more cognitively complex than those assessing similarity. In Experiment 1 we tested this hypothesis for both verbal comparisons between word pairs, and visual comparisons between sets of geometric shapes. Participants were asked to select one of two options that was either more similar to or more different from a standard. On *unambiguous* trials, one option was unambiguously more similar to the standard; on *ambiguous* trials, one option was more featurally similar to the standard, whereas the other was more relationally similar. Given the higher cognitive complexity of processing relations and of assessing difference, we predicted that detecting *relational difference* would be particularly demanding. We found that participants (1) had more difficulty detecting relational difference than they did relational similarity on unambiguous trials, and (2) tended to emphasize relational information more when judging similarity than when judging difference on ambiguous trials. The latter finding was replicated using more complex story stimuli (Experiment 2). We showed that this pattern can be captured by a computational model of comparison that weights relational information more heavily for similarity than for difference judgments.

A naïve construal of *similarity* and *difference* is that one is the inverse of the other: As things become more similar, they become less different. Cognitive scientists, however, have demonstrated that human reasoners process the two relations in a way that sometimes violates this assumed symmetry. Specifically, people tend to use distinct types of information when judging what makes things similar versus when judging what makes things different (Bassok & Medin, 1997; Simmons & Estes, 2008; Tversky, 1977). For example, Medin et al. (1990) asked participants to select which of two options was more visually *similar to* or more *different from* a standard. Across trials, one option was designed to be relationally more similar to the standard and the other more featurally similar. Participants tended to select the relationally similar option as both more similar *and* more different from the standard. Bassok and Medin (1997)

* Corresponding author.
E-mail address: nichien@sas.upenn.edu (N. Ichien).

https://doi.org/10.1016/j.cogpsych.2024.101661
Received 22 July 2023; Received in revised form 8 April 2024; Accepted 16 April 2024

found the same asymmetry using verbal stimuli. Broadly, these findings indicate that people tend to consider relations more heavily when judging similarity than when judging difference. However, the reason for this asymmetry remains unclear.

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

Proponents of this explanation noted that the featurally-similar option in the study by Medin et al. (1990) did not involve a salient relation. Accordingly, the relational difference between it and the standard would not be alignable, and hence would have been ignored in difference comparisons. However, later work found that both alignable and nonalignable differences contribute to judgments of difference, and that the latter actually exert a *greater* influence than the former (Estes & Hasson, 2004). These findings appear to undermine the core assumption that enabled structure mapping theory to potentially account for asymmetries in similarity and difference judgments.

1. Processing-demand Hypothesis

In the present paper we propose and test an alternative *processing-demand hypothesis* based on two aspects of cognitive complexity that impact judgments of relational difference. The first factor arises from the content of information that is compared when making these judgments. When human reasoners make comparisons, they tend to do so on the basis of both features of individual entities, and also relations between entities and their component parts. Extensive evidence indicates that processing and comparing relational information is more cognitively demanding than processing featural information (e.g., Bunge et al., 2005; Green et al., 2010; Halford et al., 1998; Kroger et al., 2002, 2004; Waltz et al., 2000).

The second factor arises from a corresponding asymmetry between comparisons assessing similarity versus difference, in which processing difference is more cognitively demanding than processing similarity. For example, when presented with image-word pairings involving either a *circle* or a *square* shape and either the word "square" or the word "circle", participants were much faster to accurately respond "yes" or "no" as to whether image and word were the same (e.g., "yes" to *circle* / "circle" and "no" to *circle* / "square") than when they were different (e.g., "yes" to *circle* / "square" and "no" to *circle* / "circle") (Clark, 1971; Seymour, 1969).

We note that the processing-demand hypothesis is compatible with any number of explanations for the additional cognitive complexity imposed by difference comparisons, relative to similarity comparisons. However, one compelling account for this asymmetry proposes that assessments of difference involve a complex comparison involving the *negation* of sameness (i.e., *not-sameness*), whereas assessments of similarity involve a relatively straightforward comparison of degree of *sameness*. Difference imposes greater processing demands than does similarity because, in general, processing of negation adds complexity.[1] For example, determining the truth of a proposition including a negated expression (e.g., "star isn't above the plus") takes longer than for a matched positive expression (e.g., "star is below the plus") (Carpenter & Just, 1975; Clark & Chase, 1972). Introducing additional negation into sentences makes them more difficult to interpret (e.g., "Because he often worked for hours at a time, *no one* believed that he was *not capable* of sustained effort;" Sherman, 1976). Previous research has shown that processing negation often involves multiple steps, including processing the affirmative components of negated phrases before processing the entire phrase (Hasson & Glucksberg, 2006). Although the complexity of negation is most pronounced when an explicit negative such as *not* is used, processing difficulty is also increased for expressions that incorporate implicit negation (e.g., words such as *few*, *little*, or *deny*; Clark, 1976).

More recently, Hochmann, Mody, and Carey (2016) provided specific evidence that implicit negation of *same* accounts for this added difficulty in processing the relation *different*. Participants were shown three boxes and were asked to either select which of the flanking boxes contained the same object as the middle one (i.e., same option) in a match-to-sample task (MTS); or else were asked which contained a different object from the middle one (i.e., different option) in a non-match-to-sample task (NMTS). During a given trial, box contents were revealed one-by-one before disappearing. In crucial conditions the contents of one of the options remained occluded throughout the trial: In the visible-same condition, participants were shown the same option before being shown the middle object; whereas in the visible-different condition, they were shown the different option before being shown the middle object. Participants completing the MTS task, the goal of which was to select the same option, were at ceiling in accuracy but were slower in the visible-different condition, suggesting that they adopted a straightforward strategy of seeking the same option to select it (i.e., a direct assessment of *sameness*), rather than seeking the different option to avoid it. In contrast, those completing the NMTS task, the goal of which was to select the different option, were less accurate and slower in the visible-different condition, suggesting that they adopted a strategy of seeking the *same* option to avoid it (i.e., an assessment of *not-sameness*) rather than seeking the different option to select it. Moreover, NMTS participants were less accurate and slower overall than MTS participants, providing further evidence that sameness is simpler to process than difference.

[1] We emphasize that our discussion of a processing asymmetry between similarity and difference judgments is distinct from analyses of the metaphysics of same and different (e.g., Gerson, 2004; Grier, 2007) or the semantics of the corresponding terms (Carlson, 1987; Moltmann, 1992).

Finally, the claim that assessing difference involves negating sameness implies that the ability to assess sameness is more basic than assessing difference. This analysis has been used to explain the well-established developmental lag between children's understanding of the concepts *same* versus *different* (Hochmann, 2021; Hochmann et al., 2016, 2018), and also converges with evidence that processing sameness but not difference is privileged across species (Zentall et al., 1981, 2018).[2]

These two factors contributing to the cognitive complexity of comparison (i.e., relations versus features, and similarity versus difference) imply that incorporating relational information will be particularly demanding in a task that also involves difference judgments. We predict that because of this added complexity, difference judgments are less likely than similarity judgments to take account of relational information. Unlike the explanation offered by proponents of structure mapping theory, this account makes no reference to alignability of relations.

2. Overview

In Experiment 1, we tested the processing-demand hypothesis for both verbal comparisons between word pairs and for visual comparisons between sets of geometric shapes. For both types of stimuli, we measured participants' sensitivity to featural and relational information in a 2-alternative forced-choice task, in which participants selected which of two options was more similar to or more different from a standard. In order to directly examine the relative difficulty of similarity and difference judgments, we included *unambiguous* comparisons, in which one option was unambiguously more similar to a standard than the other based either on features or on relations. We predicted that even for unambiguous trials, participants would have greater difficulty in detecting relational difference compared to relational similarity. We also included *ambiguous* comparisons, for which either of the options might be selected depending on whether features or relations are emphasized. We predicted that when judging difference as compared to similarity, participants will tend to base their choices on features rather than relations. In order to check to validity of our stimuli in Experiment 1, we formulated the processing-demand hypothesis as a computational model operating over data-driven measures of featural and relational similarity, based on human ratings, as well as model-derived representations of lexical and relational meaning, using Word2vec (Mikolov et al., 2013) and *Bayesian Analogy with Relational Transformations* (BART) (Lu et al., 2019). We show that predictions generated by this model capture the hypothesized differential weighting of featural and relational information in similarity and difference judgments, regardless of whether it operated over human similarity ratings or model-derived similarity from Word2vec and BART.

In Experiment 2, we used stories to examine whether this same asymmetry between similarity and difference judgments could also be obtained with more complex stimuli. This experiment allowed us to further disambiguate between our processing-demand hypothesis and the alignment hypothesis. We tested whether the observed asymmetry is attributable to distinct comparison processes for similarity and difference, as predicted by our processing-demand hypothesis, or to use of dissociable pools of output from a unified comparison process, as predicted by the alignment hypothesis. To preview our findings, the results supported the processing-demand hypothesis.

3. Experiment 1

3.1. Method

3.1.1. Participants

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

3.1.2. Materials and procedure

Comparison tasks. All participants completed two comparison tasks: A verbal comparison task using word-pair stimuli and a visual comparison task using geometric shape stimuli. On each trial, participants were presented with a standard at the top of the screen and two options on either side at the bottom of the screen. Fig. 1 shows example trials of the visual task (top panel) and the verbal task (bottom panel). Across both comparison tasks, some participants were instructed to select which option was more *similar* to the standard, whereas other participants were asked to select which was more *different* from the standard. Task instructions and all stimuli used in the study can be found in the Supplemental Information at this paper's OSF link: https://osf.io/szqjk/.

Each comparison task consisted of 24 trials, presented in a random order. Of these, 6 *unambiguous* trials included one option that was unambiguously more similar to the standard than the other. On half of the unambiguous trials, the similar option was more featurally similar to the standard than the other option, whereas both options were equally relationally similar to the standard. We refer to these as *featural* trials (see left side of left panel in Fig. 1 for examples), The other 3 unambiguous trials were *relational* trials (see right side of left panel in Fig. 1 for examples). On these trials, the similar option was more relationally similar to the standard, whereas

[2] There is evidence that adults can represent sameness and difference in either a categorical or a continuous fashion (for a review see Davis & Goldwater, 2021). Judgments of degree of similarity or degree of difference (our focus here) are clearly more compatible with a continuous interpretation. In any case, negation seems likely to play a special role in processing difference under either interpretation.

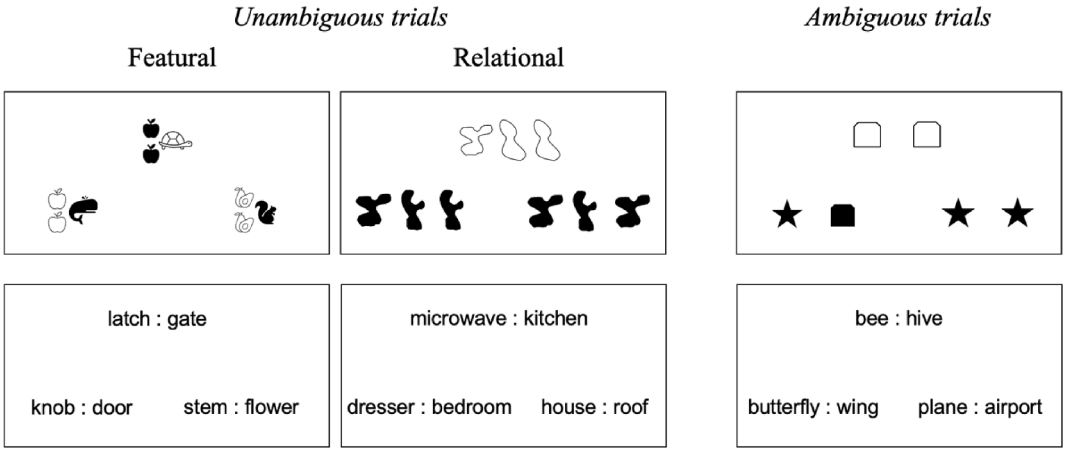


Fig. 1. Examples of each trial type for visual (top) and verbal (bottom) comparison tasks used in Experiment 1. Left panel: Unambiguous trials. Within each trial, one option is more similar (either featurally, left, or relationally, right) to the standard and the other is more different from the standard (shown on the top in each figure). (For illustration purposes, in each example trial the left option is more similar, and the right option is more different.) With respect to featural trials, the option on the left is more similar to the standard than the option of the right in virtue of having objects of the same shape as the standard (i.e., apples) in the visual example or semantically associated words (i.e., door-related words) in the verbal example. With respect to relational trials, the option of the left is similar to the standard than the option on the right in virtue of a uniquely shared visuospatial relation (i.e., A-B-B sequence) in the visual example or semantic relation (i.e., *located-in*) in the verbal example. Right panel: Ambiguous trials. On ambiguous trials, one option is more featurally similar to (and relationally different from) the standard shown on the top, whereas the other option is more relationally similar (and featurally different). The left option is featurally similar to the standard in virtue of sharing an object of the same shape (i.e., rounded square) in the visual example or semantically associated words (i.e., bug-related words) in the verbal example. The right option is relationally similar to the standard in virtue of a uniquely shared visuospatial relation (i.e., *same*) in the visual example or semantic relation (i.e., *located-in*) in the verbal example.

both options were equally featurally similar to the standard. Unambiguous trials enabled us to compare the difficulty of incorporating featural and relational information in similarity and difference judgments. Failure to select the similar option on featural trials would reflect a difficulty with incorporating featural similarity, whereas failure to select the similar option on relational trials would reflect a difficulty with incorporating relational similarity.

The remaining 18 trials were *ambiguous* trials, consisting of one option that was more featurally similar to (and relationally different from) the standard than the other option, which was more relationally similar to (and featurally different from) the standard (see right panel of Fig. 1 for examples). We refer to these trials as *ambiguous* because they were constructed so that selecting either option was reasonable, depending on a participant's criteria for judging similarity or difference. We used these trials to compare participants' preferential weighting of featural or relational information in their similarity and difference judgments. Selecting the featurally similar option as more similar indicates a preferential weighting of featural information, whereas selecting it as more different indicates a preferential weighting of relational information, and vice versa for selecting the relationally similar option.

For the verbal comparison task, featural similarity was determined by the semantic similarity among the individual words in each word pair. For instance, the bottom-right panel of Fig. 1 shows an example of an ambiguous trial for the verbal task. The individual words composing the standard (*bee* and *hive*) and those composing the right option (*butterfly* and *wing*) all refer to concepts related to garden plants, and thus are more semantically similar than the words composing the left option (*plane* and *airport*), which are generally less semantically similar to those in the standard. As a first approximation, semantic similarity was manipulated using qualitative judgments based on experimenter intuition; however, later in this paper we present computational simulations that incorporate data-driven measures of similarity, which serve as a validation check.

Relational similarity was determined by the semantic relation instantiated by each word pair. Referring again to the bottom-right panel of Fig. 1, the standard (*bee:hive*) and the left option (*plane:airport*) both instantiate the semantic relation *located-in*, and are thus more relationally similar to each other than the standard is to the right option (*butterfly:wing*), which most saliently instantiates an *object:part* relation (which does not match the standard's relation). In addition to *located-in* and *object:part* relations, verbal comparison trials featured *antonym* (e.g., *love:hate*), *synonym* (e.g., *big:large*), *category coordinate* (e.g., *broom:mop*), and *instance-of* (e.g., *shrub:bush*) relations. All featurally similar options in the verbal comparison task saliently instantiated one of the six relations listed above. On one trial, for example, participants were given the standard *hoof:horse* and asked to choose between the featurally similar option *goat:cow* and the relationally similar option *wheel:bicycle*. All three word pairs form representative examples of a semantic relation (either *part-of* or *category coordinate*). As with semantic similarity, relational similarity was also manipulated based on experimenter intuition; however, as mentioned above, computational simulations presented later in this paper use data-driven measures of similarity that serve as a validation check.

For the visual comparison task (Fig. 1, top panel), featural similarity was determined by a shared salient visual feature among individual objects, either *shape* or *shading* (filled / black or unfilled /white). Relational similarity was determined by the visual relation instantiated by each set of shapes. Most of the visual comparison trials were comparable to the one presented in the top-right panel of

Fig. 1, where the standard and the relationally similar option (right) instantiated the *same* relation and each consisted of repetitions of different shapes, while the featurally similar option (left) violated the standard's *same* relation and shared one object with the same shape as the standard. Other visual relations featured in this task included *symmetry*, consisting of two identical objects reflected about a vertical axis; *ABA sequences* consisting of three objects, of which the first and last were identical to each other; *ABC sequences* consisting of three unique objects; and *AABB sequences* consisting of two repetitions of different objects.

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

All participants completed a verbal comparison task and a visual comparison task in a counterbalanced order, and then completed the Ravens Progressive Matrices. The median duration of the entire experimental session was 9.91 min.

4. Results and discussion

4.1. Performance on unambiguous trials

Performance on unambiguous trials across conditions is depicted in Fig. 2. For participants making similarity judgments, accurate responding consisted of selecting the more similar option to the standard; and for participants making difference judgments, accurate responding consisted of selected the more different option. Overall, participants performed well on unambiguous trials. Those making similarity judgments ($n = 98$) more frequently selected the more similar option for both the verbal task ($M_{sim} = .80$, $SD_{sim} = .17$) and the visual task ($M_{sim} = .86$, $SD_{sim} = .14$). Those making difference judgments ($n = 86$) more frequently selected the more different option across both tasks (verbal: $M_{diff} = .77$, $SD_{diff} = .21$; visual: $M_{diff} = .77$, $SD_{diff} = .22$). Hereafter, we refer to the responses described above as 'accurate' responses. Of particular interest was the relative accuracy with which similarity and difference participants completed relational trials.

We hypothesized that assessing relational difference is more overtly cognitively demanding than assessing relational similarity, and so we predicted that participants making difference judgments would perform less accurately on relational trials than those making similarity judgments. On the other hand, we were uncertain as to whether the processing required by accurate performance on featural trials would be sufficiently cognitively demanding to reliably reveal an advantage for similarity judgments over difference judgments, and so we did not make strong predictions about a performance difference across similarity and difference for featural trials. Section 2 of our Supplemental Materials details analyses of human similarity norms for Experiment 1 stimuli. These analyses show that whereas human-rated relational similarity clearly discriminates between 'similar' and 'different' options to a greater extent than does human-rated featural similarity for both verbal and visual stimuli used on relational trials, only for verbal stimuli used on featural trials does human-rated featural similarity discriminate between 'similar' and 'different' options to a greater extent than does human-rated relational similarity. Our manipulation of similarity on relational trials is thus on firmer ground than is our manipulation of similarity on featural trials, and so we refrain from speculating about participant performance on featural trials.

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

We used likelihood-ratio tests to compare this full model to reduced models that each omitted a term of interest but that were otherwise equivalent to the full model. First, we tested whether performance generally differed across verbal and visual tasks. To do so, we fit a reduced model to the data that lacked the *modality* term but that was otherwise equivalent to the full model. We used a likelihood ratio test to compare the full model to the reduced model and found that removing the *modality* term did not increase model prediction error, $\Delta AIC = -1.40$, $\chi^2\,(1) = .65$, $p = .420$. We did not make any predictions about differences across modalities, and this result indicates that verbal and visual tasks did not differ in their overall difficulty.

Next, we tested the prediction of the processing-demand hypothesis that relational trials would be more difficult for participants judging difference than for those judging similarity. In order to do to so we compared the full model of unambiguous trials to a reduced model that removd the *comparison type x trial type* interaction term (but that retained the individual terms for *comparison type* and *trial type*). Dropping the interaction term did increase model prediction error, $\Delta AIC = 10.7$, $\chi^2\,(2) = 14.66$, $p < .001$, indicating that performance differences between participants making similarity judgments and difference judgments varied across featural and relational trials. To examine this interaction further, we used the *emmeans* and *pairs* functions from version 1.8.4 of the emmeans R package (Lenth, 2023) to compare the relevant estimated marginal means of the full model. Across verbal and visual tasks, similarity participants ($M = .81$, $SD = .18$) outperformed difference participants ($M = .69$, $SE = .22$) on relational trials, $z = 4.81$, $p < .001$, but not on featural trials, $z = .04$, $p = .966$ (similarity: $M = .84$, $SD = .14$; difference: $M = .84$, $SD = .20$). This result supports the prediction

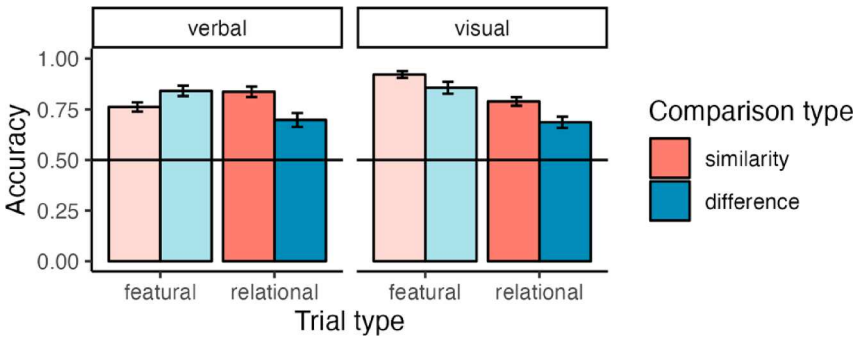


Fig. 2. Human accuracy on unambiguous trials for verbal and visual tasks (Experiment 1). Accuracy is broken down by trial type (featural vs. relational) and comparison type (difference vs. similarity). Error bars indicate ± standard error of the mean, and horizontal line indicates chance performance.

that difference judgments involve more complex comparisons than do similarity judgments, particularly impacting relational trials.[3]

A likelihood ratio test comparing the full model and a reduced model that lacked the *RPM score* term showed that removing that term increased model prediction error, $\Delta AIC = 13.5$, χ^2 (1) = 15.56, $p < .001$. Notably, the finding that similarity participants outperformed difference participants on relational trials persisted even after we accounted for individual differences in reasoning ability by including *RPM score* as a covariate in the full model. Thus, even though general reasoning ability indeed influenced performance on unambiguous trials, comparison type impacted performance specifically on relational trials, over and above individual differences in this ability.

4.1.1. Relational responding on ambiguous trials

We will now discuss response patterns on ambiguous trials in depth. Having confirmed that on unambiguous trials assessing relational difference was more difficult than was assessing relational similarity, we went on to examine participants' preferential weighting of featural and relational information in ambiguous comparisons for which the two kinds of information are pitted against each other. Fig. 3 presents item-level response rates for which the relationally similar option was selected as more similar (X-axis) and as more different (Y-axis) for verbal (left) and visual (right) comparisons. As described by Medin et al. (1990), if participants responded symmetrically across similarity and difference judgments, datapoints would lie along the solid diagonal. Instead, all points lie above the diagonal, with participants selecting the relationally similar option more often regardless of whether they were judging similarity ($M = .61$, $SD = .29$) or difference ($M = .62$, $SD = .26$). Notably, selecting this option implies different criteria based on comparison type: Selecting the relationally similar option as more similar implies an emphasis on *relational* similarity, whereas selecting that same option as more different implies an emphasis on *featural* difference.

In order to assess participant responses across comparison types (similarity vs. difference), we grouped responses according to whether they indicated an emphasis on *relational* information (see Fig. 4). We thus compared responses in which similarity participants selected the relationally similar option and in which difference participants selected the featurally similar option, and refer to these as *relational* responses.

As with unambiguous trials, we fit logistic mixed-effects models to predict relational responses on ambiguous trials. We defined a full model including *participant* and *comparison problem* as random intercept effects; *modality* (*verbal* vs. *visual*), *comparison type* (*similarity* vs. *difference judgments*) as fixed effects; and *RPM score*, *task order* (*verbal first* vs. *visual first*), and *trial number* as covariates. As was done for unambiguous trials, we used likelihood-ratio tests to compare this full model to reduced models that omitted a term of interest but that were otherwise equivalent to the full model. First, we compared the full model to a reduced model omitting the *comparison task* term. We found that dropping this term did not reduce model prediction error, $\Delta AIC = -2.0$, χ^2 (1) = .01, $p = .930$. This result indicates that relational responding did not differ across verbal and visual modalities.

Next, we compared relational response rates for similarity judgments and difference judgments, to test our main prediction that participants will preferentially weight relational information more heavily when judging similarity than when judging difference. Indeed, dropping the *comparison type* term from the full model did increase prediction error, $\Delta AIC = 33.3$, χ^2 (1) = 35.31, $p < .001$, confirming the prediction that relational response rates were affected by comparison type on ambiguous trials. This effect on ambiguous trials held even after we accounted for individual differences in reasoning ability by including *RPM score* as a covariate in the full model. Omitting *RPM score* from the full model also increased model prediction error, $\Delta AIC = 2.6$, χ^2 (1) = 4.60, $p = .032$. Thus,

[3] Note that task instructions did not impose any speed pressure, so response time was not a reliable measure of cognitive processing. However, we found no evidence of a speed-accuracy tradeoff, which would imply slower responses in the *similarity* condition and faster responses in the *difference* condition. In fact, we found a numerical difference in response time in the opposite direction: Participants tended to be slower to make accurate difference judgments ($M = 3.33$ s, $SD = 1.46$ s) than similarity judgments ($M = 3.06$ s, $SD = 1.26$ s). This pattern of response times persisted on ambiguous trials: Regardless of what option participants selected, they tended to be slower to make difference judgments ($M = 3.12$ s, $SD = 1.35$ s) than similarity judgments ($M = 2.97$ s, $SD = 1.15$ s).

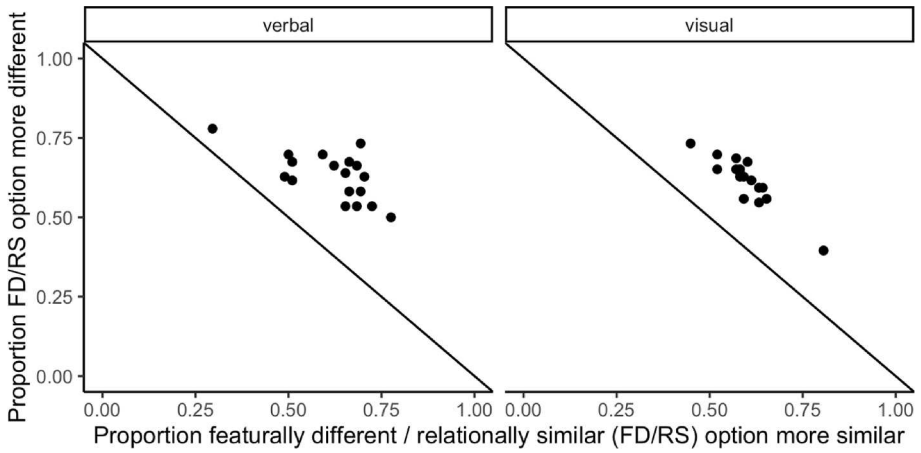


Fig. 3. Item-level plot of response rates for verbal (left) and visual (right) comparisons (Experiment 1). X-axis shows the proportion of participants selecting the relationally similar (and featurally different) option as more similar to the standard, and Y-axis shows the proportion of participants selecting this same option as more different from the standard.

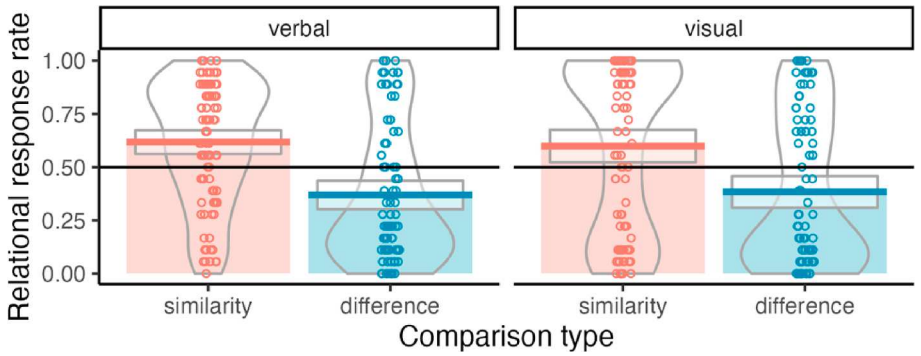


Fig. 4. Participant-level relational response rate on ambiguous trials for verbal and visual comparison tasks (Experiment 1). Response rates are broken down according to comparison type (difference vs. similarity). Unfilled circles each indicate an individual participant's response rate, dark lines indicate mean response rates, box boundaries indicate ± standard error of the mean, and horizontal line corresponds to indiscriminate selection of relational versus featural options.

even though individual differences in reasoning ability predicted relational responses on ambiguous trials, our manipulation of comparison type impacted responses over and above these individual differences in general reasoning ability.

In summary, for both visual and verbal comparisons, Experiment 1 showed that (1) human reasoners have greater difficulty processing relational difference than they do relational similarity, and (2) they tend to weight relational information more heavily when judging similarity than when judging difference. Overall, the present results provide convergent evidence for the claim that assessments of difference are more cognitively demanding than assessments of sameness (Hochmann, 2021; Hochmann et al., 2016, 2018).

4.2. Modeling and validation check for experiment 1

Experiment 1 used experimental materials that manipulated featural and relational similarity as initially determined by experimenter intuition. In order to provide a validation check for these stimuli, we formalized the human comparison process as construed by the processing-demand account in a computational model that operates over data-driven measures of similarity. Here we describe two sets of simulations of human performance on the experimental task in Experiment 1. The first set of simulations uses human ratings of featural similarity and of relational similarity. The second set of simulations uses automatically-generated measures of similarity derived from models of word representation, Word2vec (Mikolov et al., 2013), and relation representation, BART (Lu et al., 2019). In both sets of simulations, we reproduced the asymmetry in similarity and difference judgments observed in Experiment 1, thus providing evidence that our stimuli vary featural and relational similarity as intended.

In accord with the model of analogical mapping developed by Lu et al. (2022), the present model includes a weighting mechanism that controls the relative contribution of relational and featural information to a comparison judgment. We model the human comparison process as a weighted sum of featural similarity between two items i and j, $sim_{feat_{ij}}$, and their relational similarity, $sim_{rel_{ij}}$,

$$sim_{ij} = (1 - \alpha_{sim})sim_{feat_{ij}} + \alpha_{sim}(sim_{rel_{ij}}) \quad (1)$$

$$diff_{ij} = -(1 - \alpha_{diff})sim_{feat_{ij}} - \alpha_{diff}(sim_{rel_{ij}}) \quad (2)$$

where α_{sim} and α_{diff} are free parameters that respectively reflect the degree to which a similarity judgment and a difference judgment each weight relational information (where higher values of α imply greater emphasis on relations). We refer to α_{sim} and α_{diff} as relation-weight parameters. Note that both similarity, sim_{ij}, and difference, $diff_{ij}$, are based on identical computations that produce a weighted sum of featural and relational similarity: Difference judgments simply negate the output of that computation.

Crucially, while α_{sim} and α_{diff} jointly provide this model with a means to *reflect* the differential weighting of featural and relational information in similarity and difference judgments, the model offers no *explanation* for this differential weighting. We emphasize this in order to clarify that the goal of the current set of simulations is to use data-driven measures of featural and relational similarity to reproduce key empirical phenomena in Experiment 1 and provide a validation check of our stimuli. Specifically, we fit our model to individual participant data. The fitted parameters enabled the model to reproduce the pattern of results found in Experiment 1, in which on ambiguous trials participants making similarity judgments selected the 'relational' option (i.e., the option that was more relational similar and more featurally different from the standard) more often than did participants making difference judgments; whereas on relational unambiguous trials similarity participants more often selected the 'similar' option (i.e., the option that more similar to the standard, in terms of either featural similarity on 'featural' trials or in terms of relational similarity on 'relational' trials) than did difference participants. Finally, we predicted that when fit to human judgments at the level of individual participants, the fitted parameter α_{sim} will be greater than α_{diff}, reflecting the greater impact of relations on assessments of similarity than on assessments of difference.

Recall that a given trial involves three word pairs: a standard and two options, A and B. Each trial minimally involves two comparisons: One between the standard and option A and the other between the standard and option B. In order to simulate participant behavior on this two-alternative forced-choice task, we computed the similarity (or difference) for each of these two comparisons (standard vs. option A and standard vs. option B). The model's choice on a given trial was defined to be whichever option (A or B) for which the comparison with the standard yielded a higher similarity (or difference) value.

As mentioned above, we ran two sets of simulations on data-driven measures of featural and relational similarity: One based on human ratings and one on model-based measures. In the next section, we describe how we collected human similarity ratings; we describe the model-based measures in the subsequent section. Finally, we describe how we fit our model to trial-level data using each set of similarity measures, and then use the fit model to reproduce the relative weighting of featural and relational similarity observed in Experiment 1.

4.2.1. Human similarity ratings of verbal and visual stimuli

Participants. Participants were 36 UCLA undergraduate students (M_{age} = 20.48, SD_{age} = 3.27, range = [18, 34]). This sample consisted of 33 female and 3 male participants. All participants completed experimental tasks online to obtain partial course credit in a psychology class. The study was approved by the Institutional Review Board at UCLA.

Materials and Procedure. Recall that in Experiment 1, each of verbal and visual stimulus sets consisted of 24 triad trials, including 6 unambiguous trials and 18 ambiguous trials. A given trial presented a standard and two response options. For unambiguous trials, a "similar" response option was intended to be more similar to the standard than the "different" option (which was intended to be more different from the standard). For ambiguous trials, a "featurally similar" response option was intended to be more featurally similar to but also more relationally different from the standard than the "relationally similar" option (which was intended to be more featurally different from but also more relationally similar to the standard). Each triad yielded two trials for the present norming study (48 norming trials for each modality), and we collected both featural similarity ratings and relational similarity ratings for each one of these norming trials. Both of these trials shared a common standard, but they varied which of the two options was paired with the standard.

Pilot testing showed that participants distinguished featural from relational similarity more clearly when the (presumably simpler) featural judgments were made before the relational judgments (especially for visual stimuli). Accordingly, judgments were made in two blocks, using a fixed order of featural similarity followed by relational similarity. On each norming trial, the standard from one of the original Experiment 1 triads was presented above one of two response options from that same triad, and participants were asked to rate the similarity between the two stimuli by clicking on a slider from 0 ("not at all similar") to 100 ("completely similar"). In the first block of trials, all participants were instructed to rate featural similarity between pairs of stimuli by comparing how similar the individual words or objects in the top stimuli were to the individual words or objects in the bottom stimuli. They were reminded of these instructions on each trial. In the second block, participants were asked to rate the relational similarity between pairs of stimuli by comparing how similar the relation between words or objects in the top stimuli was to the relation between words or objects in the bottom stimuli. They were again reminded of these instructions on each trial. While the order of similarity judgments (featural versus relational) was fixed, modality order (verbal stimuli first versus visual stimuli first) was counterbalanced across participants; thus half of our sample judged visual-featural similarity before judging verbal-relational similarity, whereas the other half judged verbal-featural similarity before judging visual-relational similarity. Importantly, this counterbalancing ensured that each individual participant made only one similarity judgment (either featural or relational) for any single triad. Task instructions are available at this paper's OSF link: https://osf.io/szqjk.

Prior to completing each block, participants were given two example trials. Before blocks consisting of verbal comparisons,

participants were shown a pair of highly similar word pairs (*earthquake: destruction*; *drought: famine*) and a pair of much less similar word pairs (*earthquake: destruction*; *glass: fragile*). Fig. 5 shows the corresponding examples used to clarify visual comparisons, a pair of highly similar object-pairs (Fig. 5, left panel) and a less similar pair of object-pairs (Fig. 5, right panel). Depending on whether participants were asked to judge featural similarity or relational similarity, task instructions respectively emphasized the presence or absence of semantic association between individual words (e.g., natural disasters) or visual similarity of individual objects (e.g., crosses); or else the semantic relations between words (e.g., *cause-effect* relations) or visual relations between objects (e.g., *smaller-than*). Section 2 of Supplementary Information presents direct analyses of these ratings. In general, human similarity ratings cohered with the intended manipulations of featural and relational similarity for both unambiguous and ambiguous stimuli, with the exception of visual stimuli used for featural unambiguous trials. Because we did not make strong predictions about performance on featural trials, we were not concerned by this result. In order to generate similarity measures to serve as input to our weighted-sum model, we took the mean featural similarity and the mean relational similarity rated for each unique trial, and these norms are available at this paper's OSF link: https://osf.io/szqjk/.

4.2.2. Model-based similarity of verbal stimuli

We will now describe how computational models of word and relational representation were used to generate data-driven measures of featural and relational similarity for verbal stimuli. In order to represent individual word meanings, we used pre-trained word embeddings generated by Word2vec (Mikolov et al., 2013), a machine-learning model that represents word meanings as high-dimensional vectors of length 300. These vectors are based on the hidden layer of activation within a neural network trained to predict patterns of text in sequence as they appear in a large corpus consisting of Google News articles (about 100 billion words). Although these types of word embeddings are solely derived from the statistical distribution of texts in their training corpora, they have been shown to preserve the similarity structure of individual word meanings in a psychologically realistic way. These embeddings have been used to successfully model a number of cognitive processes beyond similarity judgments, including human memory search, categorization, and decision making (Bhatia & Aka, 2022; Günther et al., 2019).

To compute featural similarity, the meaning of a word pair is represented as a concatenation of the semantic vectors of the two individual words. In order to equate the contribution of each individual semantic vector to the concatenated vector, we normalized each Word2vec vector according to its L2 norm prior to concatenation. We use A to denote the first word in a word pair and B to represent the second word in a word pair, and we compute the featural similarity between two word pairs i and j as the cosine similarity between concatenated word vectors constituting i, $[f_{A_i}\ f_{B_i}]$ and those constituting j, $[f_{A_j}\ f_{B_j}]$ (see top panel of Fig. 6):

$$sim_{feat_{ij}} = cos\left(\left[f_{A_i}\ f_{B_i}\right], \left[f_{A_j}\ f_{B_j}\right]\right) \tag{3}$$

This model-derived measure of featural similarity correlated with the human featural similarity ratings described in the previous section ($\rho = .67$, $p < .001$) but not the human ratings of relational similarity ($\rho = -.21$, $p = .15$).

A basic requirement for computing relational similarity is a mechanism to accomplish the *eduction of relations* (Spearman, 1923): generating representations of the unstated semantic relations linking paired entities (e.g., *part-of* for *finger:hand*). To instantiate this mechanism, we used BART, a learning model that has been used to predict human analogy performance and graded judgments of relational similarity (Ichien et al., 2022; Lu et al., 2012, 2019). BART assumes that specific semantic relations between words are coded as distributed representations over a set of abstract relations. The BART model takes concatenated pairs of Word2vec vectors as input, and then uses supervised learning with both positive and negative examples to acquire representations of individual semantic relations. We use a version of BART that was trained on two datasets consisting of human-generated word pair examples, which were used to learn a total of 270 semantic relations.

In the present simulations, we combined two datasets of human-generated word pairs to train BART. The first dataset (Jurgens

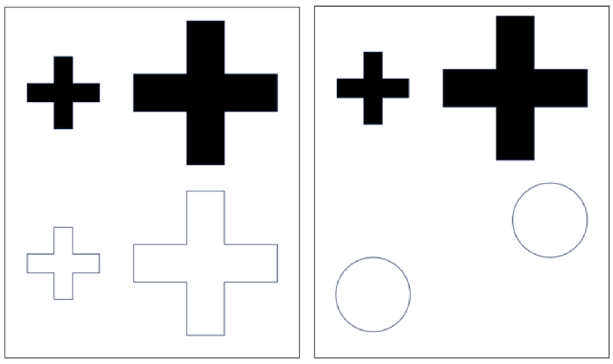


Fig. 5. Examples used to illustrate high visual similarity (left panel) or low visual similarity (right panel). Stimuli in the left panel are featurally similar because top and bottom stimuli are composed of objects of the same shape (i.e., crosses), and are relationally similar because they both instantiate a *smaller-than* relation. In contrast, stimuli in the right panel are featurally dissimilar in that they consist of distinct shapes, and are relationally dissimilar in that the bottom stimuli instantiate a spatial *diagonal-of* relation.

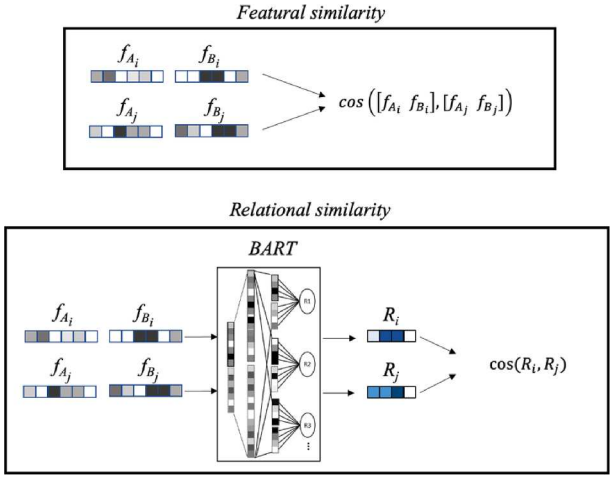


Fig. 6. Schematic computation of featural (top) and relational (bottom) similarity.

et al., 2012) consists of at least 20 word pairs (e.g., *engine: car*) instantiating each of 79 semantic relations. Each of these relations belongs to one of 10 broad relation types according to a taxonomy originally developed by Bejar et al. (1991): *class inclusion* (e.g., *X is a kind of Y*), *part-whole* (e.g., *X is a part of Y*), *similarity* (e.g., *X is a synonym of Y*), *contrast* (e.g., *X contradicts Y*), *attribute* (e.g., *X does action Y*), *nonattribute* (e.g., *something X cannot be Y*), *case relation* (e.g., *X makes Y*), *cause-purpose* (e.g., *X causes Y*), *space–time* (e.g., *X happens at Y*), and *reference* (e.g., *X indicates Y*). The second dataset consists of at least 10 word pairs instantiating each of 56 additional semantic relations (Popov et al., 2017). These relations were likewise organized into a relation taxonomy introduced in Ichien et al. (2020) and consisting of the following types: *function* (e.g., *X's job is to produce Y*), *constitution* (e.g., *X is made up of individuals Y*), *leadership* (e.g., *X mentors Y*), *opposite* (e.g., *X protects against Y*), *cover* (e.g., *X is an object whose lid is Y*), *cause* (e.g., *X develops into Y*), *part-whole* (e.g., *X is an appendage of Y*), *location* (e.g., *X is an artifact located in Y*), *measurement* (e.g., *X is a unit of Y*). Across both datasets, BART acquired weight distributions for 135 semantic relations. BART can automatically generate representations of the converse of each learned relation by swapping the relation weights associated with each individual relational role. Thus, upon learning a representation of *X is a category for Y*, BART can also form a representation of its converse, *Y is a member of category X* , effectively doubling its pool of learned relations from 135 to 270 in total.

After learning, BART calculates a relation vector consisting of the posterior probability that a word pair instantiates each of its learned relations. BART uses its pool of learned relations to create a distributed representation of the relation(s) between any two paired words *A* and *B*. The posterior probabilities calculated for all learned relations form a 270-dimensional relation vector R_{AB}, in which each dimension codes how likely a word pair instantiates a particular relation. The relational similarity between word pairs *i* and *j* is computed as the cosine similarity of the corresponding relation vectors R_i and R_j (see bottom panel of Fig. 6):

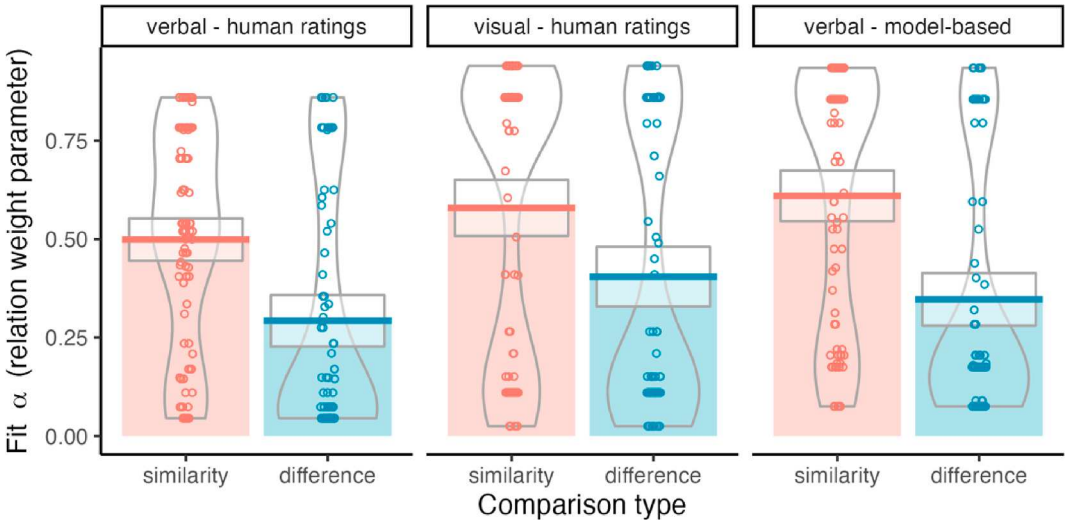


Fig. 7. Relation-weight parameter fitted to individual participant data as a function of comparison type. Different panels show the fitted results using human similarity ratings or model-derived similarity ratings for the verbal task.

$$sim_{rel_{ij}} = cos(R_i, R_j) \quad (4)$$

As with Word2vec vectors, we normalized each BART relation vector according to its L2 norm prior to computing cosine similarity. This model-based measure of relational similarity correlated with human relational similarity ratings ($\rho = .48, p < .001$) but not human featural similarity ratings ($\rho = -.03, p = .85$).

We used the model described in Equations (1) to generate trial-level predictions for participants who made similarity judgments in Experiment 1, and that described in Equation (2) to generate predictions who participants made difference judgments. We fit the relation-weight parameter to each participant's data by maximizing the accuracy with which the model predicted responses on all trials. If multiple values of the relation-weight parameter predicted a participant's data equally well, we took the mean of those parameter values. For both verbal and visual stimuli, the model that was fit using human ratings predicted participant responses similarly well across those in the similarity condition of Experiment 1 and those in our difference condition (*verbal-similarity*: $M_{Acc} = .81$; $SD_{Acc} = .11$; *verbal-difference*: $M_{Acc} = .77$; $SD_{Acc} = .11$; *visual-similarity*: $M_{Acc} = .87$; $SD_{Acc} = .09$; *visual-difference*: $M_{Acc} = .84$; $SD_{Acc} = .12$). This was also the case with the model that was fit to verbal data using model-derived measures of featural similarity from Word2vec, and relational similarity from BART (*verbal-similarity*: $M_{Acc} = .72$; $SD_{Acc} = .10$; *verbal-difference*: $M_{Acc} = .79$; $SD_{Acc} = .12$).

We predicted that the value of the relation-weight parameter would be greater when fit to participants making similarity judgments than when fit to those making difference judgments. Fig. 7 shows the distribution of the parameter, broken-down according to comparison type. Mann-Whitney U tests confirmed what is clear from visual inspection: Fit relation-weight parameters were reliably greater for participants who made similarity judgments than for those who made difference judgments (*verbal – human ratings*: $W = 5822$, $p < .001$; *visual – human ratings*: $W = 5224$, $p < .001$; *verbal – model-based*: $W = 6092$, $p < .001$). Moreover, the value of the fit relation-weight parameter predicted the rate with which participants selected relational options on ambiguous trials, both in the similarity condition of Experiment 1 (*verbal*: $\rho = .90$, $p < .001$; *visual*: $\rho = .88$, $p < .001$; *model-based*: $\rho = .83$, $p < .001$,), and in the difference condition (*verbal*: $\rho = .88$, $p < .001$; *visual*: $\rho = .79$, $p < .001$; *model-based*: $\rho = .71$, $p < .001$). This computational result supports the validity of our manipulation of featural and relational similarity and further supports our central claim: Similarity judgments elicit greater reliance on relational information than do difference judgments.

5. Experiment 2

Beyond the special emphasis that structure mapping theory places on alignability, a more general difference between that account and our processing-demand account of comparison involves the processing stage at which each explanation locates the dissociation between similarity and difference comparisons. Structure-mapping theory proposes that judgments of both similarity and difference involve an identical comparison process—structural alignment—which consistently operates over the same representations (i.e., representations of relational structure) (Gentner, 1983; Gentner & Markman, 1994). Any divergence between similarity and difference judgments is then attributed to asymmetries in use of the *output* of the comparison process (i.e., all commonalities versus alignable differences only). In contrast, the processing-demand hypothesis proposes that comparisons of similarity and difference operate on distinct representations: comparisons of similarity tend to operate on representations that incorporate more relational information than do comparisons of difference. Thus, the present explanation locates the dissociation between similarity and difference judgments observed in Experiment 1 in the representations over which comparison operates.

Structure-mapping theory and our processing-demand hypothesis thus make distinct predictions about the extent to which asymmetries in similarity and difference judgments reflect the representations compared in order to arrive at those judgments. Whereas the processing-demand hypothesis proposes a direct link between this response asymmetry and the representations compared, structure-mapping theory proposes no such link. We assessed these competing hypotheses in Experiment 2.

The stimuli used in Experiment 1 were relatively simple, consisting of pairs of words or geometric forms. To determine whether the asymmetry between judgments of similarity versus difference can also be obtained with more complex stimuli, in Experiment 2 we used naturalistic story stimuli created by Gentner et al. (1993). These stimuli consist of story sets, each including one story that is analogous to a standard story by sharing similar plot structures and event relations but with particularly dissimilar entities, and another story that is disanalogous but superficially similar to the standard in terms of using similar entities and topics included in the story. We used these sets to respectively emphasize relational and featural similarity in the two response options constituting the same type of triad task as in Experiment 1.

For such complex stimuli, significant processing is required to read each story and generate a stable representation of its meaning. The processing-demand hypothesis locates the dissociation between similarity and difference judgments observed in Experiment 1 in the representations over which comparison operates. For complex stimuli that require extended processing, it may be possible to dissociate the effective representations from the type of comparison. Accordingly, in addition to manipulating different type of comparison judgments (similarity vs. difference), in Experiment 2 we manipulated whether or not participants were prompted to process individual stories *before* comparing them. This manipulation was intended to vary the extent to which the processes involved in generating stimulus representations occur separately from or simultaneously with comparison (with these processes being more separated in participants given a pre-comparison processing step and more simultaneous in participants lacking that step). The key assumption is that relations (which are more cognitively demanding) will benefit from a pre-comparison step that allows relations to be extracted from the inputs without simultaneously requiring comparisons.

Under the processing-demand hypothesis, differences among representations subserving similarity and difference judgments should be diminished for participants who are prompted to generate representations of stimuli *prior* to making a comparison, relative

to those not prompted to do so. Specifically, for difference judgments, a reasoner is more likely to represent stimuli relationally when individual stimuli are processed *prior* to comparing them, relative to when the stimuli are first processed *during* comparison. This hypothesis thus predicts that the asymmetry in similarity and difference judgments observed in Experiment 1 will be found *only* for participants who do not receive a pre-processing step, as they must generate stimulus representations while also comparing them. In contrast, structure-mapping theory assumes that asymmetries in similarity and difference judgments do not reflect differences in the representations over which comparison operates, and hence predicts that manipulating the point at which stimulus representations are formed will have no effect on response patterns.

5.1. Method

5.1.1. Participants

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

5.1.2. Materials and procedure

Comparison task. Participants completed a story comparison task, in which they were asked to compare sets of three story stimuli drawn from Gentner et al. (1993); for examples see Table 1. As in Experiment 1, participants were asked to compare a standard to a relational option and a featural option in order to select which was more similar or else which was more different. The relational option (labeled "analogy match" in the original materials) consisted of characters (e.g., a bird and a hunter versus a pair of nations) and individual events (e.g., gift of feathers versus gift of supercomputers) that were superficially dissimilar to those in the standard but that played roles in an overall plot structure that matched the standard (e.g., an act of kindness leads to a reciprocal act of kindness). In contrast, the featural option (labeled "mere-appearance match" in the original materials) consisted of characters and individual events that were superficially similar to those in the standard but that played roles in different plots structures from the standard (e.g., an act of kindness fails to elicit a reciprocal response, in contrast to the plot structure mentioned above). Task instructions and all stimuli used in the study can be found at this paper's OSF link: https://osf.io/szqjk/.

In total, participants completed 18 trials of this task. On each trial, participants were presented with the standard at the top of the screen and were instructed to read the story carefully. They were given 10 s before they could proceed to see the two options. Participants were randomly assigned to one of two presentation conditions, an incremental-options condition and a simultaneous-options condition, which differed in the way that the two options were presented (see Fig. 8). In the simultaneous-options condition, the standard was presented first and participants were required to read it for at least 10 s. Then the two options were presented at the same time, directly after participants had proceeded from reading the standard., Participants were instructed that once the two options were revealed, they were to compare them and judge which option was more similar to or more different from the standard.

In the incremental-options condition, the standard was first presented in the same manner as for the simultaneous-options condition. Then the two options were revealed incrementally, one at a time. After being given at least 10 s to read the standard, participants were given at least 10 more seconds to read one option on the left side of the screen, before they could proceed to read the option on the right side of the screen for at least another 10 s. After having read all three stories, participants were finally asked to enter their responses as to which option was more similar to or different from the standard.

For both conditions, whether the relational or featural option appeared on the right or left side of the screen was randomized across trials; once presented, each story remained on the screen for the rest of the trial. The incremental-options condition thus gave participants an opportunity to process each option in isolation before comparing them to the standard. In contrast, the simultaneous-options condition required participants to process each option while comparing them to the standard. Crossing presentation condition (simultaneous-options versus incremental-options) with decision type (similarity versus difference) yielded four conditions; both factors were manipulated between subjects. We note that given the high cognitive demands imposed by the story comparison task used in Experiment 2, we omitted any further tasks, including the Ravens Progressive Matrices task used in Experiment 1. The median duration of the entire experimental session was 13.10 min.

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

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

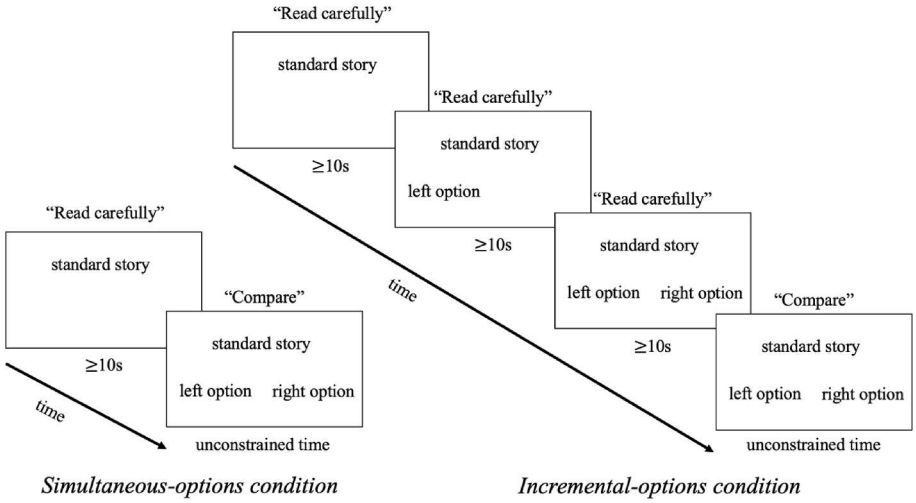


Fig. 8. Trial structure of story comparison task. In the simultaneous-options condition (left), left and right options were presented simultaneously and participants were asked to compare them upon seeing them. In the incremental-options condition (right), each story was presented one at a time.

6. Results and discussion

In general, participants assigned to the incremental-options presentation condition spent more time on each trial (*similarity*: $M_{RT} = 52.26$ s, $SD_{RT} = 16.77$ s; *difference*: $M_{RT} = 59.92$ s, $SD_{RT} = 24.53$ s) than did those in the simultaneous-options condition (*similarity*: $M_{RT} = 45.04$ s, $SD_{RT} = 22.16$ s; *difference*: $M_{RT} = 43.05$ s, $SD_{RT} = 22.56$ s). To test this, we fit a linear mixed-effects model of trial times, using the *lmer* function from version 1.1.26 of the LME4 R package (Bates et al., 2015) in R version 4.1.1 (R. Core Team, 2021). This model included *participant* and *comparison problem* as random intercept effects, *presentation condition* as a fixed effect, and *response* and *trial number* as covariates. A likelihood-ratio test comparing this model with an otherwise equivalent model that omitted the *presentation condition* parameter confirmed the reliability of this difference in trial times ($\Delta AIC = 10.0, \chi^2(1) = 11.84, p < .001$). This relative lag among incremental-options participants is sensible since they (but not participants in the simultaneous-options condition) had to spend at least 10 s reading each option individually. The fact that incremental-options participants spent more time on each trial than did simultaneous-options participants helps alleviate concerns that simultaneous-options participants may have incidentally adopted a similar approach to story processing as those in the incremental-options condition.

To analyze responses on the story comparison task, we fit a logistic mixed-effects model to human performance on this comparison task, using the *glmer* function from version 1.1.26 of the LME4 R package (Bates et al., 2015) in R version 4.1.1 (R. Core Team, 2021). We defined a full model including *participant* and *comparison problem* as random intercept effects; with a *presentation condition* (*simultaneous-options* vs. *incremental-options*) x *comparison type* (*similarity* vs. *difference judgments*) interaction term as a fixed effect, as well as *trial number* as a covariate to account for any systematic change in strategy across trials within a task.

We used a likelihood-ratio test to compare this full model to reduced models that omitted the *presentation condition* × *comparison task* interaction term but that was otherwise equivalent to the full model. As predicted by the processing-demand hypothesis, removing

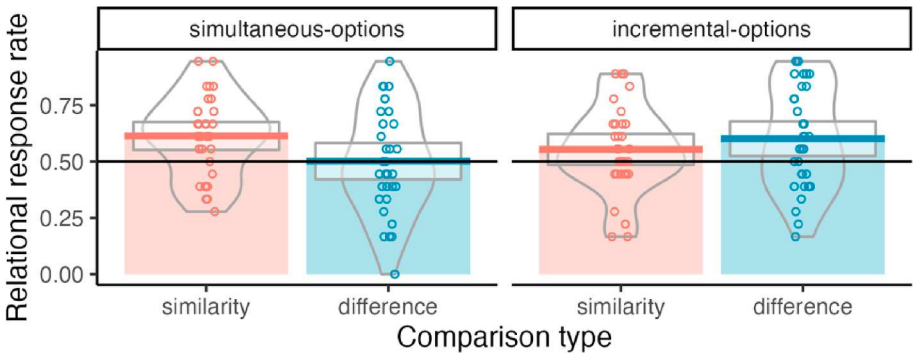


Fig. 9. Relational response rate for story comparison task (Experiment 2) in simultaneous-options (left) and incremental-options (right) conditions. Response rates are broken down according to comparison type (similarity vs. difference). Unfilled circles each represent an individual participant's response rates, dark lines indicate mean response rates, box boundaries indicate ± standard error of the mean, and horizontal line corresponds to indiscriminate selection of relational versus featural options.

the interaction term increased model prediction error, $\Delta AIC = 2.80, \chi^2\,(1) = 4.82, p = .03$. This result confirms that, as predicted by our processing-demand account, whether or not participants processed story stimuli in a pre-comparison step had an impact on the response pattern among similarity and difference judgments (see Fig. 9).

In order to test the difference between relational responding in similarity and difference judgments for each group, we compared the relevant estimated marginal means of the full model, using the *emmeans* and *pairs* functions from version 1.8.4 of the emmeans R package (Lenth, 2023). Participants in the simultaneous-options condition who made similarity judgments ($M = .61$, $SD = .18$) had higher rates of relational responding than those who made difference judgments ($M = .51$, $SE = .23$) when asked to simultaneously read and compare stories to the standard;; $z = 2.11, p = .035$; left panel of Fig. 6). However, this difference did not hold for participants in the incremental-options condition who were asked to read and then compare stories to the standard (similarity: $M = .55$, $SD = .19$; difference: $M = .60$, $SD = .23$; $z = 1.02$, $p = .307$; right panel of Fig. 6). This result confirms the prediction of the processing-demand hypothesis that processing story stimuli *during* comparison elicited asymmetric responding. This difference was eliminated when participants were given an opportunity to read and process each story prior to comparing them. As predicted by the processing-demand hypothesis (but not structure mapping theory), response differences in the simultaneous-options condition reflected differences in stimulus processing involved in comparing similarity versus comparing difference.

7. General discussion

Across a wide range of stimulus types (word pairs, sets of simple shapes, and stories), the present findings provide convergent evidence for the claim that assessments of similarity operate on distinct representations than do assessments of difference in that the former incorporate relational information more than do the latter. Our processing-demands hypothesis assumes that this dissociation is ultimately rooted in a processing asymmetry in comparisons assessing similarity and those assessing difference. Difference comparisons impose greater cognitive demands than do similarity comparisons, and a compelling explanation for these additional cognitive demands is that the former but not the latter involves processing negation (Hochmann, 2021; Hochmann et al., 2016, 2018). The joint cognitive complexity of processing relations (relative to features) and assessing difference (relative to similarity) impacts the way that human reasoners actually represent the items they compare. Because of the greater demand imposed by difference judgments, human reasoners represent the items about which they make this type of judgment in a more shallow or non-relational way. As demonstrated by the incremental-options condition of Experiment 2, this effect can be eliminated if complex stimuli (stories) are processed individually in advance of initiating the comparison process, so relations can be extracted prior to initiating the actual comparison task.

The present results argue against an alternative explanation that has been offered for the asymmetry between judgments of similarity versus difference. Under this alignment hypothesis (based on structure mapping theory; Gentner, 1983), assessments of similarity and difference both depend on analogical comparison and involve an identical process of structural alignment, in which representations of entity features and their structural relations are placed into one-to-one correspondence (Gentner & Markman, 1994; Markman, 1996; Markman & Gentner, 1993; Sagi et al., 2012). The asymmetry observed by Medin et al. (1990) is hypothesized to arise from an asymmetry in the relevant output of this comparison process. Whereas all commonalities contribute to similarity judgments, differences are split into *alignable* differences (i.e., those filling corresponding roles within a shared relational structure) and *non-alignable* differences (i.e., those not based on corresponding roles. Proponents of this explanation noted that the featurally-similar option in the study by Medin et al. (1990) did not involve a salient relation, so that any relational difference between it and the standard would not constitute an alignable difference, and hence was ignored in difference comparisons.

The results of the present study disconfirm the alignment hypothesis. In Experiment 1, all relational differences on the verbal task were clearly alignable. All word pairs instantiated one of six distinct semantic relations;[4] accordingly, all differences between relations were alignable. Structure mapping theory therefore predicts that mismatching relations (e.g., between *hoof:horse* and *goat:cow*) will contribute to difference judgments just as much as do mismatching features (Gentner & Markman, 1994; Markman, 1996). Accordingly, that theory predicts symmetric responding for similarity and difference judgments on our ambiguous trials: Participants should have selected all options with the same frequency, regardless of whether they were judging similarity or difference. But instead, we found clear asymmetries between the two types of judgments.

In addition, because the alignment account places differences between similarity and difference judgments solely in the output stage subsequent to structure mapping, it also does not explain our finding (Experiment 2) that processing story stimuli *before* comparison eliminated asymmetric responding. The present study did not directly test whether nonalignable differences contribute to difference judgments. However, when Estes and Hasson (2004) did precisely this—comparing the influence of alignable and nonalignable differences on comparison judgments—they showed not only that nonalignable differences impacted both similarity and difference judgments, but also that nonalignable differences actually had *greater*—not lesser—impact than did alignable differences. Together with these previous findings, the results of the present study disconfirm central predictions of the alignment hypothesis.

We acknowledge that some demonstrations of asymmetries between similarity and difference judgments are not obviously explained by the hypothesized processing asymmetry between similarity and difference judgments (Simmons & Estes, 2008; Tversky, 1977). In addition to considering features and internal structural relations of stimuli (our focus here), human reasoners also attend to external relations or thematic relatedness *between* stimuli (i.e., associations based on co-occurrence in some context, such as between *dog* and *leash*). There is evidence that people consider thematic relations when making similarity judgments (Bassok & Medin, 1997),

[4] For the visual comparison task used in Experiment 1, participants may not always have interpreted FS/RD options as instantiating a relation, so performance on this test does not constitute as strong a test of the alignment hypothesis as does the verbal comparison task.

but do so less often when making difference judgments (Golonka & Estes, 2009; Simmons & Estes, 2008). Galonka and Estes (2009) have argued this this type of asymmetry arises because thematic relatedness introduces commonalities between thematic associates without reducing the relevant differences between them. However, asking participants to complete a larger number of comparison trials (~60), and reminding participants of task instructions throughout the experimental session, has been shown to eliminate the effect of thematic relatedness on similarity judgments (Honke & Kurtz, 2019). Future work should aim to clarify the impact of thematic relatedness on similarity and difference judgments, and assess whether any persisting asymmetry between the two might also be explained in terms of a processing asymmetry between similarity and difference.

CRediT authorship contribution statement

Nicholas Ichien: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Nyusha Lin:** Writing – review & editing, Methodology, Investigation. **Keith J. Holyoak:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Hongjing Lu:** Writing – review & editing, Supervision, Resources, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

Preparation of this paper was supported by NSF Grant BCS-2022369. We thank Angela Kan and Jennifer Lo for help with stimulus generation and data collection. A preliminary report of part of this research was presented at the 45th Annual Conference of the Cognitive Science Society (July 2023).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cogpsych.2024.101661.

References

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

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

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

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

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

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

Carlson, G. N. (1987). Same and Different: Some Consequences for Syntax and Semantics. *Linguistics and Philosophy, 10*(4), 531–565.

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

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

Clark, H. H. (1971). *The chronometric study of meaning components*. Presented: CRNS Colloque International SUE les Probl6mes Actuels de Psycholinguistique, Paris.

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

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

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

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

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

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

Gerson, L. P. (2004). Plato on Identity, Sameness, and Difference. *The Review of Metaphysics, 58*(2), 305–332.

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

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

Grier, P. T. (2007). *Identity and Difference: Studies in Hegel's Logic, Philosophy of Spirit, and Politics*. State University of New York Press. doi: 10.1353/book5224.

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

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

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

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

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

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

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

Ichien, N., Lu, H., & Holyoak, K. J. (2020). Verbal analogy problem sets: An inventory of testing materials. *Behavior Research Methods, 52*(5), 1803–1816. https://doi.org/10.3758/s13428-019-01312-3

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

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

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

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

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

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

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

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

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

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

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

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

Moltmann, F. (1992). Reciprocals and "Same/Different": Towards a Semantic Analysis. *Linguistics and Philosophy, 15*(4), 411–462.

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

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

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

Seymour, P. H. K. (1969). Response Latencies in Classification of Word—Shape Pairs. *British Journal of Psychology, 60*(4), 443–451. https://doi.org/10.1111/j.2044-8295.1969.tb01217.x

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

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

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

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

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

Zentall, T. R., Andrews, D. M., & Case, J. P. (2018). Sameness may be a natural concept that does not require learning. *Psychological Science, 29*(7), 1185–1189. https://doi.org/10.1177/0956797618758669

Zentall, T. R., Edwards, C. A., Moore, B. S., & Hogan, D. E. (1981). Identity: The basis for both matching and oddity learning in pigeons. *Journal of Experimental Psychology: Animal Behavior Processes, 7*(1), 70–86. https://doi.org/10.1037/0097-7403.7.1.70