# Asymmetry in similarity and difference judgments results from asymmetry in the complexity of the relations *same* and *different*

**Nicholas Ichien**[1]
ichien@ucla.edu

**Nyusha Lin**[1]
nyushalin@gmail.com

**Keith J. Holyoak**[1]
kholyoak@g.ucla.edu

**Hongjing Lu**[1,2]
hongjing@ucla.edu

[1] Department of Psychology, [2] Department of Statistics
University of California, Los Angeles
Los Angeles, CA 90095 USA

**Abstract**

Explicit similarity judgments tend to emphasize relational information more than do difference judgments. We propose and test the hypothesis that this asymmetry arises because human reasoners represent the relation *different* as the negation of the relation *same*, so that processing difference is more cognitively demanding than processing similarity. For both verbal comparisons between word pairs, and visual comparisons between sets of geometric shapes, we asked participants to select which of two options 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 assessing relational similarity, we predicted that detecting relational *difference* would be particularly demanding. We found that participants (1) had more difficulty accurately 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 captured by a computational model of comparison that weights relational information more heavily for similarity than for difference judgments. Our results provide convergent evidence for a representational asymmetry between the relations *same* and *different.*

**Keywords:** comparison, similarity, relational reasoning

## Introduction

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

One attempt to explain this phenomenon invokes structure mapping theory (Gentner, 1983). Under this hypothesis, assessments of similarity and difference involve the same comparison process of structural alignment, in which representations of entity features and their structural relations are placed into one-to-one correspondence (Gentner & Markman, 1994; Markman, 1996; Markman & Gentner, 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 *nonalignable* differences (i.e., those not based on corresponding roles). For example, in a comparison between a car and a bicycle, wheel number would be an alignable difference (i.e., 4 vs. 2), whereas window number would be a nonalignable difference because this feature is only applicable to cars and not bicycles.

Proponents of this explanation noted that the featurally-similar option in the study by Medin et al. (1990) did not involve a salient relation, so that any relational difference between it and the standard did not constitute an alignable difference, and was therefore ignored in difference comparisons. However, later work found that both alignable and nonalignable differences contribute to judgments of difference; indeed, the latter actually exerted a *greater* influence than the former (Estes & Hasson, 2004). This result appears to undermine the core assumption required to explain asymmetries in similarity and difference judgments in terms of structure mapping theory.

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



In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.* 

of negation tends to place additional cognitive load on human reasoning. For example, determining the truth of a proposition including a negated expression (e.g., "star isn't above the plus") takes longer than a matched positive expression (e.g., "star is below the plus") (Carpenter & Just, 1975; Clark & Chase, 1972). Introducing extra 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).

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

We tested this hypothesis for both verbal comparisons between word pairs and 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. Participants completed two kinds of unambiguous comparisons: On *featural* trials, failure to select the similar option would reflect a difficulty in using featural similarity in comparison, whereas failure to select the similar option on *relational* trials would reflect a difficulty in using relational similarity. We expected that relational trials would be more cognitively demanding, and hence prove more difficult for participants judging difference as compared to similarity. On the other hand, since featural trials could be successfully completed without any relation processing, performance for difference versus similarity judgments was expected to be more equal. We also included ambiguous comparisons, for which either of the options might be selected depending on whether features or relations are emphasized (Bassok & Medin, 1997; Medin et al., 1990). We predicted that when judging difference as compared to similarity, participants would tend to base their choices on features rather than relations.

# Experiment

## Method

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


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

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

Each comparison task consisted of 24 trials, presented in a random order. Of these, 6 *unambiguous* trials included one option that was unambiguously more similar to the standard than the other. Correct responding on half of the unambiguous trials was more reliant on detecting the relative featural similarity of the two options, and so we refer to these as *featural* trials. The other 3 unambiguous trials were *relational* trials. On these, correct responding was more reliant on detecting the relative relational similarity of the two options.

The remaining 18 trials consisted of one option that was more featurally similar to but relationally different from the standard (FS/RD; e.g., the left option of both trials depicted in Figure 1) than the other option, which was more featurally different from but relationally similar to the standard (FD/RS; e.g., the right options of both trials in Figure 1). We refer to these trials as *ambiguous* trials because they were constructed so that selecting either option was valid, depending on a participant's criteria for judging similarity or difference. We used these trials to compare participants' preferential weighting of featural or relational information in their similarity and difference judgments. Selecting the FS/RD option as more similar indicates a preferential weighting of featural information, whereas selecting it as more different indicates a preferential weighting of relational information, and vice versa for selecting the FD/RS option.

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

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

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

**Ravens Progressive Matrices** Following the verbal comparison task, all participants completed an abridged, 12-problem version of the Ravens Advanced Progressive Matrices (RPM) (Arthur et al., 1999). On each problem in this task, participants are presented with a 3x3 array of simple geometric objects, with the object in the bottom-right corner of the array missing, and they are asked to select which one of 8 options best completes the pattern instantiated by the incomplete array. Carpenter et al. (1990) showed that individual differences in performance on these visual reasoning problems predict differences in the ability to induce abstract relations between objects and to maintain a hierarchy of problem goals and subgoals in working memory. 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.

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

## Results

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

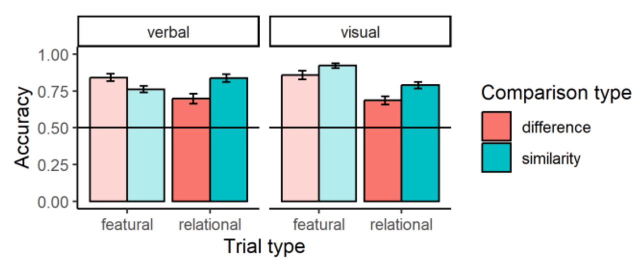


Figure 2: Human accuracy on *unambiguous* trials of verbal (left) and visual (right) comparison tasks, broken down according to trial type (featural vs. relational) and comparison type (difference vs. similarity). Error bars reflect ± standard error of the mean, and horizontal line reflects chance performance.

We used the *glmer* function from version 1.1.26 of the LME4 R package (Bates et al., 2015) in R version 4.1.1 (R. Core Team, 2021) to fit a logistic mixed-effects model to performance on unambiguous trials. We defined a full model including *participant* and *comparison problem* as random intercept effects; *comparison task* (*verbal* vs. *visual*), *comparison type* (*similarity* vs. *difference*) and *trial type* (*featural* vs. *relational*), as well as an interaction between the last two as fixed effects. As discussed previously, 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 improvement in performance across trials within each task.

We used likelihood-ratio tests to compare this full model to reduced models that omitted a term of interest but that was otherwise equivalent to the full model. First, 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 *comparison task* 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 *comparison task* term did not increase model prediction error, $\Delta AIC$ = -1.40, $\chi^2$ (1) = .65, $p$ = .420. This result indicates that the verbal and visual tasks did not differ in their overall difficulty.

Next, we tested our main 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 our full model to a reduced model that lacked the *judgment type x trial type* interaction term (but that retained the individual terms for *judgment type* and *trial type*). Dropping the interaction term did increase model prediction error, $\Delta AIC$ = 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 our 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 our hypothesis that difference judgments involve more cognitively demanding comparisons than similarity judgments, which particularly impact relational trials. Notably this difference in performance persisted even after we accounted for individual differences in reasoning ability by including *RPM score* as a covariate in our full model. A likelihood ratio test comparing the full model and a reduced model that lacked the *RPM score* term showed that removing that term indeed increased model prediction error, $\Delta AIC$ = 13.5, $\chi^2$ (1) = 15.56, $p$ < .001. Thus, even though general reasoning ability influenced performance on unambiguous trials, comparison type impacted performance specifically on relational trials, over and above individual differences in this ability.

**Relational responding on ambiguous trials**

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

As with unambiguous trials, 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; *comparison task* (*verbal* vs. *visual*), *comparison type* (*similarity* vs. *difference*) as fixed effects; and *RPM score*, *task order* (*verbal first* vs. *visual first*), and *trial number* as covariates.

As was done for unambiguous trials, we used likelihood-ratio tests to compare this full model to reduced models that omitted a term of interest but that was otherwise equivalent to the full model. First, 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, $\chi^2$ (1) = .01, $p$ = .930. This result again indicates that relational responding did not differ across verbal and visual comparison tasks.

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 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, $\chi^2$ (1) = 35.31, $p$ < .001, which confirms our main prediction that relational response rates were affected by comparison type on ambiguous trials. As on unambiguous trials, this effect on ambiguous trials held even after we accounted for individual differences in reasoning ability by including *RPM score* as a covariate in our full model. Omitting *RPM score* from the full model also increased model prediction error, $\Delta AIC$ = 2.6, $\chi^2$ (1) = 4.60, $p$ = .032. Even though individual differences in reasoning ability predicted relational responding on ambiguous trials, our manipulation of comparison type impacted responses over and above these individual differences.

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

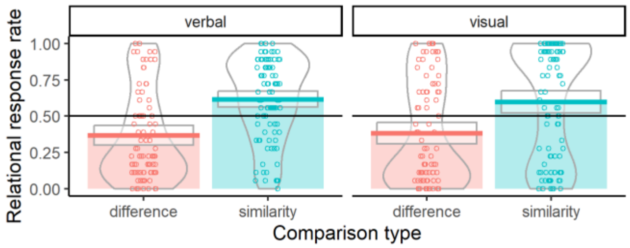


Figure 3: Relational response rate on *ambiguous* trials in verbal (left) and visual (right) comparison tasks, broken down according to comparison type (difference vs. similarity). Unfilled circles each reflect an individual participant's response rates, dark lines reflect mean response rates, box boundaries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

## Computational modeling

In order to formally characterize the human comparison process on ambiguous trials, we attempted to predict responses of individual participants on the verbal comparison task using a computational model. This model includes a weighting mechanism that controls the relative contribution of relational and featural information to a comparison judgment. We predicted that this weighting mechanism would create the observed asymmetry by altering the emphasis on relational information between similarity and difference judgments. Moreover, the computational model operates entirely on semantic representations of words and relations generated by machine learning, avoiding any hand-coding or reliance on experimenters' intuitions. Unlike computational implementations of structure mapping theory, the present model captures the *eduction of relations* (Lu et al., 2019; Spearman, 1923): generation of relations from nonrelational inputs. The same basic framework could be applied to visual judgments, given an appropriate front-end to create representations of visual stimuli.

### Model specification and approach

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

In order to represent individual word meanings, we used pre-trained Word2vec word embeddings (Mikolov et al., 2013), which represent word meanings as high-dimensional vectors of length 300. These vectors constitute the hidden layer of activation within a neural network trained to predict patterns of text in sequence as they appear in a large corpus consisting of Google News articles (about 100 billion words). Such word embeddings provide psychological models of semantic memory in that they preserve the similarity structure of individual word meanings in a psychologically realistic way. These embeddings have been used to successfully model a number of cognitive processes beyond similarity judgments, including human memory search, categorization, and decision making (Bhatia & Aka, 2022; Günther et al., 2019).

To compute lexical similarity, the meaning of a word pair is represented by a simple aggregate of the semantic vectors of the two individual words. We use $A$ to denote the first word in a word pair and $B$ to represent the second word in a word pair. 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 $j$, $[f_{A_j}, f_{B_j}]$:

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

To compute relational similarity, we used representations generated by *Bayesian Analogy with Relational Transformations* (BART), a learning model that has been used to predict human analogy performance and graded judgments of relational similarity (Ichien, Lu, & Holyoak, 2022; Lu, Chen, & Holyoak, 2012; Lu et al., 2019). BART assumes that specific semantic relations between words are coded as distributed representations over a set of abstract relations. The BART model takes pairs of Word2vec vectors as input, and then uses supervised learning with both positive and negative examples to acquire representations of individual semantic relations. After learning from datasets (Jurgens et al., 2012; Popov et al., 2017), BART can take inputs of any pair of words to calculate a relation vector consisting of the posterior probability that the word pair instantiates each of the learned relations. The posterior probabilities calculated for all learned relations form a 270-dimensional relation vector, 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:

$$sim_{rel_{ij}} = 1 - cos\left(BART(f_{A_i}, f_{B_i}), BART(f_{A_j}, f_{B_j})\right). \quad (2)$$

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

$$sim_{ij} = \alpha(sim_{rel_{ij}}) + (1 - \alpha)sim_{feat_{ij}} \quad (3)$$

$$diff_{ij} = -\alpha(sim_{rel_{ij}}) - (1 - \alpha)sim_{feat_{ij}}, \quad (4)$$

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

### Modeling results

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

Figure 4 shows the distribution of the parameter, broken-down according to comparison type. A Mann-Whitney U test confirmed what is clear from visual inspection: Fit relation-weight parameters were reliably greater for similarity participants than for difference participants, $W$ = 2540.5, $p$ < .001. This result confirms our prediction 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. Hence, this result further supports our main claim: similarity judgments prompt greater reliance on relational information than do difference judgments. Moreover, these simulations support the validity of our manipulation of featural and relational similarity.

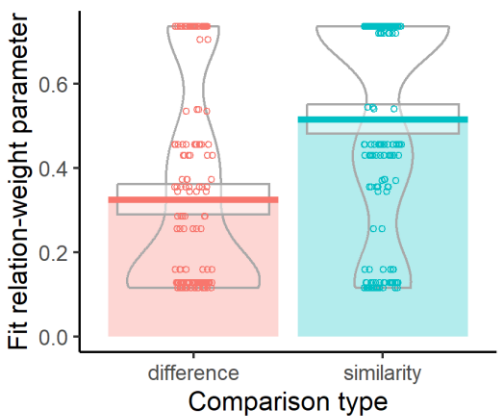


Figure 4: Relation-weight parameter values fit to individual participant data, broken down according to comparison type.

## General Discussion

For both visual and verbal comparisons, we 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. With respect to this latter finding, it is important to note that all word-pair stimuli in the verbal comparison task instantiated some binary semantic relation (either *part-of* or *category coordinate*), and so mismatching relations (e.g., between *hoof:horse* and *goat:cow*) constituted alignable differences. Structure mapping theory therefore erroneously predicts that such mismatching relations would contribute to difference judgments just as much as would mismatching features (Gentner & Markman, 1994; Markman, 1996). Participants should have thus selected all options with the same frequency, regardless of whether they were judging similarity or difference. Contrary to this prediction, we obtained an asymmetry in similarity and difference judgments even though all relational differences in our verbal stimulus set were alignable.

We acknowledge that we 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—they showed not only that nonalignable differences impacted both similarity and difference judgments but also that they had a *greater* (not lesser) impact than did alignable differences.

We were able to account for the asymmetry obtained in our experiment with verbal materials with a computational model of comparison based on machine-generated vector representations for both words and their semantic relations. When fit to human data at the level of individual participants, this model weighted relational information more heavily when fit to similarity judgments than when fit to difference judgments. Overall, this set of findings provides convergent evidence for the claim that assessments of difference are more cognitively demanding than assessments of sameness (Hochmann, 2021; Hochmann et al., 2016, 2018). This dissociation may ultimately be rooted in a representational asymmetry in the relations *same* and *different*, such that people process *different* as a negation of *same*.

## Acknowledgements

Preparation of this paper was supported by NSF Grants IIS-1956441 awarded to H.L. and BCS-2022369 to K.H.

## 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

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. https://doi.org/10.1093/cercor/bhh126

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. (1976). *Semantics and comprehension*. The Hague: Mouton. https://doi.org/10.1515/9783110871029

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

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

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

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

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)*, (pp. 356–364). Montreal, Canada. Association for Computational Linguistics.

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*. 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., 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

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

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