Ichien, N., Lu, H., & Holyoak, K. J. (2019). Individual differences in judging similarity between semantic relations. In A. Goel, C. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st Annual Meeting of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Individual Differences in Judging Similarity Between Semantic Relations

Nicholas Ichien[1] ichien@ucla.edu
Hongjing Lu[1,2] hongjing@ucla.edu
Keith J. Holyoak[1] holyoak@lifesci.ucla.edu

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

Abstract

The ability to recognize and make inductive inferences based on relational similarity is fundamental to much of human higher cognition. However, relational similarity is not easily defined or measured, which makes it difficult to determine whether individual differences in cognitive capacity or semantic knowledge impact relational processing. In two experiments, we used a multi-arrangement task (previously applied to individual words or objects) to efficiently assess similarities between word pairs instantiating various abstract relations. Experiment 1 established that the method identifies word pairs expressing the same relation as more similar to each other than to those expressing different relations. Experiment 2 extended these results by showing that relational similarity measured by the multi-arrangement task is sensitive to more subtle distinctions. Word pairs instantiating the same specific subrelation were judged as more similar to each other than to those instantiating different subrelations within the same general relation type. In addition, Experiment 2 found that individual differences in both fluid intelligence and crystalized verbal intelligence correlated with differentiation of relation similarity judgments.

Keywords: relational reasoning, similarity, semantic cognition, fluid intelligence, crystallized intelligence

Introduction

A house key and an email password are intuitively similar. This similarity is not based on any common attributes or constituent properties of individual objects; rather, it seems to be based on some common *relation* that a house key and an email password respectively bear to a house and to an email account (roughly, *providing access*). The ability to grasp and exploit similarity based on a wide variety of relations is an important and distinguishing trait of human intelligence (Penn, Holyoak, & Povinelli, 2008). This ability underlies much of human thought, including aspects of language (Gentner & Namy, 2006), categorization (Gentner & Kurtz, 2005; Goldwater & Schalk, 2016), and perhaps most prominently, analogical reasoning (Holyoak, 2012). The explicit representation of abstract relations is an indispensable explanatory construct in major computational accounts of human analogical reasoning (Doumas, Hummel, & Sandhofer, 2008; Falkenhainer, Forbus, & Gentner, 1989; Halford, Wilson, & Phillips, 1998; Hummel & Holyoak, 2003; Lu, Chen, & Holyoak, 2012; Lu, Wu, & Holyoak, 2019; Petrov, 2013). Empirical work on relational reasoning has provided compelling evidence that humans store representations of semantic relations in memory (Estes & Jones, 2006; Popov, Hristova, & Anders, 2017; Spellman, Holyoak, & Morrison, 2001).

A number of important research questions depend on finding an effective method to assess human judgments of relational similarity. A major source of complexity stems from evidence that relations are not represented as discrete all-or-none concepts, but rather exhibit internal variability. Just as instances of natural and functional object categories differ in typicality (Rosch, 1975), so too people reliably judge word pairs to be better or worse instantiations of relations (Jurgens, Mohammad, Turney, & Holyoak, 2012). For example, *fail:succeed* is considered to be a better example of the relation *reverse* than is *eat:starve*.

Given such variations in intra-relation "goodness", it is natural to hypothesize that inter-relation similarity will also have a graded structure. Indeed, a recent theory of relation learning (*Bayesian Analogy with Relational Transformations*, BART) claims that the specific relation between a pair of words corresponds to a distributed representation over multiple relations, each of which the pair instantiates with some probability (Lu et al., 2019). For example, *lid:bottle* seems to instantiate the relations *part-whole*, *on-top-of*, and *closure-of*. BART can be used to derive theoretical predictions about the degree of similarity between a wide range of word pairs that collectively instantiate multiple relations.

It would clearly be desirable to obtain reliable human judgments of relational similarity, which might then be compared to theory-based predictions. Such data could also be used to assess potential individual differences in relation representations. A great deal of research indicates that complex relational reasoning depends on working memory capacity and other aspects of fluid intelligence (for a review see Holyoak, 2012). In particular, there is evidence that performance on analogical reasoning tasks is positively related to fluid intelligence as measured by tests such as the Raven's Progressive Matrices (RPM; Gray & Holyoak, 2018). It is possible that fluid intelligence plays a role in maintaining and comparing relations in working memory in order to differentiate among relations that overlap in meaning. Similarly, crystalized verbal intelligence seems to play an important role in comprehending metaphors (Stamenković, Ichien, & Holyoak, 2019), and may be related to the differentiation of relational concepts in semantic memory.

A reliable measure of human judgments of relation similarity would clearly be very useful for testing theories of relation representation. However, in practice it is difficult to

find an efficient procedure to elicit similarities among large sets of items (since the number of pairwise comparisons becomes prohibitively large when the number of items is substantial). Here we explore the use of a *multi-arrangement* method (adapted from previous work on assessing object similarity; Kriegeskorte & Mur, 2012) for obtaining judgments that can be used to efficiently generate a map of the psychological similarity space for abstract semantic relations.

The present paper aims to offer a first step in the exploration of relational similarity, assessing the validity and reliability of a new method for collecting human judgments of relational similarity and conducting preliminary analyses of these similarity judgments. Experiment 1 sets the stage by testing whether the method can generate sensible patterns of relation similarity. Experiment 2 then extends the method to more fine-grained semantic distinctions among relations to examine potential gradations in relational similarity. Further, Experiment 2 assesses the potential association between judgments of relation similarity and individual differences in both fluid and crystalized intelligence.

Experiment 1

The major goal of Experiment 1 was to determine whether a novel method for eliciting human judgments of relation similarity is able to capture broad distinctions among semantic relations that have been posited on the basis of previous theoretical and empirical investigations.

Method

Participants

20 participants (mean age = 19.05 years; 17 female) were recruited from the Psychology Department subject pool at the University of California, Los Angeles (UCLA). All participants were self-reported fluent English speakers. Participants provided verbal consent in accordance with the UCLA Institutional Review Board and were compensated with course credit.

Stimuli

All stimuli were word pairs taken from the SemEval-2012 Task 2 dataset (Jurgens et al., 2012), which is in turn based on a taxonomy of abstract semantic relations developed by Bejar, Chaffin, and Embretson (1991). Word pairs in this dataset express one of 79 specific relations, each falling into one of 10 general types of relations. Experiment 1 tested examples drawn from relations in each of three different general relation types (*similar*, *contrast*, and *cause-purpose*). We will refer to the examples in Experiment 1 by the names of the specific relations: *synonymy*, *contrary*, and *cause:effect* (see Table 1). Each relation included 16 word pairs, consisting of one paradigm exemplar (a seed used by Jurgens et al. to define the relation) and the 15 most prototypical word pairs for that relation. Pairs were unique in that they did not include inversions of one another. Table 1 provides examples of the word pairs used in the experiment.

Relation types	Word pair examples
synonymy	car:auto
contrary	old:young
cause:effect	joke:laughter

Table 1. Relations and examples of word pairs used in Experiment 1.

Procedure

We acquired human similarity judgments of semantic relations by asking participants to perform a multi-arrangement task, a method for efficiently eliciting similarity judgments, especially for large sets of items (Kriegeskorte & Mur, 2012). The method, which can be viewed as an inverse of standard multidimensional scaling (Shepard, 1962), has previously been successfully used for judgments of object similarity (Kriegeskorte & Mur, 2012; Mur et al., 2013; Jozwik, Kriegeskorte, Storrs, & Mur, 2017). Here we extend it to judgments of relation similarity.

On each trial, participants were presented with a subset of the 48 word pairs on a computer screen. They were asked to first identify the relation between words in each pair, and then use a mouse to arrange word pairs in a two-dimensional circular space according to the similarity of their relations (see Figure 1). Participants were told, "word pairs that involve similar relations should be placed close together," "word pairs that involve very different relations should be placed far apart," and "the distance between two word pairs should represent how different their relations are." Participants were also instructed to use the entire space to arrange word pairs on each trial.

We aimed to obtain similarity judgments from each participant relating each of the 48 item pairs to each other (a total of 1128 pairwise measurements). Estimates of similarity were based on the *relative* on-screen distances between word pairs as arranged by participants on each trial. These estimates were calculated by scaling the distances between items arranged on a single trial to match a weighted average of these distances calculated across trials. This weighted average was iteratively recomputed until convergence.

On a given trial, participants were presented with a maximum of 20 word pairs. The multi-arrangement task involves adaptively selecting stimuli to present on each trial. On the first trial, participants arranged a random subset of 20 items from the entire set of 48 items. On subsequent trials, participants arranged a subset of 20 or fewer items selected based on item pairs with the weakest similarity evidence (see Kriegeskorte & Mur, 2012, for an extended discussion).

Previous uses of the multi-arrangement task have involved 1-hour sessions (e.g., Kriegeskorte & Mur, 2012; Mur et al., 2013; Jozwik et al., 2017), but these studies all asked participants to do a relatively easier task of arranging individual objects according to their similarity. Due to the higher demand on working memory in arranging word pairs according to their relational similarity, pilot experiments suggested that a 1-hour session length would likely result in fatigue and disengagement for naïve participants. Accordingly, we limited session length to 30 minutes.

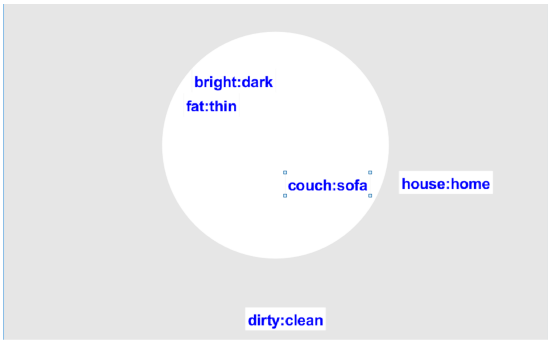

Figure 1. Example trial of the multi-arrangement task used to generate a semantic space for relations.

Participants were allowed to spend as long as they needed to complete each trial. On average, participants completed 28.5 trials (SD = 11.86, range = 4-44) within the 30-minute experimental duration.

Results

All but five participants provided a full set of pairwise similarity judgments between all combinations of the 48 word pairs. Of the five who failed to complete all possible comparisons, four provided judgments for 98% of the pairwise combinations of word pairs. The fifth participant provided judgments for just 57% of the combinations; this individual's data were excluded from analyses.

We assessed the inter-subject reliability of our relational similarity judgments by calculating the Pearson correlation coefficients between individual participants' distance matrices. The mean correlation between any two participants' distance matrices was .50 (range = .11 to .83).

We then examined whether the multi-arrangement task provided a reliable measure of relation similarity (assuming that greater inter-pair distance implies lower similarity). The results showed that participants generated smaller distances between word pairs within a relation compared to distances between word pairs instantiating a different relation. Figure 2 depicts a mean distance matrix obtained by averaging across distance matrices generated by individual participants performing the multiple-arrangement task.

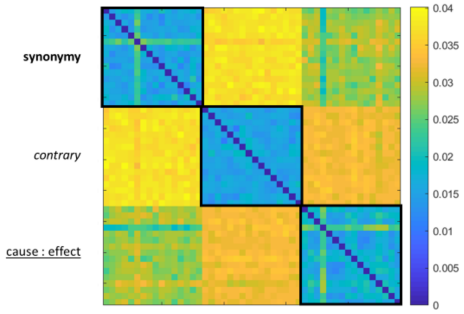

Figure 2. Mean distance matrix between the 48 word pairs used in Experiment 1. Cold colors represent smaller distances (i.e., greater pairwise similarity); hot colors represent greater distances (i.e., lesser pairwise similarity). Boxed regions represent pairwise distance measures between word pairs instantiating the same relation.

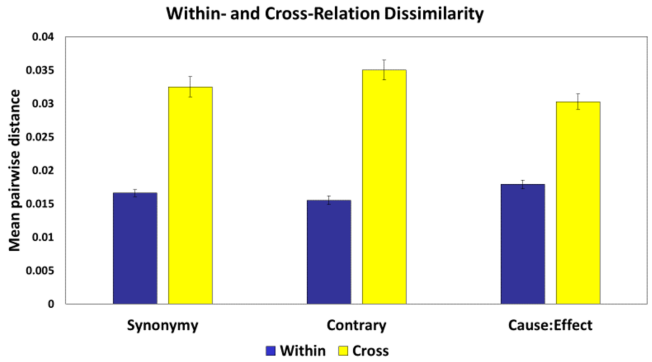

Figure 3. Mean within- and cross-relation distance measures for pairs instantiating each relation (Experiment 1). Higher bars indicate greater distance (i.e., lower similarity). Error bars indicate +/- 1 standard error of the mean.

We compared the mean distances between word pairs instantiating different relations (i.e., cross-relation distances) to the mean distances between word pairs instantiating the same relation (i.e., within-relation distances). To perform this analysis, we first calculated within- and cross-relation distances for each word pair for each individual participant. Next, we found the mean value of both of these distance measures averaged across word pairs within each relation. As depicted in Figure 3, cross-relation distances were greater than within-relation distances for each relation: for *synonymy* ($t(18) = 8.66$, $p < .001$, Cohen's $d = 1.99$); for *contrary* ($t(18) = 10.26$, $p < .001$, Cohen's $d = 2.35$); for *cause:effect* ($t(18) = 8.91$, $p < .001$, Cohen's $d = 2.5$). These findings thus establish that the multi-arrangement task is an effective method to obtain human judgments of relation similarity.

Experiment 2

Experiment 2 aimed to determine whether human judgments of relational similarity are sensitive to more fine-grained distinctions among relations than those examined in Experiment 1. In addition, we investigated whether relation judgments are systematically influenced by individual differences in cognitive capacity and/or semantic knowledge. To assess fluid intelligence, we administered a short version of the RPM (Arthur, Tubre, Paul, & Sanchez-Ku, 1999) adapted for computer administration using Matlab software. Participants are presented with a 3x3 grid of items with the item in the bottom right corner missing. They are asked to use the pattern instantiated by the presented items to select the most appropriate item to fill that bottom right corner from a set of 8 options. Prior research has shown that superior performance on this test is correlated with performance on tests of analogical reasoning (Vendetti, Wu, & Holyoak, 2014; Kubricht, Lu, & Holyoak, 2017). We hypothesized that the RPM measure would be associated with the degree to which people are able to differentiate word pairs that instantiate distinct relations.

In addition to fluid intelligence, the ability to differentiate among semantic relations may vary with crystalized verbal intelligence, particularly knowledge of semantic relations. As

a measure of semantic knowledge, we administered the Semantic Similarities Test (SST). This test was designed to be similar to the Similarities subscale of the Weschler Adult Intelligence Scale (WAIS), and is correlated with the Vocabulary subtest (Stamenković et al., 2019). Participants are presented with 20 pairs of verbal concepts and asked to describe how the concepts in each pair are similar. The concept pairs span a broad range of similarities: some are fairly specific (e.g., *bird-airplane*, which both fly), some are more general (e.g., *tavern-church*, which are both public buildings), and some are more metaphorical (e.g., *marriage-alloy*, which are both bonds between elements). Because the identification of more specific and fine-grained relations likely depends on greater semantic knowledge, we hypothesized that superior performance on the SST would also be correlated with greater differentiation of similarities among semantic relations.

Method

Participants

93 new participants (mean age = 20.17 years; 69 female) were recruited from the UCLA Psychology Department subject pool. All participants had normal or corrected-to-normal vision and were self-reported fluent English-speakers. Participants provided verbal consent in accordance with the UCLA Institutional Review Board and were compensated with course credit.

Stimuli

The multi-arrangement task in Experiment 2 used 27 word pairs drawn from the same norms as in Experiment 1 (Jurgens et al., 2012). Three word pairs were chosen from each of three specific subrelations of three general relation types (see Table 2). Note that the three relations used in Experiment 1 were included as specific subrelations used in Experiment 2. Whereas Experiment 1 did not manipulate the level of relation abstraction, Experiment 2 did. Specifically, Experiment 2 examined whether similarity judgments not only reflect broad distinctions at a high level of abstraction (i.e., between general relation types), but also fine distinctions at a lower level of abstraction (i.e., between specific subrelations within general relation types). Word pairs drawn from different subrelations of the same general type (e.g., *car:auto* instantiates *synonymy* and *rake:fork* instantiates *attribute similarity*, two subrelations of the relation type *similar*) are differentiated on the basis of relatively subtle relational differences. Each set of three unique word pairs consisted of one paradigm exemplar and the third and sixth most prototypical unique word pairs for that subrelation in the SemEval-2012 Task 2 norms (Jurgens et al., 2012).

Procedure

All participants completed three tasks in the following order: the multi-arrangement task, the Raven's Progressive Matrices (RPM) and the Semantic Similarities Test (SST).

General relation types	Specific subrelations	Word pair examples
similar	*synonymy*	car:auto
	attribute similarity	rake:fork
	change	discount:price
contrast	*contrary*	old:young
	directional	east:west
	pseudoantonym	right:bad
cause-purpose	*cause:effect*	joke:laughter
	cause: compensatory action	hunger:eat
	action/activity: goal	flee:escape

Table 2. General relation types, three specific subrelations chosen to exemplify each, and examples of word pairs used in Experiment 2.

Results

All 93 participants completed the multi-arrangement task. On average participants completed 19.51 trials (SD = 9.70, range 2-55). All but one participant provided pairwise similarity judgments for all 27 word pairs (351 pairwise comparisons). That one participant provided judgments for 86% of the pairwise combinations. Due to program failures, only 88 participants completed the SST, and 90 participants completed the RPM.

We again assessed the inter-subject reliability of our relational similarity judgments by calculating the Pearson correlation coefficients between individual participants' distance matrices. The mean correlation between any two participants' distance matrices was .38 (range = -.09 to .88).

Figure 4 depicts the mean distance matrix for all word pairs. We compared the mean distances of word pairs drawn from different general relation types (i.e., cross-type distances) to mean distances of word pairs within the same relation type (i.e., within-type distances). As depicted in Figure 5, cross-type distances were greater than within-type distances for each relation type: for *similar* ($t(92) = 10.53$, $p < .001$, Cohen's $d = 1.09$); for *contrast (*$t(92) = 18.32$, $p < .001$, Cohen's $d = 1.90$); for *cause-purpose (*$t(92) = 17.06$, $p < .001$, Cohen's $d = 1.77$).

To examine whether participants were sensitive to differences between specific subrelations within the same relation type, we compared the mean distances of word pairs instantiating different subrelations within the same general relation type (i.e., cross-subrelation distances) to the mean distances of word pairs instantiating the same subrelations (within-subrelation distances). For each relation type, mean cross-subrelation distances were greater than mean within-subrelation distances: for *similar* ($t(92) = 13.17$, $p < .001$, Cohen's $d = 1.37$); for *contrast (*$t(92) = 12.95$, $p < .001$, Cohen's $d = 1.34$); for *cause-purpose (*$t(92) = 7.35$, $p < .001$, Cohen's $d = 0.76$). These findings indicate that participants were not only able to differentiate between general relation types but were also sensitive to much more fine-grained distinctions within the same relation type. Further, these findings provide evidence of graded similarity structure among semantic relations. Specifically, word pairs instantiating the same general relation type were judged as

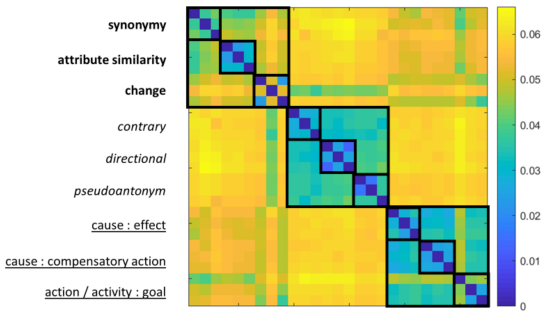

Figure 4. Mean distance matrix from Experiment 2. Cold colors represent smaller distances (i.e., greater pairwise similarity), whereas hot colors represent greater distance (i.e., lesser pairwise similarity). Larger boxed regions represent pairwise distance judgments between word pairs instantiating the same general relation type. Smaller boxed regions represent pairwise distance judgments between word pairs instantiating the same specific subrelation within a common relation type.

more similar to each other than those instantiating different general relation types, and word pairs instantiating the same subrelation were judged as more similar to each other than those instantiating different subrelations within the same general relation type.

Next, we performed analyses to determine whether individual differences in cognitive capacity (as assessed by the RPM) and semantic knowledge (as assessed by the SST) were associated with participants' sensitivity to differences among relations. Two independent raters scored the SST based on the criteria summarized by Stamenković et al. (2019). We assessed the reliability of these raters' scores by testing the average measure intraclass correlation coefficient across scores using a two-way mixed model (ICC = .971, $F(19,19) = 44.72$, $p < .001$, with a 95% confidence interval from .899 to .990). Given the reliability of these scores, we used the average score across these two raters in the following analyses.

In order to estimate individual differences in sensitivity to broad distinctions relation types, we computed a relation type discriminability index for each participant using the following steps. First, we found each participant's cross-type distance by calculating the mean distance for pairwise comparisons between word pairs instantiating different general relation types (e.g., *old:young* instantiates the relation type *contrast*, while *car:auto* instantiates the relation type *similar*). Second, we found each participant's within-type distance by calculating the mean distance for pairwise comparisons between word pairs instantiating the same general relation type (e.g., *old:young* and *east:west* both instantiate the relation type *contrast*). Third, we computed each participant's discriminability index by dividing that participant's cross-type distance by their within-type distance (range = 1.01 to 2.60). This relation type discriminability index reflects how well a participant discriminated between relation types in their similarity judgments. An index of 1 indicates complete lack of discriminability between word pairs instantiating different relation types and those instantiating the same relation type, whereas higher indices indicate judgments of greater similarity between word pairs instantiating the same relation type than between word pairs instantiating different relation types.

These discriminability indices for relation types were significantly correlated with RPM scores (Pearson's $r = .33$, $p = .005$, power = .90) and also with SST scores (Pearson's $r = .30$, $p = .014$, power = .82). Partial correlations revealed that these discriminability indices were significantly correlated with RPM scores after residualizing out SST scores (Pearson's $r = .236$, $p = .028$, power = .61), and that they were significantly correlated with SST scores after residualizing out RPM scores (Pearson's $r = .236$, $p = .028$, power = .61). These results indicate that there is an association between the discrimination of general relation types both with cognitive capacity and with semantic knowledge.

In order to estimate each participant's sensitivity to more fine-grained distinctions between specific subrelations within general relation types, we also computed a subrelation discriminability index using the following steps. First, we found each participant's cross-subrelation distance by calculating the mean distance for pairwise comparisons between word pairs instantiating different subrelations within the same general relation type (e.g., *old:young* instantiates the subrelation *contrary*, and *east:west* instantiates the subrelation *directional*, where both instantiate the relation type *contrast*). Second, we found each participant's within-subrelation distance by calculating the mean distance for pairwise comparisons between word pairs instantiating the same subrelation (e.g., *old:young* and *black:white* both instantiate the subrelation *contrary*). Third, we computed each participant's subrelation discriminability index by dividing each participant's cross-subrelation distance by their within-subrelation distance (range = .96 to 2.74). This subrelation discriminability index reflects how well a participant was able to discriminate between specific subrelations within a relation type in their similarity judgments. An index of 1 would indicate a complete lack of discriminability between word pairs instantiating different subrelations and those instantiating the same subrelation, whereas higher indices indicate judgments of greater similarity between word pairs instantiating the same subrelation than between word pairs instantiating different subrelations.

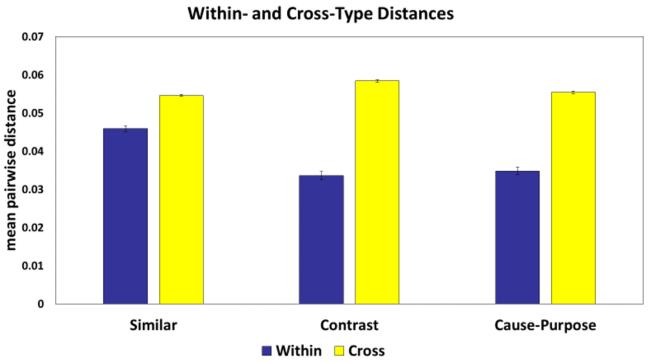

Figure 5. Mean within- and cross-type distances for each general relation type in Experiment 2. Error bars indicate +/- 1 standard error of the mean.

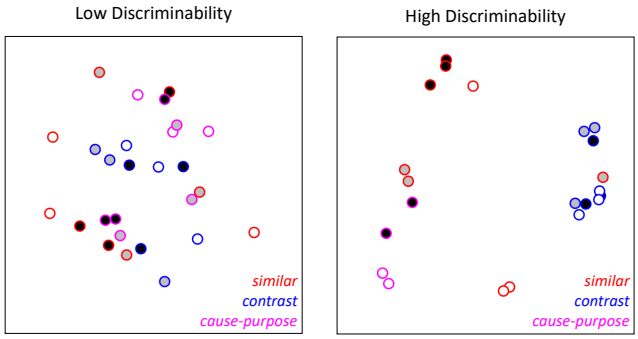

Figure 6. Visualization of relation similarities from two representative participants. Left: MDS solution for a participant with low discriminability indices (relation type discriminability index = 1.02; subrelation discriminability index = .98). Right: solution for a participant with high discriminability indices (2.08 and 2.74, respectively). Each marker indicates a single word pair. Marker outline color indicates word pair relation type, and marker shading indicates subrelation within relation type.

These fine-grained discriminability indices for subrelations showed a significant correlation with RPM scores (Pearson's r = .35, p = .003, power = .93), and also with SST scores (Pearson's r = .30, p = .014, power = .82). Partial correlations revealed that these discriminability indices were significantly correlated with RPM after residualizing out SST scores (Pearson's r = .291, p = .006, power = .79), but that they were not correlated with SST scores after residualizing out RPM scores (Pearson's r = .090, p = .408). These results indicate that there is a stronger association between the discrimination of specific subrelations within relation types with cognitive capacity than with semantic knowledge.

To provide a visualization of the difference between high and low discriminability, Figure 6 presents multidimensional scaling (MDS) solutions (Shepard, 1962) for the distance matrices of a participant with both a low relation type and a low subrelation discriminability index (left) and of a participant with both a high relation type and a high subrelation discriminability index (right). The latter solution shows a much greater degree of clustering into distinct relation types as well as into subrelations.

General Discussion

Across two experiments, we showed that a multi-arrangement task can be used to efficiently assess judgments of similarity among semantic relations. Human judgments obtained using this method have a clear interpretation. Judged similarity reflects not only broad distinctions between relation types, but also finer distinctions between subrelations within relation types. Moreover, the degree to which a participant differentiated between pairs from the same versus different relation types was positively correlated with measures of both fluid and verbal crystallized intelligence. At the more detailed level of subrelations, only fluid intelligence was a reliable predictor of discriminabilty. Future work should examine these associations further and assess directions of causality.

The present findings add to mounting evidence that semantic relations do not have discrete, all-or-none representations. Previous work has shown that word pairs instantiating a particular relation vary systematically in their *typicality* (Jurgens et al., 2012; Popov et al., 2017), much like instances of object categories (Rosch, 1975). Our findings reveal that *similarities* between relation examples (within and across subrelations) also vary in a graded fashion. In addition, the present study establishes that similarity gradients for relations show reliable individual differences across people who vary in either cognitive capacity or semantic knowledge of relations.

Note typicality judgments are importantly distinct from similarity judgments. Specifically, typicality is a relation between entities at different levels of abstraction (i.e., exemplar and category), and the typicality of a word pair is necessarily defined with respect to a particular relation. For example, *up:down* is typical of the relation *opposite*. In contrast, similarity is generally a relation between entities at the same level of abstraction (i.e., exemplar and exemplar), and relational similarity of a word pair can be defined with respect to another word pair. For example, *up:down* is similar to *light:dark*., Notably, whereas typicality judgments can be used to evaluate relational semantic representations *within* relations, similarity judgments can be used as a more holistic evaluation *across* relations.

This emerging picture of human relation concepts is consistent with models of relation learning and analogical reasoning that assume relations are coded by distributed representations (e.g., Lu et al., 2019). More generally, judgments of relation similarity provide a rich source of potential data that can be used to evaluate computational models. Specifically, a relation distance matrix generated from a theoretical model can be compared to a distance matrix obtained from human judgments of relation similarity, as described here. To the extent that a model-generated distance matrix approximates a human-generated distance matrix, the model's representation of semantic relations is descriptive of human semantic cognition. The same logic can be applied to test computational models as predictors of relational priming (Estes & Jones, 2009; Popov et al., 2017; Spellman et al., 2001), and of neural responses to relation processing (Kriegeskorte, Mur, & Bandettini, 2008).

The multi-arrangement method of collecting similarity judgments for relations may also prove useful in guiding studies of educational interventions (Goldwater & Schalk, 2016). The type of MDS solution that can be derived from similarity judgments can be related to the well-known technique of using "concept maps" to teach systematically related concepts. The degree of match between the clusters identified in an MDS solution obtained for an individual learner may provide a useful index of how well that learner's internal representation of a set of concepts maps onto the organization the teacher aimed to convey.

Acknowledgements

We thank Ali Hepps, Anvita Diwan, Lina Chan, and Zhibo Zhang for assistance in data collection. This research was supported by NSF Grant BCS-1827374.

References

Arthur, P. L., 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*, 354-361.

Bejar, I. I., Chaffin, R., & Embretson, S. E. (1991). *Cognitive and psychometric analysis of analogical problem solving*. New York: Springer-Verlag.

Doumas, L. A. A., Hummel, J. E., & Sandhofer, C. M. (2008). A theory of the discovery and predication of relational concepts. *Psychological Review*, *115*, 1-43.

Estes, Z., & Jones, L. L. (2009). Integrative priming occurs rapidly and uncontrollably during lexical processing. *Journal of Experimental Psychology: General*, *138*(1), 112-130.

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

Gentner, D., & Kurtz, K. (2005) Relational categories. In W. K. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman, & P. W. Wolff (Eds.), *Categorization inside and outside the lab*. Washington, DC: APA.

Gentner, D., & Namy, L. L. (2006). Analogical processes in language learning. *Current Directions in Psychological Science*, *15*, 297-301.

Goldwater, M. B., & Schalk, L. (2016). Relational categories as a bridge between cognitive and educational research. *Psychological Bulletin*, *142*, 729-757.

Gray, M. E., & Holyoak, K. J. (2018). Individual differences in relational reasoning. In C. Kalish, M. Rau, J. Zhu & T. T. Rogers (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 1741-1746). Austin, TX: Cognitive Science Society.

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), 803-864.

Holyoak, K. J. (2012). Analogy and relational reasoning. In K. J. Holyoak & R. G. Morrison (Eds.), *Oxford handbook of thinking and reasoning* (pp. 234-259). New York: Oxford University Press.

Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review*, *110*(2), 220–263.

Jozwik, K. M., Kriegeskorte, N., Storrs, K. R., & Mur, M. (2017). Deep convolutional neural networks outperform feature-based but not categorical models in explaining object similarity judgments. *Frontiers in Psychology*, *8*, Article ID 1726. DOI: 10.3389/fpsyg.2017.01726

Jurgens, D. A., Mohammad S. M., Turney P. D., & 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.

Kriegeskorte, N., & Mur, M. (2012). Inverse MDS: Inferring dissimilarity structure from multiple item arrangements. *Frontiers in Psychology*, *3*. DOI: 10.3389/fpsyg.2012.00245

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in System Neuroscience*, *2*(4). DOI: 10.3389/neuro.06.004.2008

Kubricht, J. R., Lu, H., & Holyoak, K. J. (2017). Individual differences in spontaneous analogical transfer. *Memory & Cognition*, *45*, 576-588.

Lu, H., Chen, D., & Holyoak, K. J. (2012). Bayesian analogy with relational transformations. *Psychological Review*, *119*, 617-648.

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

Mur, M., Meys, M., Bodurka, J., Goebel, R., Bandettini, P. A., & Kriegeskorte, N. (2013). Human object-similarity judgments reflect and transcend the primate-IT object representation. *Frontiers in Psychology*, *4*, 128. DOI: 10.3389/fpsyg.2013.00128

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, *104*(3), 192-233.

Penn, D. C., Holyoak, K. J., & Povinelli, D. J. (2008). Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. *Behavioral and Brain Sciences*, *31*(2), 109-130.

Petrov, A. A. (2013). *Associative Memory-based Reasoning: A computational model of analogy-making in a decentralized multi-agent cognitive architecture*. Saarbrücken, Germany: Lambert Academic.

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.

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika*, *27*, 125-40.

Spellman, B. A., Holyoak, K. J., & Morrison, R. G. (2001). Analogical priming via semantic relations. *Memory & Cognition*, *29*, 383-393.

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

Vendetti, M., Wu, A., & Holyoak, K. J. (2014). Far-out thinking: Generating solutions to distant analogies promotes relational thinking. *Psychological Science*, *25*(4), 928-933.