Probing the Mental Representation of Relation-Defined Categories

Ava Y. Du (du.618@osu.edu)
Department of Psychology, Ohio State University, 1835 Neil Ave., Columbus, OH 43210

John E. Hummel (jehummel@illinois.edu)
Department of Psychology, University of Illinois, 603 E. Daniel St., Champaign, IL 61820

Alexander A. Petrov (apetrov@alexpetrov.com)
Department of Psychology, Ohio State University, 1835 Neil Ave., Columbus, OH 43210

Abstract

The mental representation of relation-based concepts is different from that of feature-based concepts. In the present experiment, participants learned to categorize two fictional diseases that were defined either by a feature (e.g., short cells) or an ordinal relation (e.g., diseased cells being shorter than healthy cells). After the participants learned the categorization task to criterion, their strategies were probed in transfer task in which features and relations were pitted against one another. Finally, participants engaged in a stimulus reconstruction task. The results supported the prediction that participants who had adopted a feature-based strategy on a stimulus dimension, as identified by transfer data, tended to reconstruct values close to the means presented during training. By contrast, participants who had adopted a relation-based strategy tended to exaggerate that dimension away from the mean of the training examples and in the direction of the category-defining comparative relation. These data add to the growing literature suggesting that, unlike featural categories, relational categories are not represented in terms of the category's central tendency.

Keywords: Category learning; relations; typicality; probing the category representation; extreme-value hypothesis

Introduction

Ever since Wittgenstein's (1953) observations about the "family resemblance" nature of concepts, cognitive psychologists have thought of concepts as lists of features, each of which appears only probabilistically across various exemplars of the concept. However, many important concepts are better conceived not as collections of features but as relations between things (Barsalou, 1985; Gentner & Kurtz, 2005; Goldwater, Bainbridge, & Murphy, 2016; Murphy & Medin, 1985). For example, a *barrier* may be a concrete structure on a roadway that blocks one's usual route to work, or a financial limitation that prevents a student from attending her school of choice. Although these two kinds of barrier have few features in common, they share the relational property that each stands between an agent and her goal.

The distinction between feature- and relation-based concepts is important for several reasons. Whereas many animals use feature-based concepts such as *food* or *bed*, it has been argued that only humans understand relation-based concepts such as *series, limit, connection, closure, support, resistance*, which play a central role in many uniquely human forms of thought (Penn, Holyoak, & Povinelli, 2008). Another difference between feature- and relation-based concepts concerns how they are learned. Feature-based concepts amount to statistical associations between features and category labels and thus can be learned by simple associative systems. By contrast, relation-based categories cannot be learned by tracking simple statistical associations between individual features and labels (Doumas, Hummel, & Sandhofer, 2008; Hummel & Holyoak, 2003). Instead, these and other authors have proposed that relational concepts are learned by a process more akin to schema induction by intersection discovery over examples (Gick & Holyoak, 1983; Hummel & Holyoak, 2003; Jung & Hummel, 2015a, 2015b), which makes probabilistic relation-based concepts very hard to learn as their intersection is the empty set (Kittur, Hummel, & Holyoak, 2004).

There is also evidence that relation-based concepts are represented differently than feature-based ones (Gentner & Kurtz, 2005). The latter very consistently demonstrate prototype effects as pointed out by Wittgenstein (1953; see Murphy, 2002, for review): A "good" member of a feature-based category is one that is *typical*, sharing many features with the prototype (i.e., the central tendency) of the category. By contrast, a "good" member of the relational category "diet food" is not a typical diet food (which is low in calories but bland), but rather an extreme exemplar, which has zero calories but is delicious (Barsalou, 1985): That is, the "goodness" of a member of at least some relation-based categories may be a function not of its similarity to the prototype, but of the degree to which it instantiates extreme values of the relevant relations (Goldwater, Markman, & Stilwell, 2011; Kim & Murphy, 2011; Kittur Holyoak, & Hummel, 2006; Rein, Goldwater, & Markman, 2010). A central goal of the present study is to explore this *extreme-value* hypothesis.

Kittur et al. (2006) studied people's learning of (partially deterministic) relation-based categories and found support for the extreme value hypothesis. In their experiment, participants learned two categories, A and B, defined by the relations between two shapes. Every exemplar depicted an octagon and a square that differed in their *size*, *darkness*, *relative height*, and which was *in front of* the other. In the prototype of category A, the octagon was *larger*, *darker*, *above* and *in front of* the square, and in the prototypical B these four relations were reversed. To ensure that participants

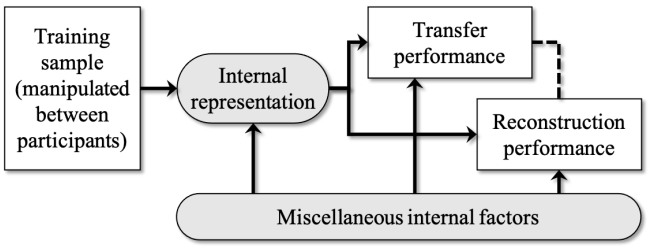

Figure 1: Schematic depiction of the three phases of our experimental design (white boxes) and the cognitive structures (shaded ovals) hypothesized to be involved. Solid arrows depict causal links; the dashed line depicts a statistical dependence between the performance measures.

learned the categories in terms of the relations between octagon and square (rather than their exact sizes and darkness), the absolute sizes and darkness varied during training in a limited range, but the *relative* size and darkness was always consistent with the category structure. Using a deterministic version of this category structure, Kittur et al. (2006) showed that the category structures participants learn do not equate goodness with typicality. During a post-training transfer task, participants were shown two members of a given category and asked to say which was a "better" member. The most striking result was that they consistently chose the exemplar with the most extreme values of a relation as the "best" exemplar, even though that exemplar depicted values outside the range that was experienced during training. Importantly, however, this extrapolation took place only for relations that had been deterministic during training.

The present study uses a novel method to further explore the extreme-value hypothesis. The experiment had three phases depicted by the white boxes in Figure 1. In the *training phase* the participants performed a classic two-category classification-learning task with feedback. The stimuli involved "diseased cells" and "healthy cells" that varied along 4 dimensions as detailed below. These gave rise to 4 features (e.g., *short* diseased cells) and 4 relations (e.g., *shorter* diseased cells compared to healthy cells). Both types of attributes – features and relations – were deliberately defined over the same underlying stimulus dimensions to minimize differences in their visual salience. The category structure was manipulated experimentally between participants in two main groups during training. In the Relation group, one relation (counterbalanced) was deterministic (i.e., perfectly predictive of the category label), whereas each of the other 7 attributes was 75% diagnostic. In the Feature group, one particular (counterbalanced) feature was deterministic, whereas each of the other 7 attributes was 75% diagnostic.

Note that each training environment afforded multiple strategies to achieve perfect categorization accuracy. The *deterministic strategy* was to identify the one deterministic attribute for the particular environment and rely on it for categorization. However, various *probabilistic strategies* that pool information across multiple attributes were available as well. These strategies can be *pure* or *mixed* depending on whether relations and features are used together. The polyvalent training environments allowed us to investigate two questions of interest. First, whether participants spontaneously prefer deterministic over probabilistic strategies. Based on earlier research (Kittur et al., 2004; Jung & Hummel, 2015a, 2015b) we hypothesize that they would, especially in the Relation group. Second, we are interested in the possibility that the Relation environment might promote a *relational mind-set* distinct from the mind-set in the Feature environment (Vendetti, Wu, & Holyoak, 2014). This would predict a preference for pure (relation only) over mixed (relation plus feature) strategies in the Relation group.

Following training, the participants' knowledge of the category structure was probed in two complementary ways (Fig. 1). During the *transfer phase* participants categorized a sequence of novel probe stimuli designed to differentiate relation- from feature-based strategies. Finally, during the *reconstruction phase* they used sliders to construct representative members of the categories they had learned previously. Of interest was whether participants who were identified as relying on a given relation (as opposed to feature) would selectively exaggerate the reconstructed value of the corresponding stimulus dimension relative to the mean value of that dimension in the training sample.

Following Kittur et al. (2006), we predicted that during reconstruction: (*i*) participants will produce more extreme values in comparison to central tendencies for the arguments of the relations they had adopted, whereas (*ii*) those who have learned deterministic features will reproduce the central tendencies they were exposed to during training.

The design of the present study is somewhat complex because it combines experimental and quasi-experimental aspects. The training sample is manipulated experimentally via random assignment of participants to groups (Fig. 1). However, because each training environment affords multiple strategies, the internal category representation of successful learners is not completely determined by the experimental manipulation. We expect individual differences in accuracy, strategy choice, and reconstruction performance. These differences reflect variability in motivation, propensity to adopt a relational mind-set (Vendetti et al., 2014), and various other internal factors (e.g., Goldwater et al., 2018). The transfer phase serves as a manipulation check and also assesses individual differences in strategy choice. This sets up the quasi-experimental aspect of the study: We predict a correlation between the transfer and reconstruction measures across participants, as depicted in Figure 1.

Method

Participants, Groups, and Inclusion Criteria

The experiment was conducted on-line during the COVID-19 pandemic. Students at the Ohio State University participated for course credit. Pilot studies showed that many students from this pool do not take their experimental participation seriously. To mitigate this problem, we fixed in advance two

criteria for inclusion in the sample. The "low" criterion was defined as follows: We excluded participants who (*i*) failed 2 or more of the 4 attention checks embedded within the trial sequence, (*ii*) had a median response time during training less than 0.3 sec, or (*iii*) produced a repetitive pattern of responses (e.g., pressing the same key on most trials). Due to COVID-related problems, we had access to 89 participants in total, 12 of whom failed the low criterion, leaving us with a sample of 72 students. All participants were randomly assigned to two training environments: Relation and Feature. Each main group was further partitioned into 4 subgroups to counterbalance the 4 dimensions of stimulus variation. The data collection protocol ensured a balanced sample at the top level: We ended up with 36 participants per group. Unfortunately, we could not maintain balance among the subgroups. There was also a "high" inclusion criterion that is defined below.

Stimuli and Category Structure

Participants were instructed to diagnose two fictional diseases – "Azolitis" (A) and "Leporidis" (L) – on the basis of fictional "micrographs" from patients. Each micrograph (i.e., exemplar) contained two populations of cells, diseased and healthy, that varied on four dimensions: the overall number of cells of each kind, the number of organelles within each cell, the number of hairs on the surface of each cell, and the length of each cell (the width of all cells was fixed to 25 pixels). "Healthy" cells were rendered in grey and "diseased" cells were rendered in pink. All cells were presented on a light grey background (Figure 2). One micrograph was presented on each trial with the content area (i.e., grey area) of 895 pixels wide x 415 pixels high. Each participant used their own computer; cell phones and tablets were not allowed.

In the following, we will use uppercase and lowercase As and Ls to refer to the typical relations (uppercase) and features (lowercase) of Azolitis and Leporidis, respectively. The "diseased" Leporidis cells always were *125 pixels in length [l]*, with *4 cells [l]*, *8 organelles [l]*, *and 4 hairs [l]*.

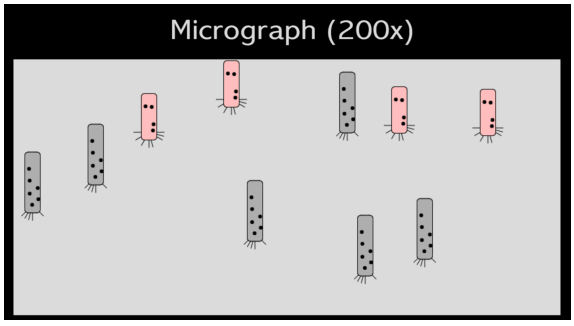

Figure 2: Each stimulus consisted of a "micrograph sample" that contained "diseased cells" (pink) and healthy cells (grey) varying in the overall number of cells, number of organelles within a cell, number of hairs, and cell length.

Table 1: Illustration of the scheme for generating the training environment for one particular subgroup (R1). The four stimulus dimensions are designated 1, 2, 3, 4 across the top; R stands for relation and F stands for feature. The rows specify 4 instances of Azolitis and 4 of Leporidis. Relations consistent with either disease are denoted "A" or "L", and features are denoted "a" or "l" as specified in the text. The two entries marked with asterisks in the R1 column make R1 the single deterministic attribute in this subgroup. The other 7 training environments are constructed analogously.

Instances	Stimulus Dimensions							
	R1	F1	R2	F2	R3	F3	R4	F4
Azolit.1	**A***	l	A	a	A	a	A	a
Azolit.2	A	a	L	l	A	a	A	a
Azolit.3	A	a	A	a	L	l	A	a
Azolit.4	A	a	A	a	A	a	L	l
Lepor.1	**L***	a	L	l	L	l	L	l
Lepor.2	L	l	A	a	L	l	L	l
Lepor.3	L	l	L	l	A	a	L	l
Lepor.4	L	l	L	l	L	l	A	a

The prototypical micrographs of Azolitis had diseased cells that were *shorter [A], more numerous [A]*, and with *fewer organelles [A]* and *more hairs [A]* than healthy cells. The prototypical Azolitis was also characterized by specific absolute feature values on the diseased (pink) cells, namely, *75 pixels in length [a]*, *8 cells [a]*, *4 organelles within each cell [a]*, and *8 hairs on the surface of each cell [a]*. By contrast, the prototypical micrographs of Leporidis had diseased cells that were *longer [L]*, *less numerous [L]*, *with more organelles [L]* and *fewer hairs [L]* than healthy cells.

To generate a training exemplar, the absolute features of the diseased cells and their relations to the healthy cells were determined from the group-specific category structure. The features of the healthy cells were then chosen to ensure that (*i*) the differences between diseased and healthy cells were easy to detect visually, and (*ii*) the absolute features of the diseased cells took values consistent with the feature-based definition of their category.

In the Relation condition, one relation (counterbalanced across subgroups) was deterministically related to the category label. For example, consider the subgroup whose deterministic relation was the number of cells. Table 1 lists the 8 concrete training exemplars for this particular subgroup. Notice that in every Azolitis exemplar the pink cells were more numerous (denoted by *A* in the R1 column) than the grey cells, and in every Leporidis exemplar they were less numerous (denoted by *L*) than the grey cells. Each of the other 7 attributes was 75% diagnostic. In the Feature condition, the deterministic attribute was always a feature (counterbalanced), while the relations and the other 3 features were 75% diagnostic. For example, in the subgroup whose deterministic feature was the number of cells (denoted F1), every Azolitis exemplar had exactly 8 pink cells (*a*) and every Leporidis exemplar had 4 pink cells (*l*).

Table 2: Prototypes and probe exemplars presented in the transfer phase. Same notation as in Table 1. "O" denotes neutral (i.e., equal) relations, and "o" denotes neutral feature values (halfway between the prototypes).

Instances	Stimulus Dimensions							
	R1	F1	R2	F2	R3	F3	R4	F4
Proto. A	A	a	A	a	A	a	A	a
T1.1	A	l	O	o	O	o	O	o
T1.2	L	a	O	o	O	o	O	o
T2.1	O	o	A	l	O	o	O	o
T2.2	O	o	L	a	O	o	O	o
T3.1	O	o	O	o	A	l	O	o
T3.2	O	o	O	o	L	a	O	o
T4.1	O	o	O	o	O	o	A	l
T4.2	O	o	O	o	O	o	L	a
Proto. L	L	l	L	l	L	l	L	l

Each of these 8 training environments (R1-R4, F1-F4) afforded multiple categorization strategies as discussed in the Introduction. A participant could adopt a purely relational or a purely featural strategy regardless of the subgroup they were assigned to. The experimental manipulation promoted one type or the other but imposed no hard constraints. A transfer phase followed the training phase and was designed to help us adjudicate which strategy (if either) an individual participant had learned during training. The transfer phase was the same for all participants in all groups (Table 2).

The transfer phase presented a sequence of *probe exemplars* such that a participant who had learned a relation-based strategy would label them one way, but one who had learned a feature-based strategy from the same training set would label them the opposite way. Each probe exemplar focused on a single critical dimension: its relation agreed with the prototype of one category, but the feature agreed with the opposite prototype. The other 3 stimulus dimensions were kept fixed at neutral values. These neutral values are denoted "O" for relations and "o" for features in Table 2. To neutralize a relation, the diseased and healthy cells were equal on the corresponding dimension. To neutralize a feature, the value for the diseased cells was set halfway between the values of the two prototypes: *100 pixels in length*, *6 cells*, *6 organelles*, *and 6 hairs.*

Procedure and Scoring

Participants were first instructed about the two diseases, shown examples of the stimuli, and asked questions designed to verify that the participant understood the stimuli and stimulus dimensions. The training phase then began. The participants were presented with a sequence of micrographs – one per trial – and required to categorize each by pressing either A (for Azolitis) or L (Leporidis). Initially, participants had to make random guesses, but they were expected to learn from the accuracy feedback, which appeared underneath the micrograph at the end of each trial.

Trials were presented in blocks, where each block presented eight exemplars, four per category (as illustrated in Table 1) in a random order. Training ended when the participant reached the "high criterion" by getting at least 87.5% (7/8) correct for three blocks in a row, or after they completed for 312 trials, whichever came first.

After training, each participant entered the transfer phase. They were instructed that they would be asked to help make diagnosis for some new samples whose correct classification may be not known. Each trial of the transfer phase presented either a probe exemplar (as defined above) or, as a form of practice, a category prototype. Each probe exemplar and prototype were presented four times, for a total of 40 transfer trials. Accuracy feedback was given only in response to the prototypes.

The transfer data were scored separately for each stimulus dimension. If a participant made a relation-based response on at least 7 of the 8 trials that probed dimension *N*, by definition she heeded relation RN (e.g., R1). Conversely, if she made at least 7 feature-based responses on the same trials, she heeded feature FN (e.g., F1). Otherwise, she did not heed this dimension. Overall, the transfer phase yielded 8 binary heeding scores per participant, organized in 4 mutually exclusive pairs.

After the transfer trials, each participant entered the final *reconstruction* phase, in which they were asked to reconstruct five "good" examples of each disease. Each reconstruction trial began with a display depicting a number of "healthy" (grey) cells, and the participant's task was to adjust the number of diseased cells, as well as their length, number of hairs, and number or organelles. A graphical user interface with four sliders allowed the participant to adjust each dimension independently. Each slider[1] supported values from 0 to 14. This range extended beyond the values that were sampled during training (3 – 10). All sliders were at their 0 (leftmost) position at the beginning of a reconstruction trial.

The reconstruction data were scored separately for each stimulus dimension. We defined *reconstruction scores* so that 0 indicates the mean of all training exemplars of the specific disease whose "good" example is being constructed, and the relationally correct direction is encoded as positive. This allows to average the scores across the two diseases. The units are unchanged, so that +1 indicates one extra cell (or organelle, etc.) than the central tendency of the training set. Averaging across the 5 reconstruction trials for each disease produced 4 *aggregated reconstruction scores* per participant – one for each stimulus dimension.

Results and Discussion

Training Phase

Sixteen participants reached the high criterion in the Relation condition and 18 did in the Feature condition. In other words, there was approximately 50% attrition in both groups. This is a serious but, we hope, not a fatal problem. We speculate that

[1] The length was rescaled linearly to the 0 – 14 range. One new unit corresponds to 12.5 pixels.

the high attrition rate is predominantly due to insufficient motivation of many participants working on-line from home during a pandemic, rather than to some intrinsic difficulty of our categorization task. This hopeful interpretation is consistent with the comparable attrition rate in the two groups. Furthermore, the transfer data indicated that all 38 people who failed to reach the high criterion heeded at most 1 of the 8 attributes, and 22 of them did not heed any attribute. It seems that many participants paid just enough attention to satisfy the low inclusion criterion, but not enough to engage seriously with the task. Others found an attribute that yielded 75% correct and settled with it, ignoring the instructions that clearly stated that perfect accuracy was possible.

Our sample size, which was limited to begin with, was cut in half by attrition. Mindful of the low statistical power that results from this unfortunate circumstance, we are reluctant to draw strong conclusions from the present data. We report them mostly as a preliminary exploration inviting further research with adequate sample sizes.

For comparison with the study of Kittur et al. (2004), we report descriptive statistics about trials-to-criterion, which was their main dependent measure. For the 34 participants who reached the high criterion in our sample, the mean time-to-criterion was virtually the same in both groups: $M = 217$. The standard deviations were nearly equal as well: $SD = 119$ for Relation, 113 for Feature. This is consistent with the results of Kittur et al. (2004) – their deterministic relational and deterministic featural conditions took comparable times to learn. As a reminder, our training environment included one deterministic attribute in all groups.

Transfer Phase

We included a transfer phase in our experimental design (Fig. 1) with two related purposes in mind: as a manipulation check and as a probe into the strategies adopted by individual participants. We tabulated how many participants *heeded* various combinations of stimulus attributes (in the technical sense defined by the transfer scores described earlier).

The participants who reached the high criterion tended (in aggregate) to heed considerably more attributes than those who did not reach the criterion. Concretely, 2, 7, 4, 1, and 2 participants in the Relation condition heeded 0, 1, 2, 3, and 4 attributes, respectively; the corresponding counts in the Feature condition were 4, 9, 3, 2, and 0. By contrast, 22 of the participants who failed to reach the criterion heeded no attribute, and the remaining 16 heeded a single attribute.

An important metric for any participant was whether or not they heeded the deterministic attribute for their training environment. Consider first the 36 participants who were assigned to the Relation group. Fourteen of them heeded their subgroup-specific deterministic relation and – remarkably – all these 14 had reached the high criterion during training. By contrast, 20 of the 22 participants who did not heed their deterministic relation had not reached the criterion either. Thus, with only a couple of exceptions (one of which was a near miss), the necessary and sufficient condition for reaching the high criterion in the Relation training

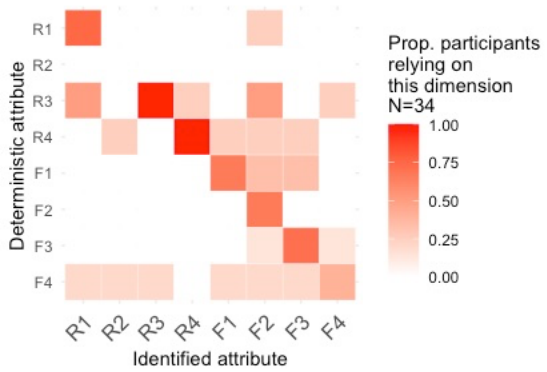

Figure 3. Proportion of the participants assigned to each training environment (rows) who *heeded* the attributes arrayed along the columns, as determined at transfer. R1~R4 and F1~F4 denote Relation and Feature subgroups.

environment was to identify and heed the appropriate deterministic relation. Turning to the Feature environment, the data suggest that heeding the subgroup-specific deterministic feature was a sufficient but not a necessary condition for reaching the high criterion. Concretely, 12 of the 36 participants in the Feature group heeded their deterministic feature, and 11 of them had reached criterion. Of the 24 people who didn't heed the deterministic feature, 7 had reached criterion and 17 had not.

From this point on, we focus our attention on the 34 participants who reached the high criterion. This restriction is presupposed throughout the remainder of the article.

The distinction between relations and features is of special interest. Nobody heeded more than 2 relations or more than 2 features. This seems to rule out most *pure* probabilistic strategies in light of the necessity to pool information across 3 non-deterministic dimensions to achieve perfect accuracy under our design (cf. Table 1).

Figure 3 explores the possibility of *mixed* strategies. The 8 rows correspond to the 8 training environments. The Relation (R1-R4) and Feature (F1-F4) subgroups are shown in the upper and lower halves, respectively. The colored squares encode the fraction of the participants in each subgroup who heeded the 8 attributes arrayed along the columns. The "hot band" along the diagonal indicates that most people heeded their subgroup-specific deterministic attribute. Note that the lower left-hand quadrant is mostly blank. This indicates that the participants in the Feature environment tended (in aggregate) to heed features at the expense of relations.

Reconstruction Phase

The reconstruction phase was designed to probe the internal representations used by the participants who successfully learned the task (as indicated by reaching the high criterion). Recall from the Method that each participant received an *aggregated reconstruction score* for each of the 4 stimulus dimensions.

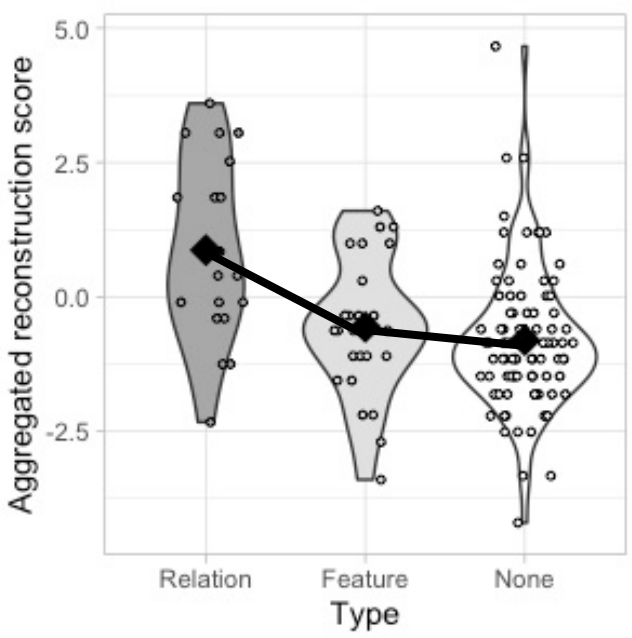

Figure 4: Aggregated reconstruction scores plotted as a function of the Type factor inferred from the transfer data. Each of the 34 participants contributes 4 points to the plot – one per stimulus dimension. Diamonds depict means

Our approach is to analyze the reconstruction data as a function of the strategies inferred from the transfer data (Fig. 1). The unit of analysis is the individual stimulus dimension (e.g., the number of cells in a micrograph). The transfer data yielded a pair of heeding scores per dimension. There are three possibilities: a given participant can heed the relational aspect of a dimension, or heed its featural aspect, or not heed it at all. We can thus re-conceptualize the transfer data as defining a factor with three levels: Relation, Feature, and None. This factor is referred to as *Type*. Each participant contributes 4 data points to the analysis – one per stimulus dimension – and each of them has a definite type.

Now, the extreme-value hypothesis predicts that stimulus dimensions that are encoded as relations in the category representation would tend to be exaggerated during reconstruction compared to dimensions that are not so encoded. Under our scoring scheme, this translates into a prediction of greater reconstruction scores for dimensions of type Relation compared to the other two types. In other words, the extreme-value hypothesis predicts a main effect of the Type factor. Furthermore, stimulus dimensions that are encoded as features in the category representation are predicted to be reconstructed near the central tendency of the training sample. Thus, the reconstruction scores are predicted to be near zero (which is the code for the central tendency).

The violin plots in Figure 4 suggest these predictions are borne out in the present data. The means of the three clusters of dots (depicted by bold diamonds in the figure) seem to conform to the profile predicted by the extreme-value hypothesis. An ANOVA analysis points to the same conclusion. There is a significant main effect of the Type factor (F (2, 132) = 12.860, MSE = 23.893, $p < .001$). A planned comparison revealed that the reconstruction scores are significantly greater for type Relation than Feature (t (37.424) = 3.275, $p < .003$). Another comparison revealed greater scores for type Relation than None (t (26.523) = 4.374, $p < .001$). Although closer to 0 than the other two types, the reconstruction scores for type Feature were significantly different from 0 (t (25) = –2.227, $p < .035$; 95% confidence interval from –1.078 to –0.04). Finally, note that the concentration of points within the None violin in Figure 4 reflects the fact that many people tended not to heed many stimulus dimensions.

General Discussion

The mental representation of relation-based concepts remains poorly understood in cognitive science. However, previous findings suggest that whereas the "goodness" of a member of a feature-based concept is a function of its similarity to the prototype of that concept, at least some relation-based concepts seem to favor extreme members (Goldwater et al., 2011; Kittur et al., 2006; Rein et al., 2010).

This difference was supported by our data. We ran a category learning experiment designed to test the hypothesis that people who learn feature-based representations of our categories will reconstruct those categories according to their feature-based mean values (as is well established in the literature; Murphy, 2002). However, people who learn relation-based representations of categories will represent those categories in a manner that is biased in the direction of the category-defining relation(s) (Kittur et al., 2006). As predicted, participants whose transfer scores indicated that they tended to categorize exemplars based on a relation tended to exaggerate that relation in their reconstruction of category members. For example, if a participant noticed that number of diseased cells in disease A tended to be larger than the number of healthy cells, then in the reconstruction of an exemplar of disease A, that participant would tend to make the number of diseased cells in the reconstructed sample larger than it had typically been in training. By contrast, also as predicted, a participant whose transfer performance suggested a feature-based strategy tended to reproduce values closer to the mean value of the relevant feature(s) during the reconstruction phase.

It seems difficult to account for the observed profile of reconstruction scores in Figure 4 in terms of any simple kind of contrast learning mechanism (e.g., Davis & Love, 2010). Recall that the deterministic features were defined by contrasting values (4 vs. 8) on all the stimulus dimensions. The theory that the representation of one category is repulsed away from (i.e., maximizes its contrast with) the alternative category cannot explain why these values seem to be drawn closer together in the case of participants who responded to the exemplars' features, whereas they were pushed further apart in the case of participants who responded to the exemplars' relations to the "healthy" cells.

An interesting aspect of the data from the transfer task (Figure 3) is that participants' strategies tend to cluster

around the broad categories Relation or Feature, as a function of which kind of attribute (a relation or a feature) was deterministic during their training, even though only one of those attributes was deterministic. For example, if a given participant, *p*, experienced *relative number* (of diseased vs. healthy cells) as deterministic, then *p* tended to focus on relative properties on all the other dimensions. But if *p* experienced that absolute number of disease cells (a feature) was deterministic, then *p* would tend to rely on the specific feature values of the other dimensions as well. This focus on Relations (or Features) *generally* as a result of experience with a single deterministic relation (or feature) is reminiscent of Vendetti et al.'s (2014) "relational mind-set": Having experienced one useful relation (or feature), people seem biased to look for other useful relations (or features).

Last but not least, we acknowledge two limitations of the current study: First, the statistical power is low due to the small sample size. Second, the reconstruction data were analyzed in terms of individual differences in a quasi-experimental design. We attribute the relationship between the two dependent measures in Figure 4 as evidence of a common unobserved cause. We hypothesize that this common cause is the internal representation depicted in Figure 1. However, in principle the observed relationship can stem from some other internal factor such as the propensity to adopt a relational mind-set. For these reasons, the current results are preliminary and should be interpreted with care.

Much remains unknown about the representation of relation-based concepts, but the present data contribute to an emerging picture of relation-based concepts as being quite different from their feature-based counterparts (Gentner & Kurtz, 2005). It is well known that associative learning algorithms, which excel at feature-based learning, have difficulty *extrapolating* outside the space of their training examples. By contrast, symbolic functions (such as mathematical equations), which are inherently relational in nature, often extrapolate to very broad set of potential input/output pairings. Many participants extrapolated beyond the training data in our reconstruction task, generating values for stimuli that were pushed *away* from the prototype in the direction that exaggerated the deterministic relation.

References

Barsalou, L. W. (1985). Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*(4), 629-654.

Davis, T., & Love, B.C. (2010). Memory for category information is idealized through contrast with competing options. *Psychological Science, 21*, 234-242.

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.

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

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology, 15*, 1–28.

Goldstone, R.L., Steyvers, M., & Rofosky, B.J. (2003). Conceptual interrelatedness and caricatures. *Memory & Cognition, 31*, 169– 180.

Goldwater, M. B., Bainbridge, R. & Murphy, G. L. (2016). Learning of role-governed and thematic categories. *Acta Psychologica, 164*, 112-126.

Goldwater, M. B., Don, H. J., Krusche, M. J., & Livesey, E. J. (2018). Relational discovery in category learning. *Journal of Experimental Psychology: General, 147*, 1-35.

Goldwater, M. B., Markman, A. B., & Stilwell, C. H. (2011). The empirical case for role-governed categories. *Cognition, 118*(3), 359-376.

Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review, 110*, 220–264.

Jung, W. & Hummel, J. E. (2015a). Making probabilistic relational categories learnable. *Cognitive Science*, *39*(6), 1259-1291.

Jung, W. & Hummel, J. E. (2015b). Revisiting Wittgenstein's puzzle: Hierarchical encoding and comparison facilitate learning of probabilistic relational categories. *Frontiers in Psychology*, *6*:110, 1-11.

Kim, S. W. & Murphy, G. L. (2011). Ideals and category typicality. *Journal of Experimental Psychology:Learning, Memory, and Cognition, 37*(5), 1092-1112.

Kittur, A., Holyoak, K. J., & Hummel, J. E. (2006). Ideals aren't always typical: Dissociating goodness-of-exemplar from typicality judgments. In R. Sun & N. Miyake (Eds.), *Proc. Of the 28th Annual Conference of the Cognitive Science Society* (pp. 429–434). Mahwah, NJ: Erlbaum.

Kittur, A., Hummel, J. E., & Holyoak, K. J. (2004). Feature- vs. relation-defined categories: Probab(alistic)ly not the same. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proc. Of the 26th Annual Conference of the Cognitive Science Society* (pp. 696–701). Mahwah, NJ: Erlbaum.

Murphy, G. L. (2002). *The Big Book of Concepts*. Cambridge, MA: MIT Press.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psych. Review*, *92*, 289-316.

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.

Rein, J. R., Goldwater, M. B., & Markman, A. B. (2010). What is typical about the typicality effect in category-based induction? *Memory & Cognition, 38*(3), 377-388.

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