Heuristics for Choosing Features to Represent Stimuli

Matthew D. Zeigenfuse (mzeigenf@uci.edu)
Michael D. Lee (mdlee@uci.edu)
Department of Cognitive Sciences, University of California, Irvine
Irvine, CA 92697 USA

Abstract

In this paper, we compare three heuristic methods for choosing which of a set of features to use to represent a domain of stimuli when we know the categories to which those stimuli belong. Our methods are based on three measures of category differentiation: cue validity, category validity, and their product, collocation. In a comparison of their ability to predict human similarity ratings in the Leuven Natural Concept Database, we find collocation to have the best performance, suggesting people use both cue and category validities in choosing which features to represent.

Keywords: Feature representation; basic-level categorization; similarity judgment.

Introduction

Of all the aspects of their world that could be represented, which do people actually choose? Imagine you are standing in front of a black dog named "Rover" with a small white patch of hair under its left eye. Which of its features do you choose to represent: its tail and four paws, its name, "Rover", and the spot under its eye? The last two of these may be useful for a representation of the family dog, but are probably less useful to representing dogs as whole. Conversely, the first two may be useful for representing dogs, but are probably less useful for distinguishing Rover.

One method of learning about which aspects of a particular set of concepts people represent is the feature generation task (Rosch & Mervis, 1975). Often in this task people are asked generate a fixed number of features for each exemplar in a domain. Moreover, in some cases, additional participants are asked to rate whether an exemplar has a feature for each combination of features and exemplars in a domain (Deyne et al., 2008). This leads to a large number of features describing each exemplar; however, not all of these features will be important to a person's representation.

Zeigenfuse and Lee (2008, in press) provide a partial computational-level solution to the problem. Similar to the theory of second-order isomorphism in perception (e.g. Shepard & Chipman, 1970), they argue that people represent those features that determine the similarity between objects and develop a model to infer which features are important using similarity judgments. Unfortunately, it does not offer a psychological rationale for why one feature is important vis-à-vis an unimportant one.

The goal of this paper is to provide an algorithmic-level solution to this problem. To this end, we propose heuristic methods for choosing important features by how useful they are in distinguishing categories in the category structure the objects-in-question are embedded and vice-versa. We use these heuristics to begin answering the question of *why* people represent one features versus another.

Representation and Basic-Level Categories

Our heuristics are based on measures of category differentiation that have been proposed to explain basic-level categorization. Basic-level phenomenology refers to people's preference to categorize objects at a particular level in a category hierarchy, known as the basic level. Researchers have found that objects are categorized into basic-level categories more quickly than sub- or super-ordinate categories, basic level objects are named faster, objects are described preferentially with basic level names, more features are listed at the basic level than at the superordinate level, basic level names are learned before names at other levels, and basic level names tend to be shorter (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). These results suggest an intimate relationship between an object's basic-level category and its mental representation.

Category-Based Measures

Category Differentiation Given a feature representation, many theories of basic-level categorization score potential categorizations of the concepts in a domain through the information its categories give about the features of category members and vice-versa. Examples include, cue validity (Rosch et al., 1976), category validity, collocation (Jones, 1983), feature predictability (Corter & Gluck, 1992), category statistical density (Kloos & Sloutsky, 2006), and strategy length and internal practicability (SLIP: Gosselin & Schyns, 2001). Inverting this logic, given a set of categories, we can score features on their usefulness in providing information about which of the set of categories a concept belongs to, the information knowing a concepts category provides about whether it has the feature, or a mixture of the two.

Usefulness Measures The heuristics described here for choosing feature representations around three measures of usefulness. The first of these is *maximum cue validity*, defined as the maximum over categories c_j of cue validity, the probability an exemplar belongs to c_j given that it has a feature f, $p(c_j|f)$. We also look at *maximum category validity*, defined similarly as the maximum over categories c_j of the category validity, the probability an exemplar has a feature f given that it belongs to c_j, $p(f|c_j)$. Finally, we look at *maximum collocation*, the maximum over categories c_j of the collocation, the product of a feature's cue and category validities, $p(c_j|f)p(f|c_j)$.

Alternative Measures

We supplement these category-based heuristics by two additional heuristics, included as baselines. The first of these is based around a measure we term *feature prevalence*, defined to be the proportion of exemplars in a domain which possess a given feature. The purpose of this heuristic is to compare the category-based heuristics to a simpler heuristic using only base-rate information. The second is a "random" heuristic, which simply selects subsets of features at random. This heuristic gives an idea of how the category-based heuristics compare to heuristics at large.

The remainder of the paper compares the five heuristics using human similarity judgments. We procede as follows. First, we describe the data on which the heuristics will be compared, the Leuven Natural Concept Database (Deyne et al., 2008), a collection of normative data for semantic concepts. We then present the selection heuristics and how the representations chosen are used to generate similarity judgments. Next, we show the results of applying the heuristics to the Leuven database. We close by discussing what these results tell us about the features people choose to represent stimuli and the difference between natural and artificial kinds.

The Leuven Natural Concept Database

The Leuven Natural Concept Database (Deyne et al., 2008) contains normative data for semantic concepts falling into one of two domains, animals and artifacts. These data consist of typicality ratings, goodness ratings, goodness rank orders, generalization frequencies, exemplar associative strengths, category associative strengths, estimated ages of acquisition, word frequencies, familiarity ratings, imageability, and pairwise similarity ratings for concepts within a single category as well as exemplar-by-feature matrices and pairwise similarity ratings between a subset of the exemplars in a domain spread across its categories.

In our comparisons we make use of the exemplar-by-feature matrices and domain similarity ratings. The exemplar-by-feature matrices describe the exemplars of a domain in terms of a number of participant-generated features. For the animals domain, 129 exemplars, split among the categories birds, fish, insects, mammals, and reptiles, are described in terms of 765 features. For the artifacts domain, 166 exemplars, split among the categories clothing, kitchen utensils, musical instruments, tools, vehicles, and weapons, are described in terms of 1295 features. These features include both high frequency features such as "is a bird" and "is made of metal" and low frequency features such as "stands in the crib at Christmas" and "stored in the cellar".

Domain similarity judgments are pair-wise similarity judgments collected between exemplars in a set of consisting five exemplars from each of the categories in a domain. This results in sets of twenty-five exemplars for the animals domain and sets of thirty exemplars for the artifacts domain. Two distinct sets of exemplars were chosen for each domain, resulting four sets of domain similarity judgments.

Feature Selection Measures

Starting with a set of features that we wish to select a feature representation from (such as the 765 animal or 1295 artifact features in the Leuven sets), each heuristic chooses a feature representation using a two step process. First, the usefulness of each feature is computed under a particular usefulness measure. Then, we select those features whose usefulness is above a pre-defined threshold. For example, suppose we wish to use the collocation heuristic to choose among the seven features representing the exemplars of the three categories in Table 1. First, we would compute the maximum collocation over categories for each of the features (shown in the "Colloc." column of Table 1). Then, we would select all those features for which the maximum collocation over the categories was above our threshold. In this example, were the threshold one-half, we would select features 1, 2, and 3. The same procedure can be used with the benchmark importance measure to select a representation.

The features selected by these heuristics to generate similarities according to a common features model (Shepard & Arabie, 1979). Let U be a set of useful features. The common features model says that similarity between concepts i and j is

$$s_{ij} = c + \sum_{f_k \in U} w_k f_{ki} f_{kj}, \tag{1}$$

where c is the universal similarity and w_k is the salience of feature f_k.

The remainder of the section is devoted to discussing for the benchmark and other heuristics in greater detail. In the first subsection, we summarize the benchmark measure of importance. In the second, we provide a rationales for each of the three category-based usefulness measures. In the final subsection, we provide rationales for the two baseline heuristics.

Benchmark

The Zeigenfuse and Lee (2008, in press) method for learning which of a set of features people use to represent stimuli is based upon latent variable selection. In this framework, those features that are included in a concept's representation are termed "important" features. For each feature, they define a variable z_k indicating whether feature f_k is used in similarity judgments. If we have n total features, the similarity between concepts i and j is then

$$s_{ij} = c + \sum_{k=1}^{n} z_k w_k f_{ki} f_{kj}. \tag{2}$$

To learn which features are included in the representation, Zeigenfuse and Lee develop a Bayesian model and sample from the marginal posterior over the z_k using Markov Chain Monte Carlo (MCMC). In this framework, a feature's importance is the marginal posterior probability the feature is represented. They found that a small number of important features are able to fit similarity almost as well as using all features.

	Category 1					Category 2			Category 3							Cue	Cat.	Colloc.
Feature 1	●	●	●	●	●											1	1	1
Feature 2	●	●	●	●	●	●										5/6	1	5/6
Feature 3	●	●	●	●												1	4/5	4/5
Feature 4			●													1	1/5	1/5
Feature 5	●	●	●	●	●	●		●			●	●	●	●		5/11	1	5/11
Feature 6	●	●	●	●	●				●	●	●	●	●	●	●	5/12	1	5/12
Feature 7							●		●			●				2/3	1/3	4/21

Table 1: Representative features illustrating behavior of the usefulness measures.

Usefulness Measures

Different measures of usefulness correspond to different assumptions about what aspects of the environment lead a person to represent a particular feature. In the opening example, the small white spot under the dog's eye and its name, "Rover", may be useful for representing the family dog, but are probably not useful for representing dogs generally. The goal of this section is to give rationales for why each of the category-based heuristics, maximum cue validity, maximum category validity, and maximum collocation, may pick out good representations.

Maximum Cue Validity Maximum cue validity measures how concentrated a feature is in a single category. Formally, let M be the total number of objects with a particular feature f and let N_j be the number of objects with the feature in category c_j. The cue validity of f with respect to category c_j is then $p(c_j|f) = N_j/M$ and the maximum cue validity is

$$u_{cue} = \max_j \left\{ \frac{N_j}{M} \right\}. \quad (3)$$

As illustrated by example features Table 1, maximum cue validity is large when most of the exemplars possessing a feature belong to the same category (Features 1 – 4), though this need not be a large number of exemplars (Feature 4). To see why, note that u_{cue} is large if and only if there exists a category for which N_j is nearly M. Since $N_{j'} \leq M - N_j$ for $j' \neq j$ and $M - N_j$ must be small, few exemplars can belong to $c_{j'}$.

Maximum Category Validity Category validity measures how diffuse a feature is within a particular category. As with maximum cue validity, let N_j be the number of exemplars in category c_j with feature f, and define a new quantity K_j to be the total number of exemplars belonging to c_j. Then, the category validity of f with respect to category c_j is $p(f|c_j) = N_j/K_j$ and the maximum category validity is

$$u_{cat} = \max_j \left\{ \frac{N_j}{K_j} \right\}. \quad (4)$$

Returning to Table 1, we see that features whose category validity is high (Features 1, 2, 5, and 6) are possessed by most of the exemplars in at least one category.

Maximum Collocation Maximum collocation is a measure of how simultaneous concentrated in and diffuse across a category a feature is. Using the terminology of the previous sections, the collocation of a feature f with respect to category c_j is

$$u_{col} = \max_j \left\{ \frac{N_j}{M} \times \frac{N_j}{K_j} \right\}. \quad (5)$$

Features with high collocation are possessed by most exemplars within a category and few outside it, as illustrated by the architypical Feature 1 in Table 1. Alternatively, Features 4 and 6 show why it is necessary for both of these to be true. Those features possessed by only a small fraction of exemplars within a single category will have high cue validity but low category validity (Feature 4). Those features possessed by most exemplars in more than one category will have high category validity but low cue validity (Feature 6).

Alternative Measures

The two baselines used here are intended to show both how well the category-based heuristics performed against heuristics derived from contrasting assumptions. The first baseline heuristic, feature prevalence, serves as a comparison to heuristics employing solely frequency information through a representative heuristic. The second baseline, the so-called random heuristic, serves to compare the category-based heuristics to heuristics generally.

Feature Prevalence Prevalence assumes that the more times a feature is encountered in the world, the more likely it is to be included in representations. Oft occurring features are not necessarily the most useful for distinguishing between categories. If usefulness in distinguishing categories is good indicator of whether a feature is represented, representations chosen on the basis of prevalence should provide poor fits to human similarity judgments.

Random The random heuristic can be thought of as selecting a heuristic at random from the set of all possible heuristics for choosing feature representations. By comparing the category-based heuristics to the random heuristic, we intend to show how well these heuristics perform against an arbitrarily one. Good heuristics will choose features that predict similarities better than an arbitrary heuristic. Should the category-based heuristics be good ones, we expect them to predict similarities better than the random heuristic.

Method Comparison

Here we describe a comparison of maximum cue validity, maximum category validity, and maximum collocation to each other as well as the benchmark and baselines using the Leuven Natural Concept Database (Deyne et al., 2008). In the first section, we enumerate the procedure used to fit the domain similarity data. In the second, we present the results of this procedure for each of the heuristics.

Procedure

The fit procedures begins with the exemplar-by-feature matrices. Before applying any of the heuristics we filter out all features possessed by zero, one, or all of the 25 or 30 exemplars included in the domain similarity comparisons, since the weights for these features are not identified under a common features similarity model. Additionally, we find all features that are possessed by exactly the same exemplars, also for just the domain similarity exemplars, and remove all but the most frequently generated representative for each unique pattern.

After this pre-processing, for the benchmark and all of the heuristics except the random heuristic, we compute its corresponding measure using all of the exemplars in the domain, not just those included in the domain similarity judgments. The features are then sorted in order of decreasing value on these measures. Starting with only the top two features, we fit the common features model to the domain similarity judgments using Non-Negative Least Squares (NNLS) and compute the correlation between the fitted similarities and the actual similarities. We repeat this process with the top three features, the top four features, etc. For the random heuristic, we generated 100 random feature orders and apply this procedure to each of the orders.

Results

Figure 1 shows the correlation between observed and those fitted using the first x percent of features ordered by either cue validity, category validity, collocation, prevalence, or the benchmark. For example, on the collocation line (shown as a solid line) the correlation at a percentile rank of 20 percent is the correlation between the observed values and those fitted using the first 20 percent of features ordered by collocation. The smaller pane in the lower right-hand corner is a blowup of the lines in rectangular region extending from $0-20$ in percentile rank and from $0.6-1$ in correlation.

The gray shaded area shows 95% confidence intervals for the correlation between the values fitted using first x percent of features chosen by the random heuristic and the observed values. These orders give an estimate of how difficult the similarity data are to fit with a heuristic choosing x percent of the available features. A heuristic whose correlation is above the upper limit of the area fits better 95 percent of heuristics at that percentage of features. Alternatively, a heuristic whose correlation is below the lower limit of the area fits worse than 95 percent of heuristics at that percentage of features.

Regardless of data set, the orders produced by the Zeigenfuse and Lee (2008, in press) measure is always able to fit the similarities in the top 5 percent of ordering, justifying its use a benchmark. The orders produced by feature prevalence nearly always perform worse than those generated by the other measures, often in the worst 5 percent of all orders. On the whole, cue validity, category validity, and collocation perform middling to well, rarely performing worse than feature prevalence.

For the animals data sets, cue validity outperforms category validity for small numbers of features (less than around 20 percent), category validity outperforms cue validity for larger numbers of features, and collocation is always commensurate to the best of these. For very small (less than around 10 percent) numbers of features, cue validity performs better than the benchmark; however, for larger numbers of features its performance is at best mediocre. After a slow start, category validity performs in the top 5 percent of orderings for larger numbers of features. Collocation always performs near the benchmark and is nearly always in the top 5 percent of orderings.

For the artifacts data sets, cue validity still performs better than category validity for very small (less than 10 percent) numbers of features, after which category validity performs better than cue validity. As with animals, collocation performs near or better than the best of these two measures. Category validity and collocation nearly always perform between the 5^{th} and 95^{th} quantiles of heuristics; however, for larger numbers of features (around 20 percent in the first set and around 40 percent in the second), cue validity performs in the bottom 5 percent of orderings.

Overall, these results suggest that both cue and category validity contain information about which features are most indicative of people's similarity judgments. Collocation always performs about the same as the best of cue and category validity, indicating that it tracks the best aspects of the two measures. This suggests that early on collocation is dominated by features with high cue validity, but later it is dominated by category validity.

Discussion

Cue and Category Validity

The major result of the previous section is that both cue and category validity seem to be important to choosing which of a set of features makes a good representation. In the case of cue validity, Murphy (1982) showed why this may be the case: cue validity cannot pick out basic-level categories because it can only increase for more inclusive categories. Consider the hierarchy of categories *physical object, animal, bird, duck*, in which *bird* is the basic-level category, and suppose we wish to compute the cue validity of the feature "has wings". Let M be the number of things with wings and N_1, N_2, N_3, and N_4 be the number of ducks, birds, animals, and physical objects with

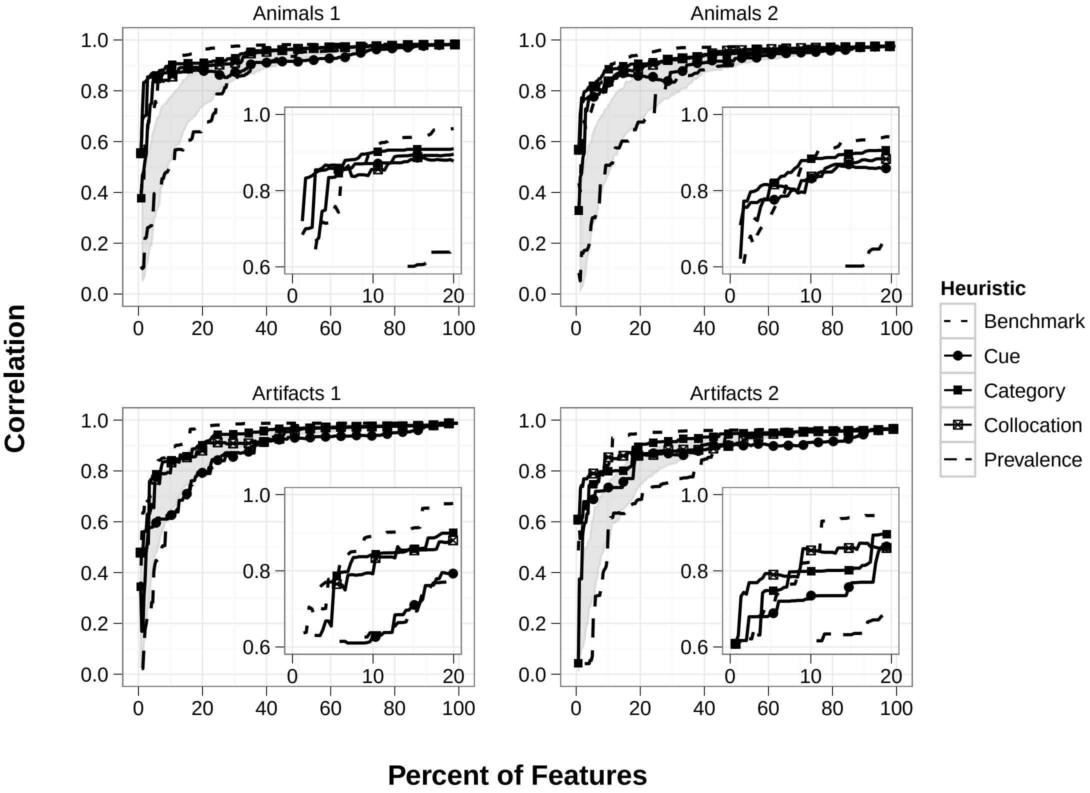

Figure 1: Model fit by the percent of features used for each of the four sets of domain similarities in the Leuven data set. The benchmark, three category-based heuristics, and feature prevalence baseline are shown as lines. In the legend, "collocation" corresponds to the maximum collocation heuristic, "benchmark" to the benchmark, "cue" to maximum cue validity, "category" to maximum category validity, and "prevalence" to feature prevalence. The gray area shows a 95% confidence interval for the fit of the random heuristic. The panels in the lower righthand corner of each of the plots enlarges the rectangular region from $0-20$ in percent of features and from $0.6-1$ in correlation in the main plots.

wings. Since the categories are inclusive, $N_1 \leq N_2 \leq N_3 \leq N_4$, so $N_1/M \leq N_2/M \leq N_3/M \leq N_4/M$. But N_i/M is just the cue validity of "has wings" for each category in the hierarchy, illustrating why, in settling on basic-level categories, people must be sensitive to more than just cue validity. Since similarity is assumed to reflect representation, this should be reflected in measures used to select representations.

Along these lines, Tenenbaum and Griffiths (2001) offer a fuller explanation for why both cue and category validities should be important to choosing good representations. They argue that people generalize properties to novel instances only in the smallest set of instances consistent with known examples, a theory known as the "size principle", and further that similarity is the degree to which the consequences of being one object generalize to another. By this logic, choosing features on the basis of cue validity will lead to categories which are overly restrictive and choosing features on the basis of category validity will lead to categories which are overly broad. Appropriate generalization, then, requires taking both types of information into account. Thus, we would expect a heuristic that does this, like collocation, to choose better representations than heuristics that do not.

Natural Versus Artificial Kinds

A final point worth mentioning is the difference in performance of the heuristics on data sets containing natural kinds versus those containing artificial kinds. Numerous authors have suggested that natural and artificial kinds are represented in fundamentally different ways (e.g. Keil, 1989). Results of Zeigenfuse and Lee (in press) support this theory, finding the ratio between the probability two stimuli within the same category have a feature and the probability two arbitrarily chosen stimuli have a feature is larger for natural kinds than artificial ones.

Here we find a similar result: for animals data sets collocation nearly always performs in the top 5 percent of heuristics, whereas for artifacts data sets, collocation performs about as well as an arbitrary heuristic. In theory this difference could come from either differences in the types of features represented or the ability of the common features model to fit similarity judgments among exemplars of that domain. The latter seems unlikely, however, given that the benchmark performs well for all four data sets it seems a common features similarity model is able to fit the data well.

This, then, suggests that the difference in fits comes from differences in the types of features people choose to represent. Among animals, people prefer features that are closely tied to a particular basic category. Among artifacts, they seem to prefer a different strategy, representing features for multiple levels in a category hierarchy or selecting features using different criteria.

Extensions

A detailed explanation of this difference may requires extensions addressing one of both of these sources. The first of these begins from the recognition that the source of the apparent distinction between natural and artificial kinds may stem not from an actual difference but from an incorrect choice of selection heuristic. Thus, it makes sense to look at heuristics based on additional measures of category differentiation. The second supposes choosing just those features associated with basic-level category structure is not sufficient for selecting good feature representations.

Additional Heuristics In order to explore the first of these extensions, we could develop heuristics based on different measures, both those that have been proposed in the basic-level literature and outside it. Such measures could include the category likelihood ratio (Zeigenfuse & Lee, in press), the mutual information between a category and a feature SLIP (Gosselin & Schyns, 2001). These last of these differs from the first two in that, in the first, each feature affects the quality of a categorization independent of all other included, whereas in the second two the effect of adding a new feature depends upon the features already included.

Category Hierarchies The second extension allows the method to deal with category hierarchies. The importance of structured representation in understanding human judgments of similarity has been illustrated by many authors (e.g. Markman & Gentner, 1993). Understanding how such structured representations influence those features represented is a crucial step towards bringing these models into contact with feature-based models such as Tversky's contrast model (Tversky, 1977). One potential method for acheiving this would be to compute the collocation, or other measure, at each level in a category hierarchy and to use a weighted combination of the collocations as the selection criterion.

Conclusion

In this paper, we have presented three heuristic methods for choosing a feature representation based on measures of category differentiation. We find these heuristics to fit human data better than heuristics that do not take this information into accounts, acheiving very good fits for natural kinds and above average fits for artificial kinds. Moreover, our results suggest both how concentrated in a particular category a feature is and how diffuse it is across exemplars in that category are important factors in whether a feature is represented as well as supporting a distinction between natural and artificial kinds. Though much still needs to be done, this work suggests people choose features in a systematic way and that these regularities can be uncovered by investigating the relationship between categories and features.

References

Corter, J. E., & Gluck, M. A. (1992). Examining basic categories: Feature predictability and information. *Pyschological Bulletin*, *111*, 291-303.

Deyne, S. D., Verheyen, S., Ameel, E., Vanpaemel, W., Dry, M., & Voorspoels, W. (2008). Exemplar by feature applicability matrices and other Dutch normative data for semantic concepts. *Behavior Research Methods*, *40*(4), 1030-1048.

Gosselin, F., & Schyns, P. G. (2001). Why do we SLIP to the basic level? Computational constraints and their implementation. *Psychological Review*, *108*(4), 735-758.

Jones, G. V. (1983). Identifying basic categories. *Psychological Bulletin*, *94*, 423-428.

Keil, F. (1989). *Concepts, kinds, and cognitive development*. Cambridge, MA: MIT Press.

Kloos, H., & Sloutsky, V. M. (2006). What's behind different kinds of kinds: Effects of statistical density on learning and representation of categories. *Journal of Experimental Psychology: General*, *137*(1), 52-72.

Markman, A. B., & Gentner, D. (1993). Structural alignment during similarity comparisons. *Cognitive Psychology*, *25*, 431-467.

Murphy, G. L. (1982). Cue validity and levels of categorization. *Psychological Bulletin*, *91*(1), 174-177.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*, 573-605.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*, 352-382.

Shepard, R. N., & Arabie, P. (1979). Additive clustering: Representations of similarities as combinations of discrete overlapping properties. *Psychological Review*, *86*(2), 87-123.

Shepard, R. N., & Chipman, S. (1970). Second-order isomorphism of internal representations: Shapes of states. *Cognitive Psychology*, *1*, 1-17.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*(4), 629-640.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*, 327-352.

Zeigenfuse, M. D., & Lee, M. D. (2008). Finding feature representations of stimuli: Combining feature generation and similarity judgment tasks. In V. Sloutsky, B. Love, & K. McRae (Eds.), *Proceedings of the 30th annual conference of the cognitive science society* (p. 1825-1830). Austin, TX: Cognitive Science Society.

Zeigenfuse, M. D., & Lee, M. D. (in press). Finding the features that represent stimuli. *Acta Psychologica*.