A Dissociation between Categorization and Similarity to Exemplars

Nolan Conaway and Kenneth J. Kurtz
Department of Psychology, Binghamton University
Binghamton, NY 13905 USA

Abstract

Research in category learning has been dominated by a 'reference point' view in which items are classified based on attention-weighted similarity to reference points (e.g., prototypes, exemplars, clusters) in a multidimensional space. Although much work has attempted to distinguish between particular types of reference point models, they share a core design principle that items will be classified as belonging to the category of the most proximal reference point(s). In this paper, we present an original experiment challenging this distance assumption. After classification training on a modified XOR category structure, we find that many learners generalize their category knowledge to novel exemplars in a manner that violates the distance assumption. This pattern of performance reveals a fundamental limitation in the reference point framework and suggests that stimulus generalization is not a reliable foundation for explaining human category learning.

Keywords: categorization, generalization, formal modeling

Introduction

The history of research on human category learning is rich and complex. Whereas early studies explored the capacity for human learners to acquire rule-defined concepts (i.e., Bruner, Goodnow, & Austin, 1956), current research and theory is now largely centered around a 'reference-point' framework, where learners are thought to master categories by learning to associate stored perceptual referents (e.g., prototypes, exemplars) with individual categories. Reference point models of categorization (e.g., Kruschke, 1992; Love et al., 2004; Nosofsky, 1986; Smith & Minda, 2000) have enjoyed wide success in explaining human behavior, and are widely considered a definitive account of how categories are learned, represented, and applied.

Although specific reference point models differ from one another in a variety of ways, they tend to make generally similar representational and process assumptions. Chiefly, all reference point models assume that categories are represented by one or more points in a psychological space. On the extremes, a prototype model represents each category in terms of its central tendency, i.e., the average across known members, while an exemplar model would represent the category in terms of the individual items themselves. Successful reference point models employ a selective attentional mechanism that allows them to weight the importance of each stimulus dimension (Medin & Schaffer, 1978; Kruschke, 1992).

Discrepancies between reference point models have been the subject of extensive debate (e.g., Homa, 1984; Nosofsky, 1992; Smith & Minda, 2000), but at present we are interested in their common design principle: that learners categorize based on proximity to reference points associated with category responses. More specifically, reference point models assume that classification decisions are based on computing the similarity of a presented cue to stored reference points, typically following an inverse exponential function of geometric distance (Shepard, 1987).

An important feature of these models is that similarity can be attentionally-mediated, but the inescapable commitment is to stimulus generalization (Nosofsky, 1986). Although legitimate concerns have been raised about the validity of this *distance assumption* in psychological models (Medin, Goldstone, & Gentner, 1993; Rips, 1989), reference point accounts have remained leaders in the field of category learning due to superior quantitative fits to behavioral data (e.g., Kruschke, 1992; Medin & Schaffer, 1978; Nosofsky et al., 1994a). Indeed, there exist few examples of empirical phenomena in the artificial classification learning paradigm that are not well described in terms of distance to reference points.

The Current Study

We report an original experiment challenging the idea that people make classification responses using distance to stored reference points. Our experiment is based on specific predictions made by two contemporary models: ALCOVE (Kruschke, 1992), and DIVA (Kurtz, 2007). ALCOVE is an adaptive network model that straightforwardly embodies the central tenets of the reference point framework: ALCOVE uses error-driven learning to optimize attention weights mediating the similarity computation and association weights between the exemplar-based reference points and category nodes. ALCOVE has been tested thoroughly for its ability to account for behavioral data (Kruschke, 1992, 1993; Nosofsky et al., 1994a) and has remained a leading account of human category learning since its publication.

DIVA (Kurtz 2007) offers a similarity-based alternative to the reference point framework by representing statistical models of categories in a DIVergent Autoencoder. Rather than learning to associate reference points with category responses, DIVA learns how to correctly reconstruct presented cues on their category channels. Classification decisions are made based on the reconstructive error observed across category channels – if one of DIVA's channels is able to reconstruct a cue without much distortion, then the cue will likely be classified as a member of that category.

Both models rely on a form of similarity to guide classification (ALCOVE uses attention-weighted distance to exemplar reference points; DIVA uses a more implicit form of similarity in that inputs are more likely to be successfully

reconstructed if they are like known category members), and accordingly their predictions are often very much alike. However, the models differ in their commitment to distance-based classification. Since DIVA does not use reference points to represent categories, its responses can diverge from predictions made by distance-based accounts of categorization. We can therefore apply the two models to generate predictions about human classification performance relative to the distance assumption.

A Priori Simulations

In a set of a priori simulations using DIVA and ALCOVE, we compared generalization predictions after training on a variant of the well-known exclusive-OR (XOR) category structure. XOR categories are commonly studied in learning and machine learning research alike, and are defined by a logical rule operating across two or more dimensions. For example, one category might consist of white squares (*00*) and black circles (*11*), and the contrast category would then consist of black squares (*10*) and white circles (*01*).

XOR categories have played an important role in the literature on human category learning. In a classic study, Shepard, Hovland, & Jenkins (1961) measured the number of errors made during classification training on six elemental category types comprised of eight exemplars that vary in three binary dimensions. The second category type (Type II) represents the logical XOR structure over two dimensions, with a third irrelevant dimension. Shepard et al. found that Type II was learned second quickest of the six elemental category types—it was even learned more quickly than the Type IV categories which are linearly separable and adhere to a minimal version of family-resemblance (Rosch & Mervis, 1975). Formal modeling work later explained the ease of acquisition differences in terms of selective attention (Nosofsky et al., 1994a): by learning to ignore the irrelevant dimension, ALCOVE was the only model that could predict a strong Type II advantage. Subsequently, fitting the observed Type II advantage has been a leading benchmark for formal models of category learning.

In a detailed investigation of the Type II advantage, Kurtz et al. (2013) found that the ease of Type II acquisition varies markedly based on a number of methodological factors. For example, Type II learning was faster when learners are provided instructions that encourage rule formation and when stimulus dimensions are more easily verbalizable. Kurtz et al. (2013) argue for a revision of the general SHJ ordering along with recognition that models should be able to account for systematic variability in Type II acquisition

To date, nearly all work on XOR has represented the categories using binary stimulus dimensions. These stimulus sets, however, are limited by the lack of a generalization set. Consequently, it is difficult to satisfactorily address the role of distance in classification. Instead, we employ a two-dimensional, continuous adaptation of the XOR structure (see Figure 1) that maintains the overall logical structure of the categories while providing a generalization set.

In our simulations, DIVA and ALCOVE were tested for generalization performance after training on the continuous XOR categories. We observed that the two models generalized similarly and the predicted classification responses for both models were consistent with the distance assumption. However, we also tested a novel variation in which the trained set included a partial version of one of the categories (i.e., one of the four quadrants was left untrained, as in Figure 1) along with a standard version of the other category. We found that DIVA often generalizes as if it had been trained on the full version of XOR. That is, DIVA often makes the prediction that the 'one-quadrant' category generalizes to exemplars in the untrained quadrant. This prediction is particularly interesting because the critical test items are closer to the members of the 'two-quadrant' category: the central exemplar in the untrained quadrant is, on average, 1.67 city blocks away from members of the two-quadrant category, and 3 blocks away from members of the one-quadrant category (consistent results arise using a Euclidean metric, though we used a cityblock metric for the current study). Accordingly, reference point models like ALCOVE have no ability to produce this pattern of results—they instead predict that generalization will be based on the more proximal exemplars belonging to the two-quadrant category.

Behavioral Experiment

We designed a straightforward study to test the predictions

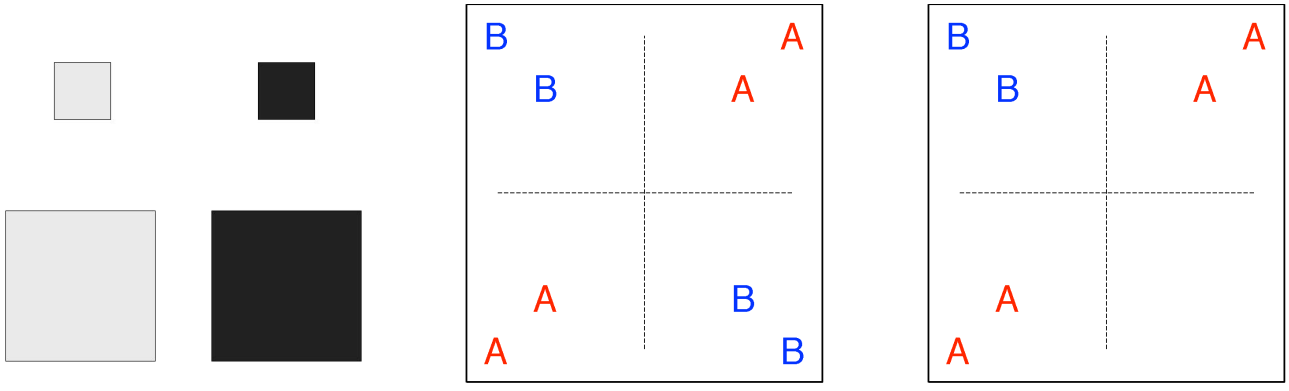

Figure 1. *Left*. Sample stimuli. *Middle*. Continuous, two-dimensional XOR categories. *Right*. Partial XOR categories.

made by DIVA and ALCOVE about generalization performance after learning the partial-XOR structure. Specifically, we were interested in the extent to which human learners would generalize the one-quadrant category to an untrained area of the stimulus space that is spatially closer to members of the two-quadrant category.

It is worth noting that our goal here is not just another example of the 'which model wins' approach. Instead, we are using these models to test assumptions about how classification decisions are reached: whereas ALCOVE presumes that responses are exclusively distance-based, DIVA is not theoretically committed to the distance assumption. We therefore use the models to evaluate the validity of the distance assumption in the psychology of category learning.

Participants and Materials. 61 undergraduates from Binghamton University participated in fulfillment of a course requirement. Stimuli were squares varying in shading and size (see Figure 1 for samples). These dimensions were chosen in order to maintain compatibility with 'standard' materials used in experiments involving XOR categories (e.g., Nosofsky et al., 1994a; Shepard et al., 1961). Exemplars were automatically generated at 7 positions on each dimension (7 shading * 7 line spacing = 49 examples). The assignment between perceptual and conceptual dimensions was randomly counterbalanced across participants.

Procedure. Each participant was randomly assigned to receive training on either the full or partial XOR category structure. In both conditions, participants completed 96 training trials (12 blocks consisting of the 8 training examples). In order to equate block size in the partial condition, the one-quadrant category exemplars were presented twice within each block. This way of handling the unbalanced category structure raises the issue of exemplar presentation frequency (Nosofsky, 1988), though little is known about how presentation frequency affects generalization. After training, participants completed 49 generalization trials consisting of items sampled at 7 positions on each dimension. All of the training examples were included (intermixed).

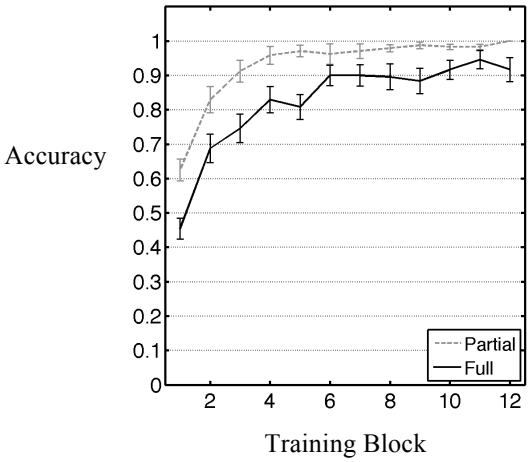

Figure 2. Aggregate training accuracy for the partial and full XOR categories.

Participants were informed that there would be test trials prior to beginning the experiment. The instructions did not encourage learners to engage in hypothesis testing to discover a rule. On each trial, a single stimulus was presented on a computer screen and learners were prompted to make a classification decision by clicking one of two buttons (labeled 'Alpha' and 'Beta'). During the training phase, learners were given feedback on their selection. Feedback was not provided during the generalization phase.

Results. One participant was excluded from analysis due to experimenter error leaving 30 participants in each condition.

Not surprisingly, the two category structures differed in terms of ease of acquisition (see Figure 2). Specifically, the partial XOR categories were learned more quickly than the full XOR categories, $t(58) = 4.06$, $p<0.001$, $d=1.03$. However, most of the learners in both conditions showed evidence of mastery of the categories by the end of training.

Our primary focus is on the generalization data. Each participant's set of responses in the test phase yields a 7x7 generalization gradient of classification performance. These data revealed a variety of individual differences in classification strategies. To formally profile each learner's generalization responses, we compared each gradient to a set of templates or idealized gradients representing idealized patterns of responses under different possible classification strategies. Each learner was profiled based on finding the template that best matched their performance according to a mean-squared error, MSE, metric.

In the full XOR condition, we identified two prevalent generalization profiles: 1) systematic XOR responses, reflecting mastery of the categories (*Learners*), and 2) random responses, reflecting failure to master the categories (*Non-Learners*). This dichotomy fits nicely with evidence from Kurtz et al. (2013) that XOR learning is bimodal—most learners either fully master the categories or do not figure them out at all.

We identified four profiles of partial XOR generalization: 1) *Extrapolation-Based* generalization in which the one-quadrant category is extended to the untrained quadrant, 2) *50/50* generalization, in which the learner randomly classifies exemplars in the untrained quadrant, but has mastered the categories otherwise, 3) *Proximity-Based* generalization, in which the learner extends the proximal two-quadrant category to the untrained quadrant, and 4) *Non-learner* generalization, reflecting the random performance of a non-learner.

The profile distribution for each category type is displayed in Table 1. We observed a substantial number of non-learners in the full XOR condition. These learners are interesting in that they achieved a reasonably high level of accuracy (81%) in the last block of training, but their poor

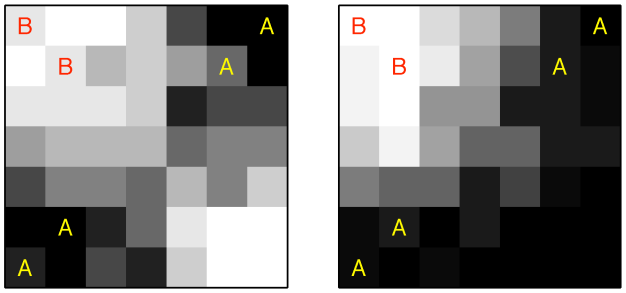

Figure 3. Aggregate generalization gradients for the extrapolation (left) and proximity (right) profiles in the partial XOR condition.

generalization performance suggests that a memorization strategy may have been used (particularly given the small size of the training set). An intriguing implication is that there exists a type of exemplar memorization that is effective during learning, but that fails to support systematic generalization. It is not clear how stimulus generalization would account for this pattern.

Nearly all of the successful learners in the partial XOR condition exhibited ether extrapolation-based (9/30) or proximity-based (19/30) generalization. This distribution is of great interest: the presence of extrapolation-based generalization suggests that classification responses are not always reached by comparing a presented cue to known reference points. In other words, these learners generalize with blatant disregard for proximity to exemplars. Aggregated generalization gradients for the extrapolation-based and proximity-based profiles are depicted in Figure 2.

Summary. To begin, we found that the partial XOR category structure was acquired more easily than full XOR. This result is interesting given previous work showing that linearly separable classifications are not always more quickly acquired that non-linearly separable ones (Medin & Schwanenflugel, 1981). In this experiment, the linearly separable partial XOR categories were learned more quickly.

Our result of primary interest was that many learners who were trained on the partial XOR categories generalized the one-quadrant category to the untrained quadrant. The presence of this extrapolation-based generalization shows that people do not universally make classification decisions based on distance to stored reference points—exemplars in the untrained quadrant are more proximal to members of the two-quadrant category. Since this is the central tenet of reference point theories of categorization, these results pose a significant challenge. In the following section, we formally evaluate DIVA and ALCOVE for their ability to fit these behavioral results.

Simulations

The overall goal of the following simulations is to evaluate the range of predictions made by DIVA and ALCOVE about generalization following training on the full and partial XOR categories. It is important to note that these models are conventionally applied to explain aggregated data. In this case, however, we are testing the models on their ability to match an individual differences distribution—we are interested in whether either model can explain the distribution of generalization profiles that was observed behaviorally (Table 1).

Procedure. For both models, we generated a large number of predictions using a wide range of parameter values. We searched over the predictions made by different parameter values using a 'grid-search' method where each model was initialized 30 times (corresponding to the number of participants in each condition) at each search point. We fit DIVA over four parameters: number of hidden units, learning rate, initial weight range, and a focusing parameter, β (Conaway & Kurtz, 2014). Likewise, we fit ALCOVE over its specificity constant, association learning rate, attention learning rate, and response mapping constant. Note that although we allowed both models to use attentional mechanisms, all dimensions are equally relevant in the categories we tested. By profiling the predictions made by each initialization, we can create a distribution of predicted generalization profiles that are linked to particular parameterizations. We are then able to assess the quality of each parameterization's predictions relative to our behavioral findings.

Our training and generalization procedure was identical to the one we used in our behavioral data. After training on full XOR, each initialization was profiled based on its match to the Learner and Non-Learner profiles. After training on partial XOR, we profiled each generalization gradient based on its match to the Extrapolation-Based, 50/50, Proximity-Based, and Non-Learner profiles.

Results. Both ALCOVE and DIVA provided a full account of the Full XOR data. In particular, both models were able to match the rate of non-learners identified at the generalization phase. Both models were able to do so under a wide range of parameterizations.

The models, however, diverged substantially when they were trained on the partial categories. Notably, and as predicted, we were unable to find any parameterization of ALCOVE that could produce extrapolation-based generalization. Instead, ALCOVE commonly produced proximity-based and 50/50 generalization gradients. ALCOVE's generalization was highly dependent on its parameter values: for example, the model was more likely to produce 50/50 gradients when the specificity value was large.

These results confirm our earlier simulations and theoretical analysis. ALCOVE can only produce classification responses based on distance to exemplar reference points, so the model will never be able to extend the one-quadrant category to exemplars that are more proximal to the two-quadrant category. Accordingly,

ALCOVE provides a sharply limited and unsatisfactory account of generalization in the partial XOR condition.

As expected, DIVA commonly predicted extrapolation-based gradients after training on the partial categories. The model was most likely to generalize from the one-quadrant category to the untrained area with a moderate to larger number of hidden units (3–10), a small to moderate learning rate (<= 0.5), and a large initial weight range (>= 1.5). No parameterization produced a strong majority of extrapolation-based gradients: in our simulations DIVA predicted a maximum of 16/30 extrapolation-based gradients. Nonetheless, DIVA's ability to produce this generalization profile sets the model apart from traditional reference point models that are limited to distance-based classification. Table 1 displays the best predictions made by each model (minimizing the mean-squared error, MSE, between the observed and predicted frequencies of each generalization profile).

Table 1. Observed and predicted generalization profile frequencies.

		Obs.	ALCOVE	DIVA
Full XOR	*Learner*	19	19	19
	Non-Learner	11	11	11
Partial XOR	***Extrapolation***	**9**	**0**	**8**
	50/50	1	0	0
	Proximity	19	23	17
	Non-Learner	1	7	5

Summary. Both models were able to accurately predict generalization after training on the full XOR categories. However, ALCOVE was unable to explain extrapolation-based generalization following training on the partial categories. DIVA provides a good account of the full distribution of generalization profiles observed behaviorally. This calls into question the reduction of human category learning to stimulus generalization that is inherent in reference point models.

Discussion

Theoretical work in human category learning has been closely tied to a reference point framework. Although reference point models differ from one another in a variety of ways, they share a common assumption: classification decisions are reached by comparing presented cues to one or more reference points. That is, all reference point models assume that classification decisions are based on *distance*.

In this paper, we reported an experiment that directly tested the distance assumption. Learners were given classification training on one of two versions of the exclusive-Or (XOR) categories. In the full XOR condition, the XOR categories were represented in a continuous, two dimensional stimulus space. In the partial XOR condition, one of the four quadrants was left untrained.

Most importantly, after classification training on the partial XOR categories, we found that a sizable proportion of learners (9/30) generalized according to the pattern of full XOR. That is, these learners extrapolated the one-quadrant category to novel exemplars that were actually more proximal, or similar, to members of the two-quadrant category. In doing so, these extrapolation-based learners violated the distance assumption, showing that classification is not always based on distance to known reference points.

We used formal simulations with ALCOVE (Kruschke, 1992) to show that traditional reference point models could not account for the presence of the extrapolation-based generalization profile in our data. Because these models are limited to classification based on distance to stored reference points, they are unable to extend the partial category into the untrained quadrant. As such, we expect this limitation to be shared by any standard reference point account including prototype models (Smith & Minda, 2000) and adaptive cluster models (Love, Medin, & Gureckis, 2004). The extrapolation-based generalization we observed is therefore a significant challenge to the assumptions underlying reference point models.

We contrasted ALCOVE's simulations with predictions made by DIVA (Kurtz, 2007), which is not a reference point model and is therefore not limited to classification based on distance. We observed that DIVA could produce extrapolation-based generalization after training on the partial categories, providing a full account of the generalization data.

Learning More Than Reference Points

Many of our partial XOR learners do not appear to have represented the categories using reference points in the input space (e.g., prototypes, exemplars, adaptive clusters). Further work will be needed to determine exactly what these individuals did learn, though their generalization responses indicate that, through training on the partial categories, they acquired a category representation that is consistent with the logical structure of XOR (i.e., white squares and black circles versus black squares and white circles). That is, our learners may have represented the partial categories using a rule, rather than similarity.

Given this interpretation, it is possible that rule-based models of category learning (i.e., RULEX; Nosofsky et al, 1994b; Nosofsky & Palmeri, 1998) can provide an account of the generalization we observed in the partial XOR condition. However, the simplest rules that characterize the two categories would not appear to offer a systematic basis for generalizing to the untrained quadrant. Hybrid or multiple-systems models that incorporate rule-based learning components (Ashby et al., 1998; Erickson & Kruschke, 1998) might be able to produce extrapolation-based generalization, but this would have to be despite the use of a category structure that requires information integration.

Alternatively, our simulation results with DIVA suggest that models may not need to represent rules explicitly in order to capture the patterns of generalization observed behaviorally. Although DIVA does not explicitly learn

rules, the model was able successfully to produce extrapolation-based generalization—it is therefore possible that the model is able to approximate rule knowledge through training on the partial XOR categories. Clearly, more work is needed to formally describe the knowledge that DIVA acquires through training on partial XOR.

Finally, it is important to note that the critical result arises in only a subset of the sample (9/30). However, we believe it is important for models to provide an explanation of the full set of commonly occurring profiles, and preferably provides a basis for explaining the variability. While reference point models can account for the majority result (proximity-based generalization), we found that DIVA can correctly predict the occurrence of both proximity and extrapolation-based generalization. These results are best explained outside of the reference point framework. In sum, by measuring generalization, removing the complicating factor of attention, avoiding aggregated outcomes, and putting a novel twist on an old favorite (XOR), we provide a clear demonstration of the need to look beyond similarity to reference points in explaining human categorization.

References

Ashby, F., Alfonso-Reese, L., Turken, A., & Waldron, E. (1998). A neuropsychological theory of multiple- systems in category learning. *Psychological Review*, 105, 442–481.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking*. New York: Wiley.

Conaway, N. B., & Kurtz, K. J. (2014). Now you know it, now you don't: Asking the right question about category knowledge. *Proceedings of the Thirty-Sixth Annual Conference of the Cognitive Science Society*. (pp. 2062-2067). Quebec City, Canada: Cognitive Science Society.

Erickson, M. A. & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General*, 127, 107-140.

Homa, D. (1984). On the nature of categories. In G. H. Bower (Ed.), *Psychology of learning and motivation* (Vol. 18, pp. 49-94). New York: Academic Press.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, 99(1), 22-44

Kruschke, J. K. (1993). Human category learning: Implications for backpropagation models. *Connection Science*, 5(1), 3-36.

Kurtz, K. J. (2007). The divergent autoencoder (DIVA) model of category learning. *Psychonomic Bulletin & Review*, 14, 560 –576.

Kurtz, K. J., K. R. Levering, R. D. Stanton, J. Romero, and S. N. Morris (2013). Human learning of elemental category structures: revising the classic result of Shepard, Hovland, and Jenkins (1961). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 39(2), 552-572.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, 111, 309-332.

Medin, D. L., & Schaffer, M. M. (1978). A context theory of classification learning. *Psychological Review*, 85, 207-238.

Medin, D. L., & Schwanenflugel, P. J. (1981). Linear separability in classification learning. Journal of Experimental Psychology: Human Learning and Memory, 7, 355-368.

Medin, D. L., Goldstone, R. L., and Gentner, D. (1993). Respects for similarity. *Psychological Review*, 100, 254-278.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. Journal of Experimental Psychology: General, 115(1), 39-57.

Nosofsky, R. M. (1988). Similarity, frequency, and category representations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14(1), 54-65.

Nosofsky, R. M. (1992). Exemplars, prototypes, and similarity rules. In A. F. Healy, S. M. Kosslyn, & R. Shiffrin (Eds.), *Essays in honor of William K. Estes* (pp. 149 –167). Hillsdale, NJ: Erlbaum.

Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., & McKinley, S. C. (1994a). Comparing models of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory & Cognition*, 22(3), 352-369.

Nosofsky, R. M., & Palmeri, T. J. (1998). A rule-plus-exception model for classifying objects in continuous-dimension spaces. *Psychonomic Bulletin & Review*, 5(3), 345-369.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994b). Rule-plus-exception model of classification learning. *Psychological Review*, 101(1), 53-79.

Rips, L. J. (1989). Similarity, typicality, and categorization. In S.Vosniadou & A.Ortony (Eds.), *Similarity and analogical reasoning*. Cambridge, England: Cambridge University Press.

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

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, 237, 1317-1323.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. Psychological Monographs: General and Applied, 75, 1–42.

Smith, J. D., & Minda, J. P. (2000). Thirty categorization results in search of a model. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, 26, 3-27.