Now you know it, now you don't: Asking the right question about category knowledge

Nolan Conaway (nconawa1@binghamton.edu)
Kenneth J. Kurtz (kkurtz@binghamton.edu)
Department of Psychology, Binghamton University
Binghamton, NY 13905 USA

Abstract

A key goal of category learning research is to describe how categories are represented. Essential to this research are measures that provide investigators insight into exactly what learners have gained from their training experience. In this paper, we review and explore three commonly used measures: A) ease of acquisition, B) generalization, and C) single feature classification. We report results of a category learning experiment in which these measures are compared side-by-side. We find that generalization and single feature classification data are the more informative measures; we also find a novel inconsistency between them. Specifically, many learners who generalize based on only a single dimension demonstrate robust knowledge of both dimensions during the single feature classification test. We discuss implications for methodology in the field, as well as for selective attention and theories of human category learning.

Keywords: generalization, categorization

Introduction

A fundamental assumption in theories and models of category learning is that a particular process leads learners to develop a core category representation that they can then put to use for subsequent categorization tasks. One of the major goals of category learning research is to describe this representation, as well as how it is acquired. Crucial to this goal are measures that provide insight into the learned representation. In this paper, we investigate three dependent measures that are commonly used in the field and explore the kind of information that is garnered from each.

Ease of acquisition measures (e.g., mean training accuracy, end-state accuracy, learning curves, blocks to criterion) are among the most frequently used measures in the field. In a classic study, Shepard, Hovland, & Jenkins (1961) analyzed the ease of acquisition of six elemental classifications based on eight stimuli varying in three binary dimensions. They observed that five out of the six elemental types were learned more quickly than would be predicted based on an identification study using the same stimuli (i.e., the mapping hypothesis; Nosofsky, 1984). Subsequent work in formal modeling provided an account of the ease of acquisition differences among the six types based on the application of selective attention to exemplar representations (Nosofsky et al., 1994).

Acquisition data, however, do not always specify what exactly learners have gained from the category learning experience. This point becomes especially apparent for classification problems that have multiple solutions. For example, the Type IV structure from Shepard et al. (1961) is often described as a linearly separable family-resemblance (Rosch & Mervis, 1975) structure whereby each category consists of a prototype and three examples that deviate by a single feature. Learners can therefore accurately classify all the training examples by comparing each to the prototypes. Importantly, Type IV mastery can also be achieved according to a rule-plus-exception solution (Nosofsky, Palmeri & McKinley, 1994) whereby each category is defined by a rule on a focal dimension along with a memorized exception to the rule. Based on the acquisition data alone, it is difficult to determine which of these solutions are actually learned. Even within the realm of similarity-based accounts, learning data alone often does not differentiate among models based on different types of reference points (prototypes, exemplars, clusters) or alternatives to reference points (e.g., DIVA, Kurtz, 2007)

Given this limitation, we need to extend our methods for probing the learned category representation. In classic work, Roger Shepard did much to establish the primacy of generalization in psychological research. Shepard put forward a universal law that describes stimulus generalization as an exponential function of distance in psychological space (Shepard, 1957, 1987). In Shepard's work, generalization was used to gain insight into how learners interpreted the dimensions of a stimulus space. Similarly, generalization after category learning offers investigators greater insight into what is gained during category learning experience (for a review of generalization, see Levering & Kurtz, 2010).

Generalization after category learning was an important testing ground in the clash between prototype and exemplar theories of categorization (e.g., Homa, 1984; Nosofsky, 1992). The 5-4 category structure (Medin & Schaffer, 1978) has been a particularly prominent structure in this regard. The 5-4 structure is based on nine training examples that vary in four binary dimensions and are divided into two categories. Crucial to studies of the 5-4 structure are seven novel examples that are classified after a set of training trials. By analyzing patterns of generalization to these items, investigators were able to describe what type of knowledge (e.g., prototypes, exemplars) was learned during the training (Johansen & Palmeri, 2002; Medin & Schaffer, 1978; Smith & Minda, 1998).

In a different approach to studying generalization, Erickson and Kruschke (1998, 2002) explored whether exemplar models could account for rule-like responding. In

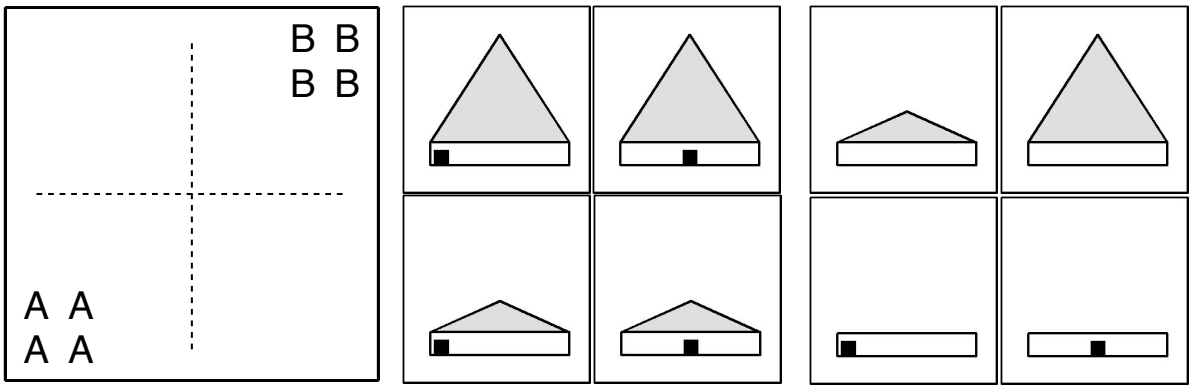


Figure 1: *Left*: A schematic of the 'minimal case' structure in which categories are separated across both stimulus dimensions. *Middle*: sample full-feature stimuli taken from the extremes of the space. *Right*: Sample partial-feature stimuli taken from the extremes of the space.

their study, learners received training on two novel categories based on a unidimensional rule with exceptions. After training, learners generalized their learning by classifying a large number of test items. Critical test items were perceptually similar to the exceptions, but also classifiable using the rule. Contrary to predictions from a pure exemplar account, learners generalized based on the rule (though see Nosofsky & Johansen, 2000 for a conflicting set of results). The authors suggested that multiple categorization systems were needed to account for the rule-based generalization that was observed – they developed ATRIUM, a hybrid of rule- and exemplar-based approaches to account for the results.

The information contained in gradients of generalization is obviously very rich and offers detailed insight into what has been learned, however there are many cases in which it is not possible to collect generalization data. The stimulus sets used in Shepard et al. (1961), for example, do not leave room for a generalization set. In situations where generalization is not possible or does not distinguish between hypotheses, yet another option is required.

Single feature classification tests offer a possibility in this regard. Such tests usually involve the presentation of a partial example (a single feature), and the learner is asked to indicate the category the feature is most likely to occur in. By testing individual features, these tests provide an opportunity to assess learners' knowledge in ways that would not be revealed in a full-item classification task (Levering, 2012).

Such tests have been used in the literature on learning via feature inference. Feature inference research has converged on the hypothesis that classification results in knowing the class assignments of exemplars, while inference learning results in knowledge about the prototypical values categories tend to have on each dimension (Yamauchi & Markman, 1998; Markman & Ross, 2003). Anderson, Ross, and Chin-Parker (2002) reported a test of this hypothesis in a study that had learners trained on a four-dimensional family resemblance structure. After the training phase, all learners were asked to complete a single feature classification and full feature classification test. The authors found that inference learners were more accurate on the single feature test, whereas classification learners were more accurate on the full feature test.

Despite their wide use in the literature, these three measures have not been compared side-by-side. In this paper, we seek to describe the type of information that can be gleaned from each measure and to explore the consistency of the conclusions that can be made from each measure.

The Current Study

In this experiment learners receive supervised classification training on a 'minimal case' category structure. The two categories are separated in opposite corners of a two dimensional space (see Figure 1). We selected this structure due to its pure simplicity – yet with the property that it is possible to learn to accurately classify all of the training items according to a variety of distinct strategies. For example, learners may focus on either one of the two stimulus dimensions, effectively forming a rule-based or unidimensional strategy. Alternatively, learners may integrate information across both dimensions (Ashby & Maddox, 1990), forming a diagonal boundary that separates the categories.

After a training phase, learners are asked to classify a generalization set of examples sampled from around the stimulus space. After generalization learners complete a single feature classification phase.

Method and Results

Participants and Materials. 75 undergraduates from Binghamton University participated in partial fulfillment of a course requirement. Stimuli were house-like figures that varied in the position of the lower box (door) and slope/height of the upper triangle (roof). Sample stimuli are shown in Figure 1. Participants were not informed that the

figures could be interpreted as houses (13.5% reported this interpretation of the stimuli in a post-experiment questionnaire). The stimuli were automatically generated at 8 positions along each dimension (8 door * 8 roof = 64 stimuli).

Procedure. Each participant completed 32 training trials (4 blocks consisting of the 8 training examples). After training, participants completed 64 generalization trials consisting of examples sampled at 8 positions on each dimension. Note that all 8 training items were presented during generalization. After generalization, each participant completed 16 single feature classification trials consisting of images containing only one stimulus dimension (door or roof), sampled at the same 8 positions on each dimension. Participants were informed that there would be test trials prior to beginning the experiment.

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 or single feature phase.

Results. 6 participants were excluded from further analysis due to A) experimenter error, or B) failing to meet at 7/8 accuracy criterion on the training items presented during the generalization phase.

From the generalization test phase, participant responses yield a generalization gradient. By comparing each gradient to three idealized gradients (rule-based generalization using exclusively the door or roof dimensions, and integration of both dimensions; Figure 3), we were able to profile each learner's generalization strategy. The majority of learners (58/69) generalized using a single dimension and a minority (11/69) generalized using both dimensions (diagonally). These results are consistent with our previous reports

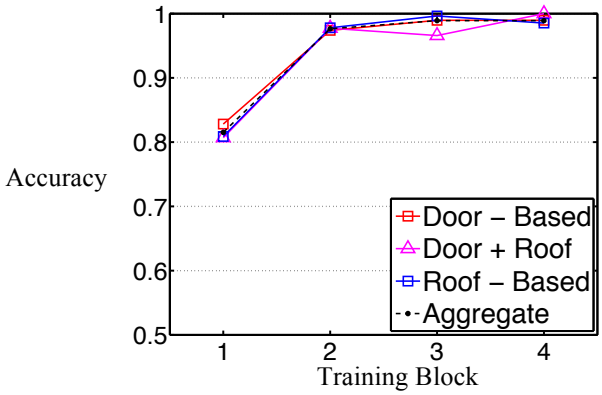
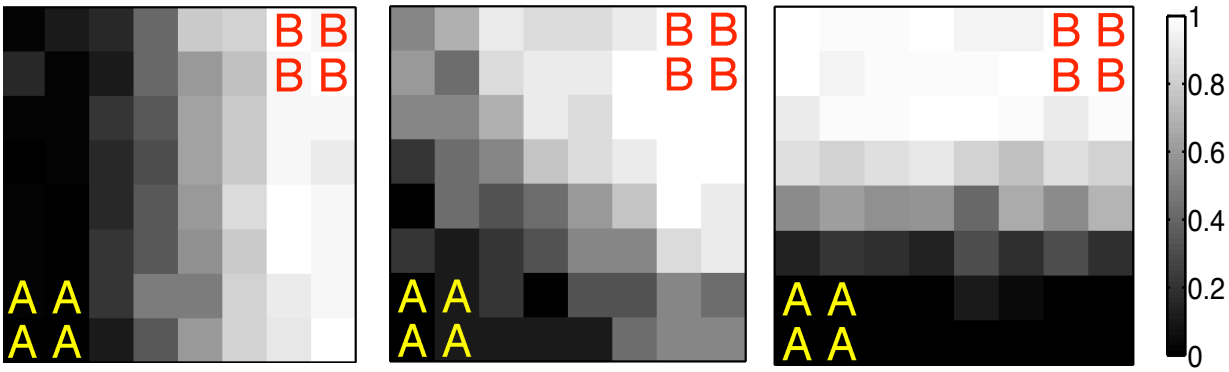


Figure 2: Training accuracy by block. Each generalization profile is shown separately.

(Conaway & Kurtz, 2013), as well as other reports of a unidimensional bias unsupervised classification (Ashby, Queller, & Berretty, 1999), supervised classification (Erikson & Kruschke, 1998, 2002), and free-sorting (Garner, 1974). Given this literature, it is interesting that a significant minority of our learners did not generalize unidimensionally and instead integrated the two dimensions for the purposes of their generalization responses.

The learners in each generalization group seem to have learned markedly different category representations. Accordingly, we analyze the training and single feature test data with generalization profile as a between-subjects variable. Generalization results are depicted in Figure 3.

The training data are depicted in Figure 2. The profile groups did not differ in the ease of acquisition of the categories: despite finding markedly different solutions to the category structure, all three groups mastered the categories at the same speed, $F(2, 66) = 0.193$, $p=0.826$. This result, however, must be qualified: all three solutions to the minimal case categories are relatively 'easy' to learn, thus we may be observing a ceiling effect in our data.



Door-Based (n=24/69) Door + Roof (n=11/69) Roof-Based (n=34/69)

Figure 3: Aggregate generalization gradients. Learners were grouped based on the profile type that best matched their gradient. The proportion of learners placed in each profile is specified below the gradient. Learners were more likely to select to the Roof dimension as their basis for generalization. *Left*: generalization based on the door dimension. Middle: generalization based on both dimensions. *Right*: generalization based on the roof dimension.

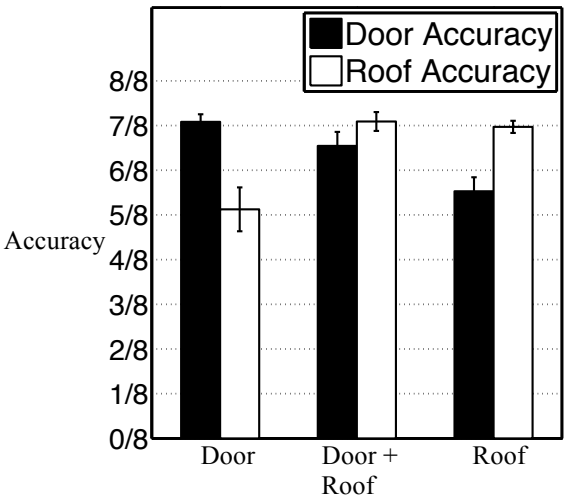


Figure 4: Single feature classification accuracy, broken down by generalization profile.

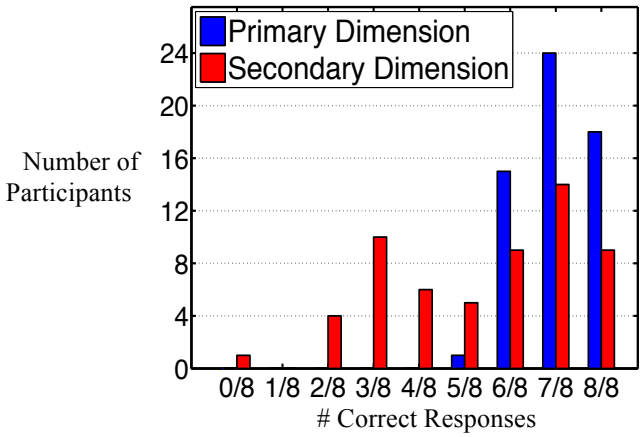


Figure 5: Single feature classification accuracy for unidimensional generalizers. 'Primary' and 'Secondary' refer to the dimension used (or not used) as the basis of generalization.

The profiles did, however, diverge during the single feature classification phase (Figure 4). Learners who generalized using a single dimension were more accurate on their selected (primary) dimension than their discarded (secondary) dimension, $t(57) = 5.7$, $p<0.001$. This results is consistent with the intuitive notion that unidimensional responding reflects or entails a lack of knowledge about one or more dimensions. That is, by selecting as the basis for generalization one dimension over another, learners forego the opportunity to learn about a secondary dimension.

Contrary to this notion, however, we found that unidimensional generalizers preformed above chance (> 4/8 correct) on the secondary dimension, $t(57) = 5.0$, $p<0.001$. This result suggests that these learners do in fact possess knowledge about the dimensions that are not being used as the basis of generalization, though it is less fully developed.

A histogram of the single feature classification data (Figure 5) shows that one subset of the unidimensional generalizers retained full knowledge of the secondary dimension, whereas a second subset did not. Indeed, 22.4% of unidimensional generalizers responded *less* accurately on single feature trials involving the feature they selected during generalization.

This result suggests that, for a notable subset of learners, the unidimensional basis for generalization reflects a strategy or convenience as opposed to a direct characterization of their category knowledge. These learners appear to possess a degree of category knowledge that remains latent during the generalization task, but becomes manifest in the single-feature classification task.

Discussion

Ease of acquisition, generalization, and single feature classification data are frequently used to probe the category representation that arises from a category learning experience. In our study, we compared these three measures directly. We trained human participants on a 'minimal case' category structure that has three qualitatively different solutions. After training, learners were asked to classify a large number of examples sampled around the stimulus space (yielding a generalization gradient of each learner's responses) and then asked to complete a single feature classification phase.

We found that most learners generalized their knowledge using only one of the two possible stimulus dimensions, though a notable subset generalized using both dimensions. We then analyzed the rest of our data with generalization profile as an independent variable.

Despite finding markedly different solutions to the minimal case structure, we found that these groups did not differ in learning speed: our ease of acquisition data do not provide insight into the learned solution. This result, however, could also be due to a ceiling effect.

We found that unidimensional generalizers responded more accurately to the dimension they selected during generalization. These result is consistent with the intuitive notion that rule-based or attentional responding implies that there is an absence of knowledge about one of the dimensions—that these learners generalized based on single dimension because they only learned about a single dimension.

Unexpectedly, further analysis of this data revealed a considerable number of unidimensional generalizers who retained full knowledge of the dimension that was not selected for generalization. Even as the generalization data would seem to reveal the essential nature of what was actually learned, we find that the single-feature classification results often revealed category knowledge that was simply not called upon for the purpose of generalizing.

Implications for Methodology

One of the central goals of research in category learning is to understand the nature of learned category representations. According to our data, conclusions that can be made about

category representations depend on the type of probe used to access the representation. This idea is most dramatically embodied by our novel result that, for many learners, unidimensional generalization is strategic and does not reflect the full extent of the learned representation. Based on the generalization data alone, we may have assumed that unidimensional generalization is the consequence of a lack of knowledge about one of the dimensions, but our single feature classification data suggest that this is not always the case. Thus, our results demonstrate that it is crucial to use a rich set of test measures that will most fully describe the type of category knowledge that has been learned.

Generalization has long been held in high regard due to its ability to describe what learners have gained through a learning experience (Shepard, 1957, 1987). Indeed, it often assumed that gradients of generalization provide a comprehensive depiction of learning. Our data suggests that this may not always be the case—many of our learners engaged only a subset of their knowledge for the purposes generalizing. More research, however, is needed in order to determine the mechanisms underlying this result.

Implications of Theories of Category Learning

One of our core findings is that many learners who generalized using a single dimension were also able to respond with high accuracy on both dimensions during the single feature test. That is, the rule-based or attentional responding we observed during generalization did not always indicate that there was a lack of knowledge about one of the dimensions. This novel result may have broad implications for theories and models of category learning.

Our results can be interpreted in terms of a generative/discriminative methods distinction made in the machine learning literature (Ng & Jordan, 2001). Discriminative methods are used to efficiently learn task-specific representations that can discriminate among categories. Generative methods allow models to learn multifaceted representations that describe each category as fully as possible. Whereas a discriminative account of category learning would predict a static application of category knowledge, many of our learners seemed to apply their category knowledge differently in response to new task demands.

Contemporary reference point theories (such as the prototype and exemplar views) do not immediately provide an explanation of this finding. Many reference point models (e.g., Kruschke, 1992; Minda & Smith, 2002) assume that selective attention is used to optimize a subsequent similarity calculation—that attention is used to increase the similarity of reference points within each category and decrease similarity between categories. To many of these theories, attention is regarded as a fundamental aspect of category acquisition (e.g., Kruschke, 2005)—attending to diagnostic dimensions allows observers to optimize the scope of their learning, thereby expediting the learning process. Reference point theories would therefore predict that these learners acquire very little (or no) knowledge about how the categories vary on the dimension they discarded for the purposes of generalization. In order to explain our findings, the manner in which attention is formulated would thus have to be altered substantially.

DIVA (Kurtz, 2007) offers a somewhat different approach to modeling category learning. Rather than prototypes or exemplars, DIVA represents categories as coordinated statistical models using dedicated output channels in a *DIV*ergent *A*utoencoder. An updated version of the model utilizes a late-focusing mechanism that allows the model to more heavily weigh dimensions that are disparate across categories. Crucially, DIVA's focusing weights are not learned, but generated dynamically on each trial, according to the following formula:

$$W_i = \exp(\beta * (|A_i - B_i| - k)) \quad (1)$$

Where A and B represent output channels dedicated to the categories, $_i$ indexes the stimulus dimension, β is a free parameter ($0 \leq \beta \leq \infty$) that determines the degree of focusing, and k is a constant set to (max-min) over dimension values.

DIVA's focusing mechanism serves to mediate between the generative knowledge learned by the divergent autoencoder and the classification decisions that are being made at a later stage. DIVA therefore suggests that learners gain a full representation of the categories, and that focusing is applied for the purposes of making decisions. Given this framing, DIVA provides an additional interpretation of our results—it is possible that our learners gain a full, non-attentional representation of the categories during the training phase. After learning, this representation can be applied in different ways depending on the constraints placed on learners by the task.

A somewhat more radical account of the present findings could be ventured: perhaps even the most simple category representations are not the stable entities we presume them to be. Such an account of the data suggests that learners differentially access or apply their categories depending on the way they are asked to use them. This account dovetails nicely with the literature on category learning by different modes (e.g., Yamauchi & Markman, 1998; Markman & Ross, 2003). The common theme is that the nature of the category knowledge is very much dependent on the manner in which the categories are used: either during learning or at test. More research is clearly needed to elucidate this theoretical interpretation.

Acknowledgments

We would like to thank the members of the Learning and Representation in Cognition (LaRC) Laboratory at Binghamton University.

References

Anderson, A. L., Ross, B. H., & Chin-Parker, S. (2002). A further investigation of category learning by inference. *Memory & Cognition*, 30, 119–128.

Ashby, F. G., & Maddox, W. T. (1990). Integrating information from separable psychological dimensions. *Journal of Experimental Psychology: Human Perception and Performance*, 16, 598–612.

Ashby, F., Queller, S., & Berretty, P. M. (1999). On the dominance of unidimensional rules in unsupervised categorization. *Perception & Psychophysics*, 61, 1178–1199.

Conaway, N. B., & Kurtz, K. J. (2013). Models of human category learning: Do they generalize? Proceedings of the Thirty-Fifth Annual Conference of the Cognitive Science Society. (pp. 2088-2093). Berlin, Germany: Cognitive Science Society.

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

Erickson, M. A. & Kruschke, J. K. (2002). Rule-based extrapolation in perceptual categorization. *Psychonomic Bulletin & Review*, 9(1), 160-168.

Garner, W. R. (1974). *The processing of information and structure.* Potomac, MD: Erlbaum.

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.

Johansen, M. A., & Palmeri, T. J. (2002). Are there representational shifts during category learning? *Cognitive Psychology*, 45, 482–553.

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

Kruschke, J. K. (2005). Learning involves attention. In: G. Houghton (Ed.), *Connectionist Models in Cognitive Psychology*, Ch. 4, pp. 113-140. Hove, East Sussex, UK. Psychology Press.

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

Levering, K. (2012). *Generative processing as a framework for human category learning*. (Order No. 3522557, State University of New York at Binghamton). *ProQuest Dissertations and Theses*, 160. Retrieved from http://search.proquest.com/docview/1040706252?accountid=14168. (1040706252).

Levering, K., & Kurtz, K. J. (2010). Generalization in higher-order cognition: Categorization and analogy as bridges to stored knowledge. In M. T. Banich & D. Caccamise (Eds.), Generalization of knowledge: Multidisciplinary perspectives(pp. 175–196). New York, NY: Psychology Press

Markman, A. B., & Ross, B. H. (2003). Category use and category learning. *Psychological Bulletin*, 129, 592–613.

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

Minda, J. P., & Smith, J. D. (2002). Comparing prototype-based and exemplar-based accounts of category learning and attentional allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28, 275–292.

Ng A.Y. & Jordan M. On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems,* volume 14, pages 841–848, Cambridge, MA, 2001. MIT Press.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 10, 104–114. doi:10.1037/0278-7393.10.1.104

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

Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., McKinley, S. C., & Glauthier, P. (1994). Comparing models of rule-based classification learning: A replication and extension of Shepard, Hovland, and Jenkins (1961). *Memory & Cognition*, 22, 352–369. doi:10.3758/BF03200862

Nosofsky, R. M. & Johansen, M. K. (2000). Exemplar-based accounts of "multiple-system" phenomena in perceptual categorization. *Psychonomic Bulletin & Review*, 7, 375-402.

Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, 101, 53–79. doi:10.1037/0033-295X.101.1.53

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. Cognitive Psychology, 7, 573–605. doi:10.1016/0010-0285(75)90024-9

Shepard, R. N. (1957). Stimulus & response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika*, 22, 325-345.

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. doi:10.1037/h0093825 and Applied, 75, 1–42. doi:10.1037/h0093825

Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24, 1411–1436. doi:10.1037/0278-7393.24.6.1411

Yamauchi, T., & Markman, A. B. (1998). Category learning by inference and classification. *Journal of Memory and Language*, 39, 124–149.