Observational Category Learning as a Path to More Robust Generative Knowledge

Kimery R. Levering (kleveri1@binghamton.edu)
Kenneth J. Kurtz (kkurtz@binghamton.edu)
Department of Psychology, Binghamton University
Binghamton, NY 13902-6000 USA

Abstract

Models and theories of category learning may exaggerate the extent to which people adopt discriminative strategies because of a reliance on the traditional supervised classification task. In the present experiment, this task is contrasted with supervised observational learning as a way of exploring differences between discriminative and generative learning. Categories were defined by a simple unidimensional rule with a second dimension that was either less diagnostic (than the simple rule on the first dimension) or non-diagnostic. When the second dimension was less diagnostic, observational learners were more sensitive to its distributional properties compared to classification learners (though classification accuracy at test did not differ). Observational learners were also consistently more sensitive to distributional information about the highly diagnostic dimension. When the second dimension was non-diagnostic, neither learning group showed sensitivity to the distributional properties of this dimension.

Keywords: concepts and categories; category learning; supervised classification task; representation;

Introduction

Representations of concepts and category knowledge must be robust enough to be applied across a broad range of psychological tasks. Some of the most important functions of such knowledge include being able to make inferences about unknown properties, satisfy goals, communicate ideas, and generate new instances. In short, we need to have comprehensive and flexible mental models of information corresponding to categories in the real world (Markman & Ross, 2003; Solomon, Medin, & Lynch, 2003).

Some well-known models of human category learning (e.g. ALCOVE, RULEX) do not address the robustness and flexibility of real psychological concepts. Instead, they focus on the time course of learning to assign items to one category or another. This neglects the learner's ability to use the resulting representation to do anything else (though see Kurtz, 2007; Love, Medin, & Gureckis, 2004).

Theoretical claims about the mechanisms behind category learning can be similarly focused on the core task of artificial classification learning. One example of this is the prevalence of selective attention as the proposed primary mechanism by which category learning occurs. In light of evidence such as Shepard, Hovland, & Jenkins' (1961) classic study, researchers have tended toward the view that people learn by allocating attention (either initially or gradually) exclusively to dimensions that are diagnostic of category membership (Nosofsky, 1984). This type of explanation has been successful in explaining classification learning as studied in the lab, but it may not go far enough to account for the flexibility with which categories are used in the real world. One example of this is the ability to perform tasks requiring knowledge that was not needed while learning to distinguish between classes.

Generative and Discriminative Methods

A useful framework for understanding this shortcoming in leading accounts of category learning can be found in the distinction between *generative* and *discriminative* classifiers in the machine learning literature (Ng & Jordan, 2001). Formally, a discriminative model (e.g., linear regression) learns to classify examples by optimizing a function to partition the space and correctly segregate category members. A *generative* model (e.g., naïve Bayes classifier) learns the distributional properties within each category. It classifies by determining the likelihood of each category generating the given input. Both types of models can learn to correctly categorize, but the generative model does so by modeling more of the data than the task requires per se.

A generative framework may be more appropriate to apply to the psychology of human categorization. According to the dichotomy, purely discriminative category learning would result in good classification performance, but not much knowledge beyond the ability to distinguish the categories. Purely generative category learning would result in complete statistical models of each category that are sensitive not only to diagnostic properties, but also to the internal structure and distributional characteristics of its features. As suggested earlier, the latter may be closer to the type of category learning that occurs in the real world.

The nature of laboratory materials and tasks may induce discriminative and generative learning to different extents. However, the task that is most commonly used, modeled, and accounted for in category learning experiments – the traditional supervised classification task – may foster learning that is especially discriminative. In a traditional supervised classification task, a series of objects or feature sets are displayed one at a time and learners respond by choosing one of the (typically two) mutually exclusive categories. Corrective feedback is given on each trial and participants learn to assign each member to the appropriate category.

In addition to other factors (binary-valued stimulus dimensions, mutually exclusive categories, small number of categories and examples), this task itself invites the learner to engage in discriminative learning behavior. Guess-and-correct learning about category membership supports hypothesis testing that is specific to distinguishing between categories and may encourage focusing on individual

features that define membership rather than information about the internal structure of the categories. As a consequence of the field's reliance on this task as a primary source of data, our explanatory accounts may be overly focused on an idiosyncratic mode of learning.

By comparing classification to other learning modes that may be less focused on distinguishing between categories, we may be able to identify shortcomings and spur development of better models and theories. For example, research on inference learning (learning by predicting missing features) shows this learning mode to be more in line with the development of generative knowledge. In contrast to classification, inference learning has been shown to make learners more aware of non-diagnostic, prototypical features (Yamauchi & Markman, 1998) and correlations between features not necessary for classification (Chin-parker & Ross, 2002). Of particular relevance to the current experiment, Hoffman and Rehder (2010) found that inference learners were better able to adapt and appropriately attend to a novel classification contrast in which a previously irrelevant dimension was suddenly relevant.

Observational Task to Induce Generative Learning

In this experiment, we employ a supervised observational category learning task in an attempt to induce more generative learning. In an observational task, the learner does not explicitly generate a guess about category membership. Instead, the example and category label are presented simultaneously (or the category label is presented just before the example). Cases of this type of learning are readily available in the real world. For example, imagine a child and parent walking by a lake when the parent says "Look! A duck." In cases like this, the focus of learning may be taken off distinguishing between categories. Instead, learning may involve trying to understand the properties of the category itself. The child who is shown a duck is likely to develop knowledge about what it means to be a duck, or what features are associated with being a duck, but probably not explicitly what distinguishes a duck form some other animal.

There has been little research comparing the traditional supervised classification task with observational category learning. Most of this research has evaluated learning based on classification accuracy and has not investigated more subtle ways that resulting representations may differ (i.e. differences in within-category knowledge). For example, using two-dimensional continuous stimuli, Ashby, Maddox, and Bohil (2002) found no difference in learning accuracy between the two types of tasks when categories were defined by a simple unidimensional rule. Although this suggests (as they took it to mean) that there is no qualitative difference between learning rule-based categories through classification or observational learning, the design was not sensitive to possible differences in representations of internal structure.

There is one study (that we know of) in which the representation of internal structure is compared between the two tasks. Hsu and Griffiths (2010) found that for a simple rule-based category structure, observational learners were more likely to be sensitive to the variability of examples within the category. They conducted an experiment in which participants learned to distinguish between two categories of lines varying in length. Category membership was based on a simple one-dimensional rule (short lines in one category, long lines in the other category). The categories were designed so that the variability of one category was greater than that of the other category. At test, participants were asked to classify lines of intermediate length. They found that observational learners were more likely than classification learners to place intermediate examples into the category with higher variability. This suggests that observational learners may possess greater sensitivity along relevant, diagnostic dimensions. This is consistent with our intuition that observational learning may have a more generative basis than classification.

Current Experiment

In this experiment we investigate differences between learning categories via classification and observation. We are particularly interested to see if these differences are consistent with the distinction between generative and discriminative methods of learning. We use two-dimensional stimuli with categories defined by a simple unidimensional rule along one of the dimensions. The other dimension is distributed bimodally with the center of the distribution vacant. Typicality ratings and an adapted inference task after learning are used to evaluate acquired representations of internal structure.

The general premise of our predictions is that, without a decrement in classification performance, observational learners will be sensitive to information beyond that needed to classify items according to the unidimensional rule. We expect that this sensitivity will include distributional information along the most diagnostic dimension and also along the other dimension.

Increased sensitivity could arise either because: 1) observational learners are able to look beyond the most obvious unidimensional rule to consider other differences between the categories, or 2) because they are attending to the distributions of the individual categories with reduced regard for distinguishing the categories. To address the cause of a possible advantage, we manipulated the degree of diagnosticity of the second dimension.

Method

Participants

200 Binghamton University undergraduates participated in this experiment for partial fulfillment of course credit. Participants were randomly assigned to one of four conditions based on the learning task (*Classification* or

Observational), and diagnosticity of the second dimension (*Less-Diagnostic* or *Non-Diagnostic*).

Materials and Design

Stimuli. The stimulus materials were adapted from a "creature" produced with the SPORE Creator tool from the videogame SPORE (2001). Figure 1 shows a sample creature. The creatures varied along two dimensions: the number of "spike" protrusions from the sides of the body, and the ratio of gray dots (always above) to black dots (always below). The number of spikes on the right and left side could range from 3 to 14. There were always 150 total dots, but creatures ranged from having only 20 gray dots (and therefore 130 black dots) to having 130 gray dots (and therefore 20 black dots) in intervals of ten. Hence, there were 12 possible values along each of the two dimensions, creating a possible stimulus space of 144 items representing every combination of feature (see Figure 2 for the entire range of possible feature combinations).

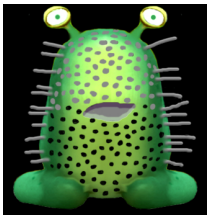

Figure 1: Example of stimuli

Twenty-four creatures (12 for each category) taken from the center of the stimulus space were used for training. The rest of the stimulus space was reserved for exploring how category representations generalized. At test, four examples within each category and ten additional items for each category were used to collect typicality ratings.

Category Structure. Figure 2 depicts the feature space, specific examples, frequency of examples, and the category assignment of examples used in the experiment. As can be seen in the figure, the assignment of stimuli to categories was determined based on a simple unidimensional rule (examples to the right of the solid black line were in one category, examples to the left were in the other). The dimension to which the rule applied was counterbalanced. We will refer to the dimension in which the simple unidimensional rule determined category membership as the HI relevance dimension, and the second, less diagnostic dimension as the LO relevance dimension.

The degree to which the LO relevance dimension was diagnostic differed between the conditions (see Figure 2). In the *Less-Diagnostic* condition, one category consisted of creatures with extreme values along the LO relevance dimension, while the other category was made up of creatures possessing central values along the LO relevance dimension. In this condition, the LO relevance dimension could be used to predict category membership, although the rule for membership was more complex than the unidimensional rule along the X-axis (e.g., one category has less than 50 gray dots or greater than 100 dots, while the other has between 50 and 100 dots). In the *Non-Diagnostic* condition, both of the categories consisted of creatures possessing extreme values along the Y- dimension. Because both categories possessed the same structure along this dimension, learners could not use it to distinguish between categories. It is important to note that the same category used in the *Less-Diagnostic* condition was also used in the *Non-Diagnostic* conditions (the "common category" on the left in Figure 2). This allows us to compare representations of the same category under conditions in which one dimension is either diagnostic or not diagnostic.

Although there were only 12 unique creatures for each category in the training set, some items were presented more than once in each block. Therefore, each block of training included 24 presentations for each category. The frequency of presentation of each item reinforced the robustness of each category's internal structure and provided another avenue for learners to learn non-diagnostic information about the categories. The most frequent items (displayed 4 times every block) were always at the center of the distribution of examples along the HI relevance dimension. Along the LO relevance dimension, the most frequently occurring items were at the center of only one category distribution in the *Less-Diagnostic* conditions. All other categories had bimodal distributions along the LO relevance dimension and the most frequently occurring items were at the extremes.

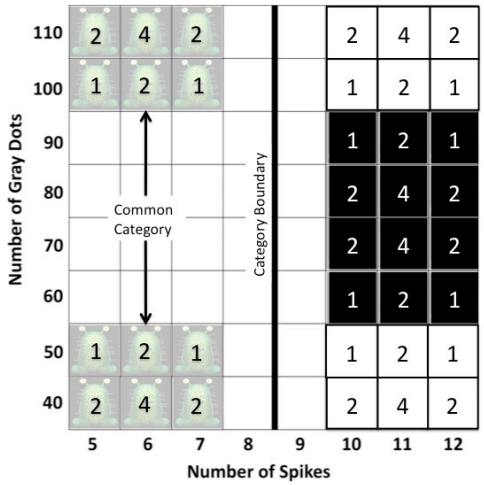

Figure 2: Feature space, examples, and frequency of each example in the training set for the *Less-Diagnostic* (black squares) and *Non-Diagnostic* (white squares) conditions. All conditions learned the common category.

Procedure

The instructions given to classification and observational learners were slightly different in conjunction with the induction of a discriminative or generative approach. All participants were told that a new planet was recently discovered and that it was their job to learn about the two types of creatures living there. In the observational condition, participants were told that researchers traveled

around the planet taking a picture of each creature they found and that they would be learning from this catalog of pictures. This was intended to give the impression that the examples were drawn from a population and their properties directly reflected the distribution of properties of creatures on the planet. Classification learners were merely told that they would be seeing pictures of creatures, but no mention was made of the pictures actually representing the population of creatures.

Learning Phase. In classification trials, items were presented one at a time randomly on the screen along with buttons labeled for creatures of type "Yugli" and "Zifer." On each trial, participants were asked to choose via mouse click the category to which the example belonged. After making a guess, they received feedback indicating whether they were right or wrong and were shown the correct answer. The feedback remained on the screen until they clicked to proceed to the next trial.

Before a stimulus was displayed in observational trials, learners were presented with the correct category label for 1500 ms. The image and label were then shown together on the screen for another 1500 ms – at which point the learner clicked to confirm the correct category name and continue to the next trial.

Both classification and observational learners completed two blocks consisting of all 48 items. After the first block, all learners were given a brief *endorsement* task asking them to classify 8 selected items. In this task, each item was presented on the screen with a correct or incorrect label. Participants were asked to indicate whether they agreed or disagreed with the classification of the item. The purpose of this test phase was to determine the progress of learners. This was critical for the observational learning condition in which no learning accuracy data could be recorded. In an effort to match the number of trials between the observational and classification conditions, there was no criterion level of performance that would move learners on to the test phase. Instead, all classification and observational learners completed a total of 96 trials before progressing to the test phase.

Test Phase. The test phase consisted of a series of three types of test trials. The first type of test trial was an endorsement task (as above) designed to assess category knowledge of the 24 trained examples. Each item was presented once with a corresponding label that was either correct or incorrect (randomly determined). This task was used instead of a traditional classification test phase to ensure that the test task did not match the task of either learning group.

While the first test phase was included to compare classification accuracy, the next two test phases were included to evaluate whether there was a difference in sensitivity to the distribution along both dimensions depending on the type of learning task. The second test phase was designed to assess representations of internal structure via typicality ratings. Learners were presented with a series of 24 creatures (shown in Figure 3). They were given the correct category label and asked to rate how typical or 'good an example' each item was of its category. Category membership of each item was provided to try to ensure that typicality ratings were not based on confidence in category membership. Only four of these items had been presented during training. The other 20 items had never been seen. Despite never having been seen, some items were closer to the distribution of items in their category than others. Along the HI relevance dimension, sensitivity to internal structure would manifest in typicality ratings reflecting the range of values presented during training. In other words, the items with feature values along the HI relevance dimension that were presented many times during training (those items labeled with an "a" in Figure 3) would be rated higher than those items outside of the distribution of presented feature values (those items labeled with a "b" in Figure 3). In the category common to all conditions, items in the center of the distribution were not sampled during training. If learners were sensitive to this, they would presumably rate the typicality of examples in this region (those items labeled with circles in Figure 3) lower than examples in the extremes of the distribution along that dimension (those items labeled with squares in Figure 3).

The last test phase was an inference task that more explicitly assessed the learner's knowledge of the distribution of features along the LO relevance dimension. Participants were given a category label and images representing two possible feature values along one dimension. They were asked which value (out of the two provided) was more likely given the category information. In addition to the category labels, they could answer "equally likely" if they believed that there was no greater probability of one category over another. To ensure that responses were made because they believed them to be accurate, not because they were forced to choose, they could also answer "not sure."

In the inference task, three key single-feature distinctions for each category were of interest along the LO relevance dimension. Of these three, two asked learners to decide between an extreme value and a central value (e.g. 70 gray dots vs. 110 gray dots). In the common category in which the distribution was bimodal, the correct answer was always that the extreme value was more likely. The third trial asked them to decide which of two extreme values was more likely (the correct answer being "equally likely"). We averaged accuracy across these three judgments for the common category to determine a score for each participant based on their sensitivity to the distribution along the LO relevance dimension.

Predictions. We believed that learners in the observational conditions would display greater sensitivity to internal structure along both dimensions. More specifically, we expected typicality ratings and selection of single feature values to reflect a greater sensitivity to the missing central values of the category that was common between all conditions.

If the observational learners were at an advantage because they were developing a representation of the features within each category separately, their sensitivity to the internal structure of the Y-distribution should not change depending on the distribution of the contrast category. If, however, the advantage of the observational learner was due to an increased ability to develop a more comprehensive combination of diagnostic information, then their sensitivity to the internal structure should depend on how diagnostic that information is. We can directly test this because we have the same exact category under a condition in which the contrast category is distributed the same along the LO relevance dimension (*Non-Diagnostic)* and a condition in which it is not (*Less-Diagnostic).*

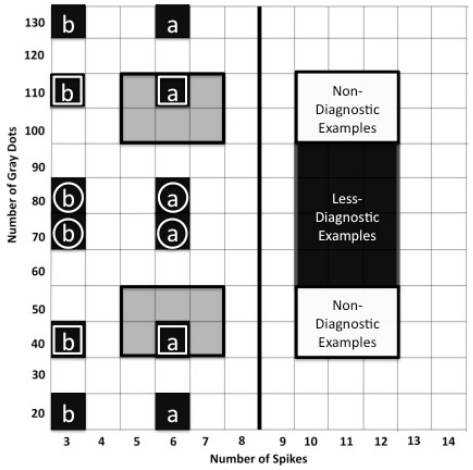

Figure 3. Items presented in typicality phase are labeled "a" and "b". Training items shaded gray for reference.

Results and Discussion

Classification of Trained Examples

To begin, we were interested in determining if classification accuracy was affected by whether the learning mode was classification or observational learning. As in Ashby, Maddox, and Bohil (2002), this was not found. For the endorsement task of trained items, performance of the classification learners (M = .98, SD = .05) and observational learners (M = .97, SD = .08) were not significantly different for the endorsement test after two blocks (*p*s>.25). This was also true after just one block. Therefore, there was no apparent advantage for classification learning despite the fact that they were queried on each trial.

Our next goal was to see whether observational learning resulted in greater sensitivity to distributions of features along both dimensions. Subsequent analyses focus on the one category that was common between conditions.

Distribution Along HI relevance dimension

Through typicality ratings, we confirmed our predictions that observational learners would be more sensitive to the range of values along the HI relevance dimension presented during learning.

Typicality ratings. Sensitivity to the range of values presented in the training set would result in typicality ratings that decreased as distance from the training examples increased. Typicality ratings were collected for 6 examples in the common category at the center of the distribution along the X-axis (indicated by the letter "a" in Figure 3) and 6 examples that were beyond the distribution along the X-axis (indicated by the letter "b" in Figure 3). The extent to which the average typicality ratings for items further away from the presented values were lower than those at the center of the distribution determined each participant's sensitivity along this dimension.

A 2 (task) x 2 (diagnosticity) x 2 (example type) ANOVA revealed a main effect of example type, $F(1, 196) = 12.501$, $p = .001$, $\eta^2 = .060$. Overall, center items were rated more typical ($M = 5.336$, $SD = 1.191$) than extreme examples ($M = 4.77$, $SD = 1.825$). An interaction between task and example type drove this main effect, $F(1, 192) = 12.281$, $p = .001$, $\eta^2 = .056$. For classification learners, there was no difference between their ratings of the central items and their ratings of the extreme examples, $t < 1$. However, observational learners rated the central items significantly higher than the extreme examples, $t(99) = 4.740$, $p < .001$, indicating that they were sensitive to the range of the distribution along this dimension.

Taken together, these results are consistent with Hsu and Griffiths (2010) in that people in a classification task try to find a boundary between categories along diagnostic dimensions, while observational learners develop a representation of the diagnostic dimension that includes distributional information.

Distribution Along LO relevance dimension

As predicted, inference judgments and typicality ratings of observational learners reflected greater knowledge of the distributional gap in the LO relevance dimension. This effect depended on whether or not the second dimension was diagnostic.

Inference test. We averaged together the accuracy for the three key questions asked of the LO relevance dimension within the common category (see Figure 4). A 2 (task) x 2 (diagnosticity) ANOVA revealed a significant main effect of task, $F(1, 196) = 8.566$, $p = 0.004$, $\eta^2 = .042$. Observational learners were more accurate ($M = .403$, $SD = .265$) than classification learners ($M = .300$, $SD = .244$) at determining which feature value was correct. This was despite the fact that classification and observational learners did not differ in their likelihood of selecting "not sure".

There was also a main effect of diagnosticity, $F(1, 196) = 7.497$, $p = .007$, $\eta^2 = .037$, indicating that the degree of diagnosticity to some extent determines sensitivity to distributional properties. More specifically, people in the *Less-Diagnostic* conditions were significantly more accurate ($M = .400$, $SD = .321$) than those in the *Non-Diagnostic* conditions ($M = .303$, $SD = .165$).

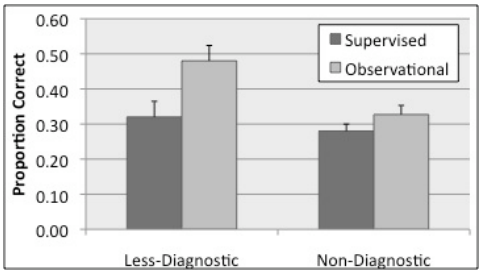

Figure 4. Proportion correct on Y-dim inference items

Typicality ratings Along the LO relevance dimension, we averaged the typicality of the four examples at the center of the distribution in each category (indicated by circles in Figure 3) and the four examples at the bimodal location in each category (indicated by squares in Figure 3). Figure 5 shows the average difference between these measures for the four conditions. We conducted a 2 (task) x 2 (diagnosticity) x 2 (example type) and found a main effect of example type, F (1, 196) = 14.403, $p < .001$, $\eta^2 = .068$. Overall, the items at the bimodal location were considered more typical (M = 5.219, SD = 1.159) than those at the central location (M = 4.835, SD = 1.397). However, we did not find a significant main effect of task or an interaction between task type and example type. In sum, classification learners were just as likely as observational learners to rate central items lower than the bimodal items.

We did find an interaction between diagnosticity and example type, F(1, 196) = 13.662, $p < .001$. Learners in the less diagnostic conditions rated the center items lower than the bimodal items, t(99) = 3.850, $p < .001$. However, in the non-diagnostic condition, there was no significant difference between ratings of the center and bimodal items, $t < 1$.

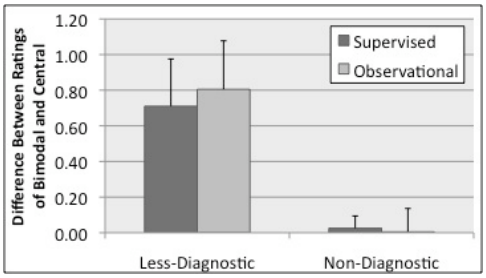

Figure 5. Average difference between typicality of central and extreme items along Y-dim.

Summary of Findings

Typicality ratings and inference judgments indicate that: 1) observational learners, but not classification learners, were sensitive to the range of distribution along the most diagnostic dimension; 2) when a second dimension was also somewhat diagnostic, observational learners were sensitive to its distribution despite the fact that it was not necessary for distinguishing between categories; 3) when a dimension was not diagnostic, neither classification nor observational learners were sensitive to its distributional properties. The advantages of observational learners were present without a corresponding difference in classification performance.

These data are consistent with observational learners being generative learners who are sensitive to information beyond that which is required for distinguishing between classes. However, there is conflicting evidence about the extent to which classification learners are sensitive to this information. While performance on the inference tasks indicates a decrement compared to observational learners, typicality ratings imply that classification learners are aware on some level of the distribution along the second dimension. It may be that typicality ratings are a more sensitive measure, or that classification learners were aided by the presentation of whole exemplars rather than individual features.

Although unexpected, we find it important to emphasize that neither observational nor classification learners were sensitive to the distribution along the second dimension when it was non-diagnostic. This suggests that distributional information is not as strongly represented for features that do not differentiate between categories. Further work will investigate more specifically the types of information that are represented through discriminative and generative learning modes.

References

Ashby, F. G., Maddox, W. T., & Bohil, C. J. (2002). Observational versus feedback training in rule-based and information-integration category learning. *Memory & Cognition, 30*, 666-677.

Chin-Parker, S. & Ross, B. H. (2002). The effect of category learning on sensitivity to within-category correlations. *Memory & Cognition, 30*, 353-362.

Hoffman, A. B. & Rehder, B. (2010). The costs of supervised classification: The effect of learning task on conceptual flexibility. *Journal of Experimental Psychology: General, 139*, 319 – 340.

Hsu, A. & Griffiths, T. L. (2010). Effects of generative and discriminative learning on use of category variability. In R. Camtrabone & S. Ohlsson (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society,* 242-247.

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

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

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

Ng, A. I., & Jordan, M. I. (2002). On discriminative vs. generative classifiers: A comparison of logistical regression and naïve Bayes. *Advances in Neural Information Processing Systems 17*. Cambridge, MA: MIT Press.

Shepard, R. N., Hovland, C. L, & Jenkins, H. M. (1961). Learning and memorization of classifications. Psychological Monographs, 75, 1-42.

Solomon, K. O., Medin, D. L., & Lynch, E. (1999). Concepts do more than categorize. *Trends in Cognitive Science, 3*, 99-105.

Yamauchi, T. & Markman, A. B. (1998). Category learning by inference and classification. *Journal of Memory and Language, 39*, 124-148.