The Influence of Learning to Distinguish Categories on Graded Structure

Kimery Levering (kleveri1@binghamton.edu)
Kenneth J. Kurtz (kkurtz@binghamton.edu)
Department of Psychology, Binghamton University
Binghamton, NY 13905 USA

Abstract

Research on the internal structure of categories has shown that central tendency, frequency of instantiation, and ideals are determinants of graded structure (Barsalou, 1985). Vast differences in the relative roles of these factors have been observed depending on the type of category, the immediate context of evaluation, and the domain knowledge or experiential history of the individual. In two experiments, using the simplest possible domain of novel artificial categories (lines of varying lengths), we tested whether the way in which a category is learned can shape its internal organization. Specifically, the impact of *co-learning* multiple categories within a classification scheme was assessed using typicality ratings. We found that the internal structure of a category took qualitatively different forms (i.e., graded structure based on ideals or central tendency; or even no graded structure at all) depending on the nature and number of the contrast categories during learning.

Introduction

A foundational aspect of the scientific understanding of human concepts and categories is the phenomenon of graded structure. Specifically, people systematically judge some members of a category to be better (i.e., more typical or representative) than others (Rosch & Mervis, 1975; Rips, Shoben, & Smith, 1973). A traditional example is that *robin* is seen as a good example of the category *bird,* while *penguin* is not considered typical. It is widely agreed that natural categories are organized in terms of family resemblance (Rosch & Mervis, 1975). Direct and indirect measures of item typicality are well-predicted by the proximity or degree of feature overlap between an example and the central tendency of the category. On one view, the graded structure of psychological categories closely parallels and reflects the correlational structure of the environment (Rosch & Mervis, 1975). Taking the *bird* category for example, the high typicality of *robin* arises from data showing that robins possess many of the properties common to birds (i.e., they fly, are of average size, coloration, feathers, etc.). An additional factor beyond centrality is the frequency of instantiation of a category member (Barsalou, 1985; Nosofsky, 1988).

Counter to the view that typicality fundamentally mirrors the environment, it has been shown that the basis for graded structure varies between different types of categories (Barsalou, 1985) and, further, that the graded structure of a particular category varies across contexts and judges (Barsalou, 1987; Roth & Shoben, 1983). This is to say, the typicality of an example is not strictly determined by the centrality of its feature values and its frequency. In addition, goal-optimizing properties called ideals have been shown to powerfully predict the organization of natural taxonomic and goal-derived categories (Barsalou, 1985; Lynch, Coley, & Medin, 2000; Bailenson, Shum, Atran, Medin, & Coley, 2002; Burnett, Medin, Ross, & Block, 2005). In many cases, the examples of a category that best serve the goals or interactions between a person and object are the ones seen as most typical.

This can be seen most dramatically in the case of goal-derived categories (Barsalou, 1985). For example, in a category such as *foods to eat on a diet*, the items judged most typical are those which contain the least number of calories (i.e., the category is organized around the ideal of *zero calories*). For goal-derived categories, Barsalou (1985) found that graded structure depended systematically on proximity to ideals, not to central tendency. Further, ordinary taxonomic categories were determined both by central tendency and ideals (as well as frequency). As seen in the work of Medin and colleagues, experts' categories are largely shaped by ideals and different populations with varying ideals or perspectives on a domain show different typicality profiles. For example, Burnett, et al. (2005) reported a study of fishermen from two cultures in the same geographical region with different notions of desirability about fish. Participants from the two cultural groups produced typicality ratings for various kinds of fish, and the ratings were best predicted by goal-relevant ideals grounded in cultural desirability.

Medin and colleagues suggest that the evidence supporting the priority of central tendency in judgments of typicality has arisen mostly in cases in which participants know relatively little about the domain. Lynch, et al. (2000) found that for tree experts, the best examples of the concept *tree* were those most consistent with the ideals of maximum height and minimal weediness – as opposed to the best examples being those with the most central values on these two (or other) features. Further, even the novices with little domain knowledge did not base their typicality judgments on central tendency; for this group, familiarity was the best predictor. Finally, Lynch, et al. (2000) found that the organization of expert categories systematically reflected the type of expertise (e.g., landscapers versus taxonomists).

Given this background, we point out two unique properties about ideals: (1) they appear to show an influence of top-down feature construction or selection based on knowledge or experience (Wisniewski & Medin, 1994; Murphy & Medin, 1985); and (2) they represent extreme points rather than central points of feature values. To

illustrate the first point, examples of *foods to eat on a diet* do not offer readily available (i.e., perceptual) information about their calorie levels, yet this is the critical ideal underlying the category organization. To illustrate the second point, note that the ideal for the example above is *zero calorie*s, an extreme value along the dimension that is optimal with regard to the goal.

In accord with the theorists mentioned above, we believe that ideals emerge and mediate category structure as a result of goal-directed experience. More specifically, we suggest that ideals are based on properties that are highlighted (or constructed) via classification learning. For example, in the context of landscaping, one makes classification judgments with regard to the desirability of particular trees for particular goals – the learning process associated with such category decisions picks out ideals such as *height* and *weediness*. We do not discount the potential role of theory-like knowledge, but the current proposal has the advantage of being fairly clear to articulate and test. Our core claim is that the internal organization of a category is at least partially determined (above and beyond the statistics of the environment) by the process of learning to distinguish among task-relevant contrast categories.

The interrelatedness of categories has already received some research attention. Nosofsky (1988) found that the frequency of presentation of category instances during learning not only affected the typicality of those items, but also influenced the typicality of members of the contrast category. Goldstone (1996) offered a theoretical analysis entailing a continuum between isolated and interrelated concepts. One especially interesting aspect of interrelated categories is that (in contrast with isolated categories) extreme examples or caricatures are more accurately categorized than central examples (Goldstone, 1996; Goldstone, Steyvers, & Rogosky, 2003; Palmeri & Nosofsky, 2001). The latter study using ill-defined categories is unique in that the extreme examples were extreme only in their (learned) psychological representations; that is, the examples were actually prototypical of their category as physical stimuli.

We focus on a basic component of classification learning – the number and the nature of the contrast categories against which a target category is learned – in order to evaluate the role of *co-learning* as a determinant of graded structure. Co-learning refers to the presence of contrast categories in a classification training set and choice set. We predict that the exact same set of category members will show systematically different graded structure depending on its status relative to contrast categories.

An extremely simple set of novel, artificial categories are used to evaluate the influence of co-learning on graded structure. We expect essentially perfect accuracy in this task and direct our focus toward the ratings of item typicality collected after the learning phase using single-item presentations. Learning a single category in isolation is likely to produce graded structure based on central tendency (since the learners are novices who lack domain knowledge). There is no evidence presently available, to our knowledge, as to whether or in what way the graded structure underlying a category representation will be affected by co-learning with one or more categories from the same domain. The closest related evidence comes from the study of learned categorical perception where it has been shown that examples become more or less easily discriminated depending on their status relative to the category boundary (Tajfel & Wilkes, 1963; Livingston, Andrews, & Harnad, 1998).

Experiment 1

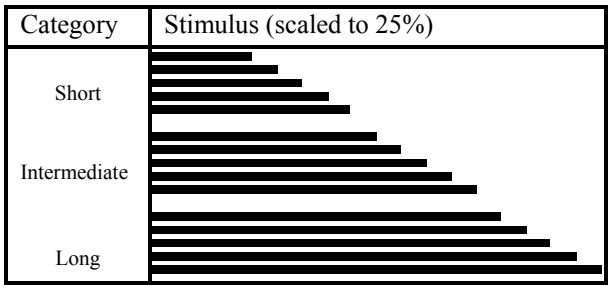
This study was designed to evaluate the internal structure of categories after either 2-way or 3-way classification learning of simple categories of examples varying in line length. A preliminary data set revealed that participants who were introduced to only one category of lines exhibited a central tendency organization in their item-by-item typicality ratings. Specifically, on average the mid-length line was rated as the most typical and the extreme lines (the shortest and longest) were the least typical. In the current experiment, the same target category was investigated under two conditions of co-learning: (1) with a second category of continuously varying line lengths; and (2) with two additional categories (one consisting of longer length examples and one consisting of shorter length examples). Graded structure based on central tendency would be evidenced by an inverted U-shaped curve (as in the preliminary data), while ideal-based typicality would be evidenced by monotonically decreasing ratings with a high point located at the extreme value of line length (i.e., the longest of the long or the shortest of the short).

Method

Participants. 158 undergraduates from Binghamton University participated in this experiment in order to fulfill a course requirement.

Material and Design. Stimuli consisted of 15 black lines separated by length into three categories. All lines were 15 pixels wide and lengths (ranging from 200 to 900 pixels) were constructed to reflect a consistent difference (of 50 pixels) between each item length within a particular category as well as between the closest members of two consecutive categories (see Figure 1). Short lines had lengths of 200, 250, 300, 350, and 400 pixels, intermediate lines had lengths of 450, 500, 550, 600, and 650 pixels, and long lines had lengths of 700, 750, 800, 850, and 900 pixels. The experiment used a between-Ss design with two conditions. In the 2-way learning condition (n = 94), only the short and intermediate lines were learned. In the 3-way learning condition (n = 64), all three sets of lines were learned. Note that the preliminary data showing centrality-based organization was collected using the intermediate lines.

Figure 1: Stimulus Materials for Experiment 1.



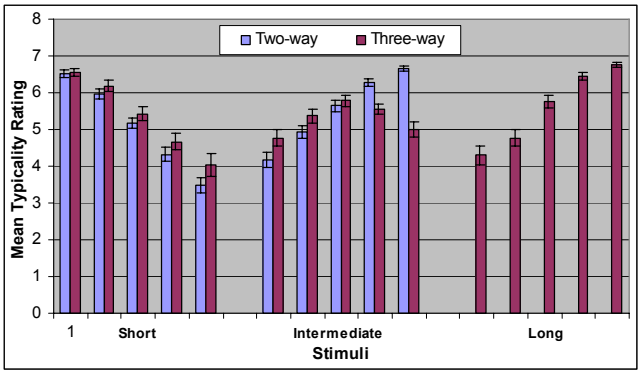
Procedure. Participants were randomly assigned to one of the two conditions. Participants were instructed that they would be exposed to a series of lines and asked to categorize them into types. Each participant completed a series of learning blocks in which all items were presented one at a time in a random order. The lines were displayed horizontally in the center of the computer display. Below the example, participants were asked to click with the mouse on the button with the correct category label. The categories were labeled with arbitrary names (Type K, Type H, and Type R) that were selected to convey no meaning to the learner. After making a classification response, participants were given correct/incorrect feedback including the correct category. A learning criterion was set at twenty consecutive correct responses. Participants who did not reach criterion proceeded to the test phase after eight blocks of training (80 trials in the two-way learning condition and 120 trials in the three-way learning condition).

During the test phase each of the training items was displayed one at a time in a random order. Participants were asked to judge the typicality of the item, specifically: "How good an example is this of the category?" Ratings were collected on a scale from 1 (poor example) to 7 (very good example). Before providing each rating, the participants were asked to select the correct category (no feedback was given). In addition to the typicality ratings, classification accuracy and response times were recorded for the learning and test phases.

Results and Discussion. In the two-way learning condition, 70% of participants reached criterion (in an average of 50 trials). The two-way learners were highly accurate in classification at test (M = .95, SD = .06). In the three-way learning condition, only 22% of participants reached criterion (in an average of 100 trials), but still showed high accuracy at test (M = .90, SD = .07). No meaningful differences were observed when the learners who failed to reach criterion were removed from the analyses. Accordingly, all participants (regardless of whether they met the criterion) were included in the analysis of the test phase data.

Unlike in the preliminary data based on learning only a single category, ideal-based graded structure (see Figure 2) was found for both categories (lines of short and intermediate length) of the two-way learning condition as well as in both of the extreme categories of the three-way learning condition (lines of short and long length). By contrast, the category of intermediate lines in the three-way learning condition showed graded structure relative to the central tendency of the category. The salient characteristic of these data is that the category of intermediate line lengths showed a qualitatively different internal structure depending on whether or not an additional contrast category (the long length lines) was included in the training phase.

Figure 2: Mean Typicality Ratings by Condition and Item in Experiment 1.



In order to quantitatively distinguish the profiles of graded structure, a standard regression formula was used to calculate slopes of the typicality ratings across the five items in a category. This measure effectively identified the degree to which the typicality ratings were influenced by ideals (greater slope values, either positive or negative, shows graded structure relative to the extreme point) or by central tendency (approximately zero slope). The slope was averaged over participants for each category in each condition and compared using independent t-tests. All ratings generated after an incorrect classification were excluded from the analysis.

In both conditions the ideal of minimum line length had a great influence on the graded structure of the category of short lines. The mean slope of the typicality distributions for the short category was highly negative in both the two-way learning condition (M = -.76, SD = .55) and the three-way learning condition (M = -.77, SD = .56). In a consistent manner, the mean slope for the category of long length items in the three-way learning condition was highly positive M = .73, SD = .52). The intermediate length items showed a highly positive slope as an extreme category in the two-way learning condition (M = .65, SD = .48), but showed near-zero slope as the intermediate category in the three-way learning condition (M = .06, SD = .51). The difference between the slopes for the category of intermediate length items was significant, $t(156) = 7.413$, $p < .01$, between the two conditions. In addition the slope for the intermediate category in three-way learning was obviously different than that of the extreme categories.

In a supporting set of analyses, the mean typicality ratings for the items of each category in each condition were

compared. Specifically, independent *t*-tests revealed that the typicality of the extreme example (i.e., the shortest of the short or the longest of the long) was significantly greater than the central example for the ideal-based categories (those with highly positive or negative slopes), and for the one remaining category (the intermediate length category of the three-way learning condition), the central example was rated significantly higher than the extreme examples (the details of this analysis are not included due to limited space).

Our most surprising finding is that the very same target category (the intermediate length lines) showed an entirely different, yet in all cases systematic, graded structure depending upon the learning conditions. After exposure to only the target category, a central tendency basis for graded structure was observed. However, when the category represented a set of extreme values relative to a contrast category, typicality ratings were ideal-based (as in the short and intermediate length categories after two-way learning and the short and long length categories after three-way learning). Finally, when the target category was co-learned such that its membership occupied the mid-range of the critical feature (i.e., the intermediate length category under three-way learning), an inverted U-shaped profile was observed was preserved. The typicality profile in this case is actually open to two interpretations. One possibility is that this category was unaffected by co-learning and simply showed the same pattern as in the single-category case. Alternatively, it is possible that this category also showed ideal-based organization cohering around an ideal of extreme *middleness*. This strikes us as unlikely, but either account predicts the inverted U-shape form of the distribution that was observed.

We note that while the classification accuracy was quite high at test, the errors that were made tended to occur near the category boundaries. Response times also showed longer latencies for the border items presumably due to greater difficulty. While the typicality ratings for incorrect classifications were not included in the analyses of graded structure, it remains possible that lower ratings were given to the border items due to low confidence about their category membership (rather than due to their low proximity to a reference point). The following study was designed to address the issues of the ambiguous interpretation of the intermediate length category under three-way learning, as well as the issue of reduced confidence at the category boundaries.

Experiment 2

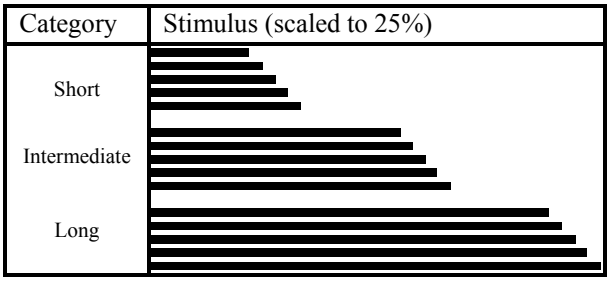
Our goal was to replicate and extend the findings of the first experiment by evaluating the role of category distinctiveness. Specifically, we tested whether the effect of co-learning would be influenced by the ease with which the categories could be distinguished from one another. Additionally, we sought to ensure that the lower levels of typicality for boundary items were not due to participants' uncertainty with regard to category membership. A set of materials was developed with increased length differences between the categories. This use of clustered, rather than continuously varied, line lengths across the categories was expected to improve the levels of classification accuracy. In addition, the easily differentiated categories provided a means of better understanding the impact of co-learning. Specifically, would the intermediate category under three-way learning continue to show the inverted-U profile?

Method

Participants. 117 undergraduates from Binghamton University participated in this experiment to fulfill a course requirement.

Material and Design. Stimuli for the second experiment were 15 black lines of the same width as in the first experiment. Although the lengths of these lines also ranged from 200 to 900 pixels, they were chosen to reflect a greater difference in length between the categories. As a consequence, there was a smaller difference of 25 pixels between each category member (see Figure 3). Short lines had lengths of 200, 225, 250, 275, and 300 pixels, intermediate lines had lengths of 500, 525, 550, 575, and 600 pixels, and long lines had lengths of 800, 825, 850, 875, and 900 pixels. This experiment, like the first, used a between-Ss design with two conditions. In the two-way learning condition (n = 58), short and intermediate line categories were learned. In the three-way learning condition (n = 59), all three sets of lines were learned.

Figure 3: Stimulus Materials for Experiment 2.



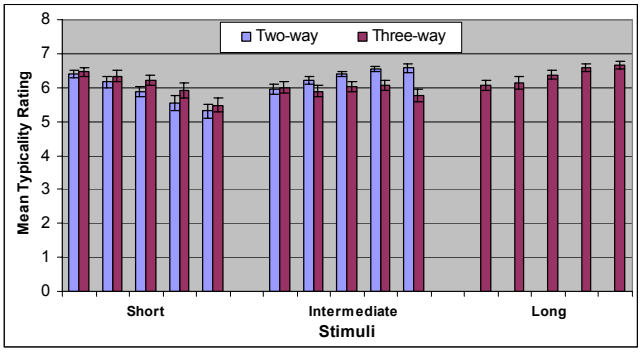
Procedure. The procedure was exactly the same as in the previous experiment.

Results and Discussion. As expected, participants learned the categories more quickly and effectively when the categories were clustered rather than continuous. In the two-way learning condition, 95% of participants reached criterion (in an average of 31 trials) with 100% overall accuracy at test. Participants in the three-way learning condition performed similarly, with 98% of participants reaching criterion (in an average of 38 trials), and categorizing with 99% overall accuracy at test.

Consistent with the results of the first experiment, the typicality ratings for the categories of clustered stimuli were again determined by ideals rather than central tendency – and again there was one important exception in the case of the intermediate category under three-way learning (see

Figure 4). In the current results, however, the typicality ratings for the intermediate of the three categories exhibited neither ideal-based nor central tendency-based graded structure. Instead, to our surprise, the distribution was essentially flat.

Figure 4: Mean Typicality Ratings by Condition and Item in Experiment 2.



A possible explanation for this finding emerges from the differences between the two experiments. One difference between the data from the previous experiment (conducted using continuous stimuli) and the present experiment with clustered stimuli is that the range of the mean typicality ratings for each category was smaller in the current study. As more distance was created *between* categories in order to evaluate category distinctiveness, the difference in the lengths *within* each category was decreased. This change in length difference (from 50 pixels to 25 pixels) may have resulted in a decreased overall ability to distinguish between the members of each category. This speculation is supported by the fact that 10% of the participants in the two-way condition and 15% of the participants in the three-way condition responded with the same rating for all of the stimuli, compared to only 3% and 1%, respectively, in the first experiment (note that the experimental findings were fully preserved when these uniform responders were removed from the analyses).

However, the existence of systematic ideal-based graded structure for the extreme categories suggests that reduced item distinctiveness was probably not an overall determining factor. Graded structure was extinguished only in the intermediate category. We interpret the current evidence to suggest that the internal structure of the intermediate category under three-way learning was based on central tendency with the continuous categories of Experiment 1, but was essentially incoherent with the clustered categories of Experiment 2. The incoherent internal structure arose because the boundaries were so clear between the categories that there was no need to become familiar with the distributional properties of the intermediate category. Essentially, the intermediate category may have acted as a kind of *junk* category lacking a positive definition; any item that was neither short nor long would belong to this class, and the common core among these items did not require analysis. Evidence favoring this interpretation emerges from an analysis of individual rating data. The flat profile of the mean typicality ratings was not actually representative of the performance of most of the individual learners. Instead the flat distribution reflected the average of a wide-range of ostensibly unsystematic profiles. However, this lack of coherence does not appear to reflect an inability to tell the examples apart. Compared to the other two categories, a smaller percentage of participants rated all of the intermediate length items as equally good examples of their category. Therefore, it appears that learners were sensitive to the item differences, but lacked a reference point for systematically assessing typicality.

Quantitative analyses support our interpretation of the data. Most importantly, the typicality ratings for the category of intermediate length items after two-way learning had a mean slope (M = .16, SD = .30) that was significantly greater than the slope for the same items in the three-way learning condition (M = -.02, SD = .30), $t(116) = 3.26$, $p <$ .01. Therefore, the category of intermediate items was organized according to the ideal of maximum length under two-way learning, but showed no graded structure under three-way learning.

While the clustered intermediate category lacked systematic organization when learned along with both shorter and longer lines, the pattern of graded structure for the extreme categories was consistent with the findings of Experiment 1. The main difference from the previous results was that the slopes were less steep – an apparent reflection of the reduced within-category distinctiveness. The mean slope of the item typicality ratings for the category of short length items was negative for two-way learning (M = -.28, SD = .39), as well as for three-way learning (M = -.24, SD = .44), indicating ideal-based graded structure. Likewise, the mean slope for the category of long length items in the three-way learning condition was positive (M = .16, SD = .31). Pairwise t-tests revealed that for the long and short categories, the extreme examples were rated as reliably more typical than the central examples.

While the internal structure of the intermediate category was found to be qualitatively different between experiments, it is important to note that that the mean slope of typicality distributions for this category of clustered stimuli (M = -.02, SD = .31) did not differ significantly from that found with corresponding continuous stimuli (M = .06, SD = .51). This is due to the fact that a measure of slope does not distinguish between a lack of graded structure and a graded structure based on central tendency (both represented by a slope approaching zero). In order to clearly quantify the difference between these two profiles of typicality ratings, a measure of variability (the standard deviation) of the typicality ratings for each of the five items was calculated. An inverted-U shape curve will show much greater variability than a flat line, despite their similar slopes. In a cross-experiment comparison, we found that the average standard deviation for the intermediate category of the three-way condition in the first experiment (M = 1.06, SD = .64) was significantly greater than that of the intermediate

category under three-way learning in the current experiment (M = .63, SD = .60), $t(121) = 3.82$, $p < .01$.

General Discussion

We found that: (1) learning one simple artificial category produces a central tendency-based organization; (2) under co-learning conditions, the bookend (i.e., extreme-valued) categories always showed ideal-based graded structure; (3) the intermediate category in three-way learning did not show ideal-based graded structure, but instead showed a central tendency basis when the between-category similarity was high (difficult boundaries) and showed a lack of systematic internal structure when the between category similarity was low (easy boundaries). One could argue that the difference in the intermediate categories between the clustered and continuous version was attributable to the difference in the within-category similarity, however, the fact that ideal-based organization was observed in both experiments for the extreme categories speaks against this interpretation.

The primary implication of our results is that co-learning is a powerful determinant of graded structure for categories varying along a single dimension. The same target category was organized differently depending on whether it was learned alone, as an extreme category, or as an intermediate category. More broadly, these results suggest that the role of ideals in the organization of natural categories may be grounded in task-driven classification experience and the relative properties of co-learned categories. Our goals for future work include extending this paradigm to multidimensional and naturalistic category domains, elucidating the underlying mechanism through which the internal structure of a category takes form, and further evaluating the relationship between learning tasks, learning modes, and goal-related activity (see Markman & Ross, 2003).

Acknowledgements

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

References

Bailenson, J. N., Shum, M. S., Atran, S., Medin, D. L., & Coley, J. D. (2002). A bird's eye view: Biological categorization and reasoning within and across cultures. *Cognition, 84*, 1-53.

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

Barsalou, L. (1987). The instability of graded structure: Implications for the nature of concepts. In U. Neisser, (Ed.), *Concepts and Conceptual Development: Ecological and Intellectual Factors in Categorization*. Cambridge: Cambridge University Press.

Burnett, R. C., Medin, D. L., Ross, N. O., & Blok, S. V. (2005). Ideal is typical. *Canadian Journal of Experimental Psychology, 59,* 3-10.

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

Livingston, K. R., Andrews, J. K., & Harnad, S. (1998). Categorical perception effects induced by category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*, 732-753.

Lynch, E. B., Coley, J. D., & Medin, D. L. (2000). Tall is typical: Central tendency, ideal dimensions and graded category structure among tree experts and novices. *Memory & Cognition, 28, 41-50.*

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

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

Nosofsky, R. M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 700-708.

Palmeri, T. J. & Nosofsky, R. M. (2001). Central tendencies, extreme points, and prototype enhancement effects in ill-defined perceptual categorization. *The Quarterly Journal of Experimental Psychology, 54A,* 197-235.

Rips, L. J., Shoben, E. J., & Smith, E. E. (1973). Semantic distance and the verification of semantic distance. *Journal of verbal learning and verbal behavior, 12*, 1-20.

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

Roth, E. M. & Shoben, E. J. (1983). The effect of context on the structure of categories. *Cognitive Psychology, 15,* 346-378.

Tajfel, H. & Wilkes, A.L. (1963). Classification and quantitative judgement. *British Journal of Psychology, 54*, 101-114.

Wisniewski, E. J., & Medin, D. L. (1994). On the interaction of theory and data in concept learning. *Cognitive Science, 18*, 221-281.