(Not) Learning a Complex (but Learnable) Category

David Danks (ddanks@cmu.edu)
Department of Philosophy, Carnegie Mellon University, 135 Baker Hall
Pittsburgh, PA 15213 USA; and
Institute for Human & Machine Cognition, 40 S. Alcaniz St.
Pensacola, FL 32502 USA

Abstract

Recent theoretical research has argued that multiple psychological theories of categorization are mathematically identical to inference in probabilistic graphical models (a framework developed in statistics and computer science). These results imply that the major extant psychological theories can all be represented mathematically as special cases of inference in (subclasses of) chain graphs, a particular type of probabilistic graphical models. These formal results suggest that people should be capable of learning significantly more complicated category structures than can be expressed in the standard psychological theories. In this paper, we present an experiment in which people apparently *failed* to learn the complex category, though a significant group of participants seemed to have learned something about the contrast category. Although inferences to cognitive failure are notoriously problematic, these results suggest that the hyper-general theory useful for the mathematical equivalencies does not accurately describe human categorization.

Introduction

Much of the experimental research on category learning has aimed to distinguish between various theories by presenting people with categories that can be learned according to one theory, but not another. After some learning period with the categories, experimental participants are tested to determine how closely their category representations match the true category structure. In contrast, there has been relatively little experimental research investigating the limits of category learning: specifically, whether there are categories that can be learned (in principle) by some statistical procedure, but not by people. Most research on the limits of category learning has focused on constraints from other cognitive systems (e.g., memory bounds). In contrast, this paper is a preliminary attempt to ask whether some category structures are sufficiently complex (in a statistical sense) that people are unable to learn them, even though they are sufficiently structured that they are (in principle) learnable.

We first describe a set of theoretical and mathematical results that connect psychological theories of categorization with inference to model structure in the computational framework of probabilistic graphical models. Those formal results suggest that people might be able to learn categories with a (previously unstudied) complex statistical structure. We thus conducted a category learning experiment using this statistical structure, and found suggestive evidence that people were in fact unable to learn this category.

Theoretical Background

This section outlines the mathematical/theoretical results that motivate the experiment reported later in the paper. The central claim of this section is that most psychological models of categorization can be represented as inference in various types of probabilistic graphical models. Due to space constraints, this exposition is necessarily at a high level. The precise formulations of the frameworks, theories, and equivalencies can all be found in the cited works.

Probabilistic Graphical Models

At a high level, probabilistic graphical models use a graph to encode independencies in a probability distribution. This compact encoding of the independence relations can then be used to dramatically speed up inference, learning, and prediction. The two most common types of graphical models are Bayesian networks and Markov random fields.

Bayes nets (e.g., Pearl, 1988, 2000; Spirtes, Glymour, & Scheines, 1993) represent a probability distribution using a directed acyclic graph (where the variables are nodes in the graph). For example, $A \rightarrow B \leftarrow C$ encodes the independence pattern in which A and C are unconditionally independent, but no other pairs are independent (conditionally or unconditionally). Bayes nets are widely used to model causal relationships, and have more recently been used to model people's psychological representations of causal structure (Gopnik, Glymour, Sobel, Schulz, Kushnir, & Danks, 2004; Griffiths & Tenenbaum, 2005; Lagnado & Sloman, 2004; Tenenbaum & Griffiths, 2001; Waldmann & Martignon, 1998).

Markov random fields (Lauritzen, 1996) represent the independence structure of a probability distribution using an undirected graph. For example, A — B — C implies that "A independent of C given B" is the only independence in the probability distribution. Markov random fields have frequently been used to model spatially correlated data (e.g., image data). They have not been widely used to model cognitive phenomena, in part because the edges do not have an obvious interpretation (in contrast with Bayes nets).

In general, we focus on probability distributions that can be *perfectly represented* by some graphical model: that is, cases in which the graphical model (whether Bayes net or Markov random field) predicts all and only the independencies that are found in the probability distribution. The set of probability distributions that can be perfectly represented by a Bayes net only overlaps with the set of those that can be perfectly represented by a Markov random field. That is, there are probability distributions that can only be perfectly represented in one of the two formalisms (but also some that can be perfectly represented by both).

Finally, *chain graphs* (Lauritzen & Richardson, 2002; Lauritzen & Wermuth, 1989) provide a unifying framework for Bayes nets and Markov random fields, which are the two most common types of graphical models. Specifically, chain graphs can contain both directed and undirected edges in the same graph, and so can perfectly represent a richer set of probability distributions.

Standard Accounts of Categorization

Four significant (classes of) psychological theories of categorization are: exemplar-based, decision bound, prototype-based, and casual model theories. *Exemplar-based models* (Kruschke, 1992; Medin & Schaffer, 1978; Nosofsky, 1984, 1986) represent a category by a set of exemplar instances, where each exemplar must previously have been observed to be a member of the category. To categorize some novel instance, one first computes the "similarity" of the instance to each possible category. This similarity is the weighted average distance in "similarity space" between the novel instance and each exemplar in the category. The similarities for each category are then integrated into a probabilistic response using the Shepard-Luce rule (Luce, 1963; Shepard, 1957).

Decision bound models (Ashby & Townsend, 1986) represent categories as regions of "feature space," and categorization decisions are made by determining the region in feature space in which some instance most likely resides (given some model of perceptual noise). From a mathematical point-of-view, decision bound models are closely connected to exemplar-based models. Ashby and Maddox (1993) extensively explored the formal connections between exemplar-based categorization models and decision bound models in general recognition theory. In general, they found that there are strong equivalencies between these model-types, and so for space reasons, we do not explore decision bound models more closely in the remainder of this paper.

Prototype-based models (Minda & Smith, 2001, 2002) represent a category by a prototypical instance (i.e., a specific point in the relevant feature space). Standard prototype models are formally similar to exemplar-based models with only one exemplar, though the prototypical instance does not have to be observed. Probabilistic categorization responses are then generated using the Shepard-Luce rule. These models typically contain only first-order (observed) features, and so have no interaction terms. In order to capture the intuition that the prototype can encode (in some sense) a "summary" of the observed data, prototype models can also contain second-order features: variables whose value is entirely determined by two first-order features (as in Rehder, 2003a; Rehder, 2003b).

Finally, *causal model theory* (Rehder, 2003a, 2003b; Rehder & Hastie, 2004) holds that some categories are defined by causal structure: individuals are members of the same category just when their observable features are generated by the same underlying causal structure. Formally, causal structures are represented using Bayes nets, and the similarity (of a novel instance to a category) is equal to the probability of the category's causal structure generating a case with the given features.

Categorization Theories and Graphical Models

Consider the problem of "categorization" from the graphical models point of view. In particular, suppose that we have some set of categories that are described by probability distributions, where each can be perfectly represented by a suitable probabilistic graphical model. Given these models and some novel case, we can straightforwardly determine (using Bayesian updating) the probability that the novel case was drawn from each of the probability distributions. That is, for each graphical model G under consideration, we can compute $P(G \mid X)$ for the instance X. At least superficially, these conditional probabilities resemble the predictions of the various categorization theories (which all have the form 'P(Say "A" | X)').

Danks (2004) proves that the resemblance is more than superficial. The psychological theories of categorization are each mathematically equivalent to computing $P(G \mid X)$, where G is restricted to be a particular type of graphical model. Differences among (the graphical model versions of) the different psychological theories arise because the possibility space is restricted in different ways. The relationship between categories in psychological theories and graphical model classes can be summarized as:

- Exemplar-based categories are equivalent to (a subclass of) Bayes nets with the structure shown in Figure 1;
- Prototype-based categories are equivalent to (a subclass of) Markov random fields; and
- Causal model categories are equivalent to Bayes nets with arbitrary structure.

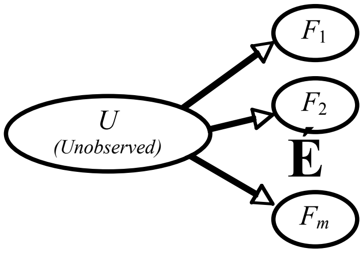

Figure 1: Bayes net structure for exemplar categories

Danks (in press) explores some methodological and theoretical implications of these mathematical equivalencies (e.g., enabling connections with causal learning research, or explaining recent experimental results). Our focus here is on one particular possibility raised in that chapter. Since categories in the psychological theories correspond to special cases of chain graphs, perhaps all psychological categorization is simply Bayesian categorization on various chain graphs. That is, the extant psychological theories might simply be particularly salient special cases of categorization that arise because Bayes nets and Markov random fields are the simplest types of chain graphs.

If this suggestion is correct, then people should be capable of learning categories that are perfectly representable only by a full-fledged chain graph (but not a Bayes net or

Markov random field). These probability distributions are quite complex, and so learning such a category should be challenging.

Experiment

The experiment reported here had a very simple central question: Can people learn a category whose distribution (over features) can be represented by a chain graph, but not a Bayes net or Markov random field?

Experimental Design

Consider the chain graph: F1 → F3 — F4 ← F2. This chain graph implies: F1 and F2 are unconditionally independent; F1 and F4 are independent conditional on {F3, F2}; and F2 and F3 are independent conditional on {F4, F1}. Probability distributions with these independencies cannot be perfectly represented by any Bayes net or Markov random field. That is, any graphical model with only directed edges, or with only undirected edges, will necessarily imply strictly more or strictly fewer independencies than actually obtain in this probability distribution. This chain graph is the simplest one that cannot be perfectly represented by a Bayes net or a Markov random field.

The formal equivalencies between the psychological theories of categorization and inference in probabilistic graphical models enable us to conclude immediately that any probability distribution perfectly representable by this chain graph cannot be perfectly represented by a prototype or causal model learner. The distribution could be learned by exemplar-based category learning, but only by encoding all of the exemplars (twelve in this experiment). A category based on this probability distribution thus provides a critical test case: none of the current psychological theories of categorization predict that people will easily learn this category, and only the exemplar theories predict that anything at all could be learned. At the same time, since the distribution can be perfectly represented with the chain graph, it is (in principle) learnable.

The distribution in the Target column of Table 1 can only be perfectly represented by this chain graph, and so serves as the target category for our learning experiment. For our contrast class, we use a multiplicative prototype category in which the central prototype is F1 = F2 = 1, and F3, F4 are irrelevant. F1 and F2 have equal weights in the category, and so cases with only one of the two features occurs half as frequently as the prototypical cases; cases without either feature are one-fourth as frequent. The case distribution for the contrast category is also provided in Table 1.

Four of the cases listed in Table 1—those with F3 = F4 = 1—were not shown to participants, and thus provide an instrument for measuring category generalization. For completeness, we have provided the implied counts for those four cases using numbers in double brackets.

In deterministic learning scenarios, high performance can result simply by memorization of salient or common exemplars, and not from any understanding of the category structure. Therefore, we deliberately used a probabilistic categorization task because we wanted to know whether people could learn the underlying *distribution*. Because the task was probabilistic, perfect performance was not possible. Optimal performance—i.e., correctly choosing the more likely category for each case—results in 66.67% correct classification (on average).

Table 1: Experimental case distribution

F1	F2	F3	F4	Target	Contrast
0	0	0	0	4	1
0	0	0	1	2	1
0	0	1	0	2	1
0	0	1	1	0 [[1]]	0 [[1]]
0	1	0	0	2	2
0	1	0	1	4	2
0	1	1	0	1	2
0	1	1	1	0 [[2]]	0 [[2]]
1	0	0	0	2	2
1	0	0	1	1	2
1	0	1	0	4	2
1	0	1	1	0 [[2]]	0 [[2]]
1	1	0	0	1	4
1	1	0	1	2	4
1	1	1	0	2	4
1	1	1	1	0 [[4]]	0 [[4]]

Participants and Materials

Eighty-eight Carnegie Mellon students were compensated $10 for participation in a group of experiments containing this one. The full set of experiments took approximately 45 minutes to complete.

The experiment was done on computers. Participants were placed in the role of biologists classifying two novel species of insects. To help them learn how to categorize, they were presented with a sequence of "already classified" insects. For each case in the learning sequence, participants were presented with an image of the four-featured insect and asked to classify it as a "Marbock" or "Wermer." After categorizing the insect, participants were given feedback about the actual insect name, as well as whether their answer was correct.

Participants were told that the learning phase would last until they were "sufficiently good at classifying these bugs." Due to the difficulty of the learning task, however, we did not actually use a specific performance criterion. Instead, we simply presented every participant with two complete sets (108 total cases). Within each block of 54 cases, the cases were presented in a randomized order.

After completing the learning phase, participants were presented with all sixteen possible insects (randomized order) and asked "how likely is it that this bug is a [TARGET]?" where 'TARGET' was replaced by the chain graph category name. Participants provided a 0-100 rating using a slider movable only in increments of five (i.e., the rating was functionally on a 0-20 scale). The lower and upper ends of the rating scale were respectively labeled "Cannot be a [TARGET]" and "Definitely a [TARGET]."

We have found in other experiments that participants can successfully learn other categories using the same interface and functionally similar images (Zhu & Danks, in prep). Thus, there is no *a priori* reason to think that the interface impedes category learning (to any significant degree).

Results and Discussion

One measure of successful category learning is performance in the learning phase prediction task. Since we randomized the presentation order within each 54-case block, we cannot directly compare performance in smaller intervals (since the optimal performance might differ across individuals). A histogram of percent correct responses in the second half of the learning phase is given in Figure 2.

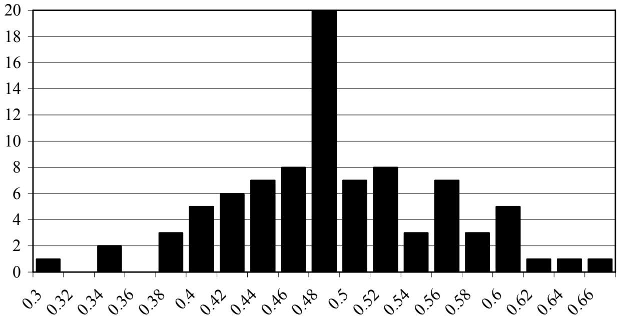

Figure 2: Histogram of second half learning rates

This performance distribution is not significantly different from normal ($p > .70$; Shapiro-Wilk test), and the mean is not significantly different from 0.5 ($p > .70$; one-sample t-test). Thus, we have *prima facie* evidence that participants did not (in general) learn the categories: their performance in the second half of the learning phase is not statistically distinguishable from what one would expect from chance performance in a population of this size.

The sixteen likelihood ratings also provide information about learning performance: specifically, did any of the participants learn (something like) the correct category structures? The mean ratings (error bars indicate 95% confidence intervals), as well as the correct likelihood for observed cases, are shown in Figure 3.

There are 120 pairwise comparisons of ratings, and so we applied the Benjamini & Yekutieli (2001) false discovery rate correction (henceforth, BY-FDR) to the two-sample paired t-test p-values. After applying this correction, there was only one significant difference in ratings: 1010 vs. 1111 ($p < .05$). This analysis of the ratings for the full population thus supports the previous analysis: people do not seem to have learned the categories in this experiment.

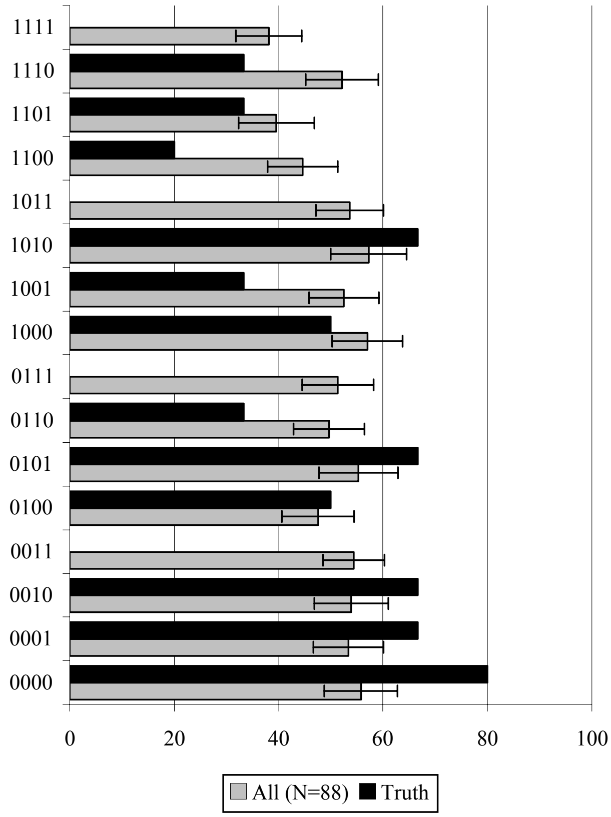

Figure 3: Mean ratings for all participants

At a more fine-grained level, however, the data present a subtler picture. If the performance distribution in the learning phase is due primarily to chance, then we would expect to find statistically similar ratings from participants with (i) above-chance performance, and (ii) below-chance performance. In the same spirit, if any individuals actually did learn something about the category structures, then we would expect that this knowledge would translate into above-chance performance. Thus, if any significant learning occurred, then the above-chance performers should exhibit a better understanding of the category structures.

To test for increased understanding, we split the participant population into two groups: (i) those with second-half learning phase performance ≥ 0.5 ($N = 45$); and (ii) those with performance < 0.5 ($N = 43$). This split was based on learning phase performance, and so we cannot perform any meaningful analyses on differences in performance.

The mean group ratings are shown in Figure 4. On the surface, these groups are not particularly different: None of the likelihood ratings are significantly different (two-sample t-tests with BY-FDR correction).

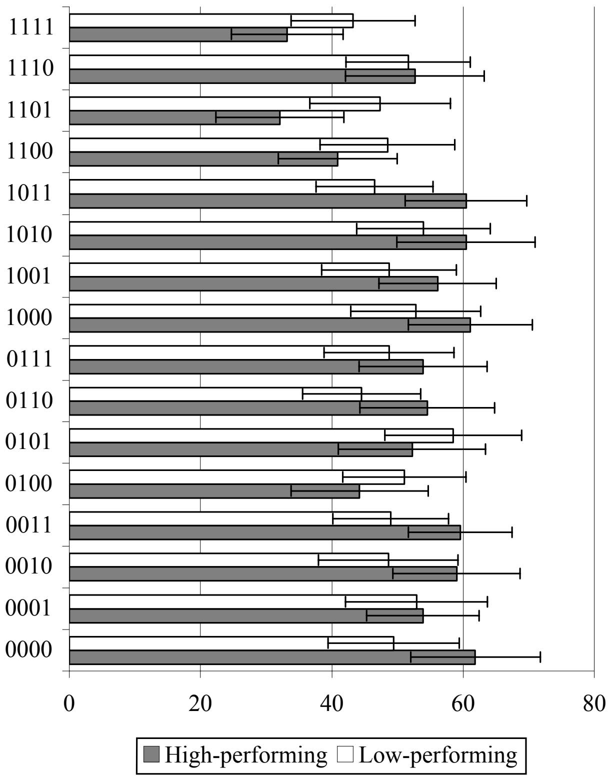

Figure 4: Mean ratings for subgroups

However, although the ratings in the two groups are similar, there is an important difference between them: the ratings of the low-performing group are much more tightly clustered around 50 than the high-performing group's ratings. The standard deviation for the low-performing group's mean ratings is only 3.74, and none of the likelihood ratings are different from one another (after BY-FDR).

In contrast, the high-performers seemed to draw important (and accurate) distinctions among some of the cases. The standard deviation for the high-performers' mean ratings is 9.63, and there are 21 pairwise significant differences in the high-performers' ratings (after BY-FDR).[1]

Interestingly, all of the significant differences involve a 11** case (principally 1111 and 1101) being judged much less likely than a non-11** case to be in the target category. That is, it seems that the high-performers were not learning the structure of the target category, but rather something about the *contrast* category structure: namely, that 11** cases (i.e., the prototypical cases for the contrast category) were highly likely to be in the contrast category, and so received much lower likelihood ratings for the target category. The ratings for the 1111 case are particularly interesting, since it never appeared in the learning phase. Moreover, according to the "true" distributions, 1111 was equally likely to be in either the target or contrast category. However, if one learned only the contrast prototype, then 1111 should receive a low rating. Thus, we argue that these ratings are better explained by the hypothesis that participants understood only some of the contrast category structure, than by the hypothesis that they correctly learned (some of) both category structures.

Conclusions

It is notoriously difficult to determine cognitive limits, since there are typically many different reasons why participants might have failed at some task. We thus must be careful about drawing any particularly strong conclusions from this one experiment. Nevertheless, by a variety of measures, it seems that participants did not learn much about the target (chain graph) category. Learning performance was not statistically distinguishable from chance responses, and the only significant differences in test phase ratings (for only one sub-group) seem to be due to an understanding of the *contrast* category structure, not the target category structure.

There are at least three obvious alternative explanations for the apparent failure of participants to learn the target category. First, participants might not have had sufficient experience with the categories. That is, perhaps performance would improve substantially given more observations. Pilot experimental results do not suggest a substantial increase in performance over time, but more investigation is needed.

Second, the two categories used in this experiment might be overly similar. If there were more contrast between the categories (i.e., if they were more separable), then people's performance and understanding might significantly increase. This concern is particularly salient given the significant dependence of target category learning on the structure of the contrast category (Goldstone, 1996).

Third, the measures used here might not have accurately revealed people's category learning. The "high performers" were identified using mean correct responses over the last 54 cases, so people who learned quite late in the sequence could have been excluded. Use of a fixed presentation order could enable us to determine more accurately the subset of individuals who actually learned the target (and contrast) categories. Alternately, in the test phase rating collection, we could ask about only crucial cases, rather than all cases.

Given these alternatives, the most definitive evidence that people are unable to learn categories with these complex statistical structures would be a series of experiments using categories that are progressively more difficult to learn (in theory). In particular, the categories would vary along the dimension of: complexity of a graphical model that perfectly represents the underlying probability distribution. By tracking the changes in participant performance, we could potentially determine something about the learnability of various classes of probability distributions. There are, of course, many studies testing progressively harder categories—perhaps most famously the canonical study of Shepard, Hovland, & Jenkins (1961)—but none of those studies vary the category complexity along the dimension proposed here. We are currently developing an appropriate series of categories (from the graphical models perspective), and we hope that experiments using those categories will

[1] Ordered by BY-FDR corrected *p*-value, the significantly different pairs were: *p* < .01: {0000, 0011, 0110} vs. 1111, 0000 vs. 1101; *p* < .02: {1010, 1011, 1000, 0001, 0010} vs. 1111, {0011, 1000} vs. 1101; *p* < .05: 0111 vs. 1111, {0001, 1001, 1011, 1010, 0010, 0101, 0110, 0111} vs. 1101, 1000 vs. 1100.

help us to understand better the nature of the limits on the category structures that can be learned.

Acknowledgments

D. Danks was partially supported by grants from the Office of Naval Research, and the National Aeronautics and Space Administration.

References

Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology, 37*, 372-400.

Ashby, F. G., & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological Review, 93*, 154-179.

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics, 29*, 1165-1188.

Danks, D. (2004). *Psychological theories of categorization as probabilistic models* (Technical Report No. CMU-PHIL-157).

Danks, D. (in press). Theory unification and graphical models in human categorization. In A. Gopnik & L. E. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation*. Oxford: Oxford University Press.

Goldstone, R. L. (1996). Isolated and interrelated concepts. *Memory & Cognition, 24*, 608-628.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review, 111*, 3-32.

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology, 51*, 334-384.

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

Lagnado, D. A., & Sloman, S. A. (2004). The advantage of timely intervention. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 30*, 856-876.

Lauritzen, S. L. (1996). *Graphical models*. Oxford: Clarendon Press.

Lauritzen, S. L., & Richardson, T. S. (2002). Chain graph models and their causal interpretation. *Journal of the Royal Statistical Society, Series B, 64*, 321-361.

Lauritzen, S. L., & Wermuth, N. (1989). Graphical models for association between variables, some of which are qualitative and some quantitative. *Annals of Statistics, 17*, 31-57.

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush & E. Galanter (Eds.), *Handbook of mathematical psychology*. New York: Wiley.

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

Minda, J. P., & Smith, J. D. (2001). Prototypes in category learning: The effects of category size, category structure, and stimulus complexity. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 27*, 775-799.

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, & Cognition, 28*, 275-292.

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

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

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Francisco: Morgan Kaufmann Publishers.

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge: Cambridge University Press.

Rehder, B. (2003a). A causal-model theory of conceptual representation and categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 1141-1159.

Rehder, B. (2003b). Categorization as causal reasoning. *Cognitive Science, 27*, 709-748.

Rehder, B., & Hastie, R. (2004). Category coherence and category-based property induction. *Cognition, 91*, 113-153.

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

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

Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, prediction, and search*. Berlin: Springer-Verlag.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Structure learning in human causal induction. In T. Leen, T. Deitterich & V. Tresp (Eds.), *Advances in neural information processing systems 13* (pp. 59-65). Cambridge, MA: The MIT Press.

Waldmann, M. R., & Martignon, L. (1998). A Bayesian network model of causal learning. In M. A. Gernsbacher & S. J. Derry (Eds.), *Proceedings of the 20th annual conference of the cognitive science society*. Mahwah, NJ: Lawrence Erlbaum.

Zhu, H., & Danks, D. (in prep). Effects of presentation and goal in category learning: Carnegie Mellon University.