Using Ideal Observers in Higher-order Human Category Learning

Aniket Kittur (nkittur@ucla.edu)
Keith J. Holyoak (holyoak@lifesci.ucla.edu)
Department of Psychology, University of California, Los Angeles CA 90095

John E. Hummel (jehummel@cyrus.psych.uiuc.edu)
Department of Psychology, University of Illinois at Urbana-Champagne, Urbana IL 61820

Abstract

Ideal observer models have proven useful in investigating assumptions about human information processing in a variety of perceptual tasks. However, these models have not been applied in the area of higher-order category learning. We describe a simple Bayesian ideal observer and apply it to empirical data on category learning. We describe an experiment in which we found that acquisition of family resemblance categories was drastically impaired if the categories were defined by relations between features rather than by the features themselves. An ideal observer was used to test whether this effect could be accounted for by inherent information differences between the conditions. A comparison of participants' performance to the model found a significant difference in efficiency of learning even after accounting for information differences between conditions. This analysis illustrates how ideal observer methods can provide useful tools for analyzing higher-order category learning.

Keywords: categorization, category learning, relations, features, ideal observer

Introduction

An ideal observer model makes optimal use of a set of given information in performing a task. Such models have traditionally proven useful in investigating human information use in various perceptual tasks by providing an upper bound or benchmark by which to measure performance. If a human can perform at the same level as (or better than) the ideal model, then we know that the human is making use of all of the available information in the situation (or, in the case of humans outperforming the ideal, more information than was available to the model). If humans underperform the ideal model, the difference can often highlight specific constraints that limit human information processing. The degree to which human performance approaches that of an ideal observer can provide a measure of processing efficiency.

Ideal observers have most commonly been applied to understanding human low-level visual tasks involving detection and discrimination (see Geisler, 2003), though they have also been applied to tasks such as reading (Legge, Klitz, & Tjan, 1997), object recognition (Liu, Knill, & Kersten, 1995), and reaching (Trommershäuser, Gepshtein, Maloney, Landy, & Banks, 2004). However, few studies have applied ideal observer methods to higher-order cognitive tasks, at least in part because of the difficulty of specifying exactly what is ideal. Instead, most studies of human category learning compare conditions against each other and assume that differences in performance capture theoretically-central differences between conditions. However, there may be differences between conditions that are not relevant to the variable being measured (e.g., noise). Ideal observers can provide a theoretical upper bound on human performance (given a set of assumptions), and can be used to control for some of these extraneous variables.

In this paper we describe a simple method for creating an ideal observer model that takes as input features, and relations between features, of the sort commonly used in category learning studies with artificial stimuli. The ideal observer assumes that the experimenter (but not necessarily the learner) has full knowledge of the generating model used to construct stimuli based on these features and relations. This assumption is typically met in experimental category learning paradigms in which artificial stimuli are used. We will first describe the model and then apply it by simulating performance in an actual category learning experiment.

The Model

The model uses a Bayesian framework to assign stimuli to categories and to learn from labeled feedback. We use a version of a naïve Bayesian classifier, one of the simplest probabilistic classifiers, which is optimal when all input features are independent (and can even be optimal in certain less restricted circumstances; see Domingos & Pazzani, 1997). The naïve Bayesian classifier makes the assumption that all features of a given category are generated independently, that is:

$$p(C, F_1, ..., F_n) = p(C)\, p(F_1 \mid C) ... p(F_n \mid C) \qquad (1)$$

for class variable C (which represents all possible categories) and feature variables 1 through *n*. Applying Bayes rule results in the following equation:

$$p(C \mid F_1, ..., F_n) = \frac{p(C)\prod_{i=1}^{n} p(F_i \mid C)}{\sum_C [\, p(C)\prod_{i=1}^{n} p(F_i \mid C)]} \qquad (2)$$

The denominator in Equation 2 is a normalization constant that is identical for all categories and thus often ignored for simplicity (though implemented in the model). With two equally-probable categories (as is most common in category learning paradigms) $p(C)$ is also constant (.5), and thus the main determinant of classification is $p(F_i|C)$. This probability is calculated in the following manner:

$$p(Fi \mid C) = \frac{n_{F_i|C} + \alpha_{F_i}}{n_C + \alpha_C} \quad (3)$$

where $n_{F_i|C}$ is the number of items with feature F_i in category C, n_c is the total number of items in category C, and α_{F_i} and α_C are uniform priors[1]. Equation 3 can be interpreted as updating a uniform prior with new information, with the prior eventually overwhelmed as more features are observed.

The classifier can be extended to reflect underlying dependencies between features that are not independently generated. This refinement can often be useful in categorization studies when one feature constrains the values of other features. When these dependencies are known, they can be incorporated into the model by retaining the relevant conditional probabilities. For example, Equation 4 is a toy model with two features in which feature 2 is dependent on feature 1 (the normalization constant is left out for simplicity):

$$p(C \mid F_1, F_2) = p(C)\,p(F_1 \mid C)\,p(F_2 \mid C, F_1) \quad (4)$$

Equation 3 can also be extended for features that are dependent on other features, becoming:

$$p(F_i \mid C, F_j) = \frac{n_{F_i|C,F_j} + \alpha_{F_i}}{n_C + \alpha_C} \quad (5)$$

where F_i is dependent on feature F_j.

The algorithm for category learning operates as follows. First, a new example is presented to the model without category label information. The probability of its being in each category is calculated based on previously-observed labeled examples, and the resulting probabilities are used to assign a predicted category to the example[2]. The typical way to classify a new example is to choose the category with the maximum probability of generating the example (Geisler, 2003). Once category assignment is complete the example is placed into the observed set along with its category label. This step simulates the effect of feedback, with the new example now affecting future classification judgments. Order of presentation is important: like the participants, the model's predictions can only be based on previously seen exemplars.

There are many ways to compare the model's performance with humans. One possibility is to use a metric based on the number of correct and incorrect trials, which is perhaps the closest analog to how ideal observers have been used in recent studies (Geisler, 2003). However, many category learning paradigms focus on differences in learning rates, with a common metric being the number of trials needed to reach a certain performance criterion. Viewed in terms of statistical sampling (e.g., the number of samples needed to learn a certain distribution), this metric provides a natural comparison of human and model learning. Specifically, we can define *sampling efficiency* as the ratio of the number of trials the model needed to learn to criterion to the number of trials a human needed (see Scholkopf & Smola, 2002; Stankiewitz, 2003):

$$\frac{ttc_{\text{mod}}}{ttc_{par}} \quad (6)$$

where ttc_{mod} is the trials to criterion needed for the model and ttc_{par} the trials needed for the participant. The closer human performance comes to the ideal, the higher the efficiency.

In many ways the ideal observer described here is similar to Anderson's (1991) "rational" theory of categorization. However, Anderson focuses on determining the optimal categorization given a general environment in pursuit of a descriptive theory of human categorization. In contrast, our ideal observer simply aims to be normative in a specific environment for which the structure and generating model is known, and to provide a benchmark or upper bound on human performance. Thus many of the goals and assumptions of Anderson's model are very different from the ideal observer described here. For example, since we know and capture the dependencies between features, we do not make the simplifying assumption that all features are conditionally independent. Dropping this assumption is necessary in order to maintain optimality for the types of generating models commonly used in higher-order category learning, where features are often constrained by the values of other relations or features. Also, instead of predicting an unseen feature (such as the category label) through chained inference, we focus on the simpler task of predicting a category class given a set of features. This simpler goal allows us to avoid using weighted category averages and only requires computation of the maximum likelihood category. Finally, we avoid the need for an empirically-determined variable governing the probability of creating a new category (Anderson's "coupling probability").

We will now apply this ideal observer in order to model learning in a categorization experiment in which the different conditions may have different types and amounts of information associated with them.

The Experiment

A fundamental shift in the understanding of categorization resulted from the "family resemblance" view of categories, which argued that many categories have a graded structure based on shared features (Medin & Schaffer, 1978; Rosch, 1976; Rosch & Mervis, 1975; Wittgenstein, 1953). The family resemblance view has had great success accounting for peoples' learning and generalization of categories that can be represented as simple lists of features. Such

[1] More specifically, the α_{Fi}s describe the parameters of a Dirichlet prior in which all values are set to 1, with their sum being α_c.

[2] One issue with the algorithm is how to get it started. Although there are a number of justifiable methods, here we start the model with the smallest number of examples for which there is one example for each category.

categories can be learned implicitly and automatically, with feature-category associations not necessarily available to conscious verbalization (Ashby, Maddox, & Bohil, 2002; Ashby & Waldron, 1999).

However, much of human conceptual knowledge is composed of categories that cannot be represented as simple features (Barsalou, 1983; Keil, 1989; Murphy & Medin, 1985; Rips, 1989; Ross & Spalding, 1994). Rather, many concepts are based on the *relationships* between things rather than the literal features of the things themselves. For example, a *barrier* is a relational concept that can be as concrete as a wall or moat or as abstract as a lack of money or the color of one's skin. Relational concepts abound in everyday life, with examples including social understanding (a love triangle), law (breach of contract), religion (atonement for sins), science (conservation of energy), as well as basic perception (recognizing arrangements of objects as scenes) (e.g., Gentner & Kurtz, 2005, Holyoak & Thagard, 1995).

Although relational concepts are fundamental to human intelligence, our understanding of how we learn them is poor compared to our understanding of feature-based categories. A reasonable and parsimonious hypothesis is that relational categories act just like feature-based categories with the features replaced by relations—that is, concept learning may be a single unified process that can take either features or relations as input. This view predicts that relational categories should show the same kind of family resemblance structure evidenced by feature-based categories, thus generalizing what we have learned about category learning from feature-based to relational categories.

However, there is evidence that relations and features may be psychologically distinct. For example, Medin, Goldstone, and Gentner (1990, 1993) demonstrated strong empirical differences between relational and feature-based similarity, suggesting that relations and features may rely on separate, competing processes for assessing similarity. Consistent with these findings, some researchers have argued that feature lists are fundamentally inadequate to represent relational concepts, and that such concepts must instead be mentally represented as relational structures such as "schemas" or "theories" (Gentner, 1983; Holland, Holyoak, Nibett, & Thagard, 1986; Hummel & Holyoak, 2003; Keil, 1989; Murphy & Medin, 1985). In such accounts, learning a relational category is more akin to inducing a schema than to learning a list of diagnostic features. Most accounts of schema induction assume that a shared, deterministic cohesive element is necessary to create the schema in the first place (Hummel & Holyoak, 2003; Kuehne et al., 2000).

We conducted an experiment to test whether relational and feature-based categories were learned in similar ways[3]. Specifically, we hypothesized that relational categories in which no single defining element existed—as is the case in family resemblance categories—would prove drastically more difficult to learn than feature-based categories with an identical family resemblance structure. Whereas learning family resemblance categories based on simple features may be done implicitly through tracking and averaging the features of the exemplars of each category, learning relational categories will be much more difficult because the same feature(s) may be associated with multiple categories, depending on the relations involved. To test this hypothesis we used a 2x2 between-subjects design, in which categories either had a single dimension perfectly predictive of category membership (deterministic) or had a family resemblance structure in which three out of four dimensions were characteristic of the category but no single dimension was perfectly predictive. Dimensions were defined either by individual feature values or by the relations between features. We predicted an interaction between category structure and type, such that the relational family resemblance condition would be much more difficult to learn than any of the other three conditions.

[3] The experiment described here is based on pilot data reported in Kittur, Hummel, and Holyoak (2004), which describes in more detail the methods used and additional measures collected.

Method

Subjects. 96 University of California, Los Angeles undergraduates participated for partial fulfillment of course requirements.

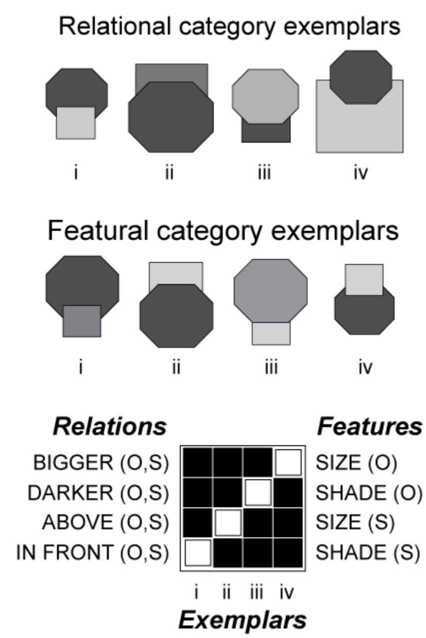

Figure 1. Examples of family resemblance categories. Deterministic relational categories were formed by removing one exemplar from each category. The table depicts the dimensions categories were defined on, as well as the value of each exemplar on the dimensions (filled = value 1, empty = value 2). For example, in relational category exemplar i, the octagon (O) is bigger, darker, above, and behind the square (S). Shown are only a small subset of all instantiations of the four exemplar types for a category.

Stimuli and Procedure. All stimuli were composed of an octagon and a square set in a fixed background resembling a computer chip. Either the relations between the two shapes (relational condition) or the individual features of each shape (feature-based condition) determined category membership of each exemplar. Relational categories were defined by whether the octagon was 1) larger, 2) darker, 3) vertically above, and 4) in front of the square (see Figure 1). Feature-based categories were defined by individual absolute feature values: 1) size of the octagon, 2) color of the octagon, 3) size of the square, and 4) color of the square.

Crossed with the feature-based and relational conditions was the structure of each category. In the family resemblance condition, each category member had three out of four dimensions consistent with its category and one inconsistent dimension. In the deterministic condition one dimension was perfectly diagnostic across all exemplars. This design yielded four conditions to which participants were randomly assigned: relational family resemblance or deterministic (R-FR or R-D) and feature-based family resemblance or deterministic (F-FR or F-D).

On each trial of the acquisition phase, a participant viewed one exemplar, categorized it as a “math” or “graphics” chip, and received accuracy feedback. Acquisition continued until the participant reached criterion (>88% correct for two consecutive blocks[4]).

Behavioral Results

The relational family resemblance condition proved much more difficult to learn than the other three conditions: 22% of participants in the relational family resemblance condition did not learn to criterion within 600 trials (no participants in any other conditions failed to learn). All results make the extremely conservative assumption that participants who failed to learn would have succeeded on trial 601.

The mean number of trials to criterion for each condition is shown in Figure 2a. There were main effects of both category type (relations vs. features, $F(1, 95) = 4.71$, $p = .032$, and category structure (family resemblance vs. deterministic, $F(1, 95) = 9.83$, $p = .002$; importantly, there was also a significant interaction of category type and structure, $F(1,95) = 6.14$, $p = .015$, due to extremely impaired acquisition when the category was defined by relations and had a family resemblance structure.

Ideal Observer Analysis

One explanation of these results is that relations and features are represented and processed differently in the brain, and that relational categories may not have access to the machinery that is used to learn feature-based family resemblance categories. However, another explanation could be that the selective impairment of the relational family resemblance condition is instead due to a difference in the amount of available diagnostic information. In other words, are people worse only because some conditions are inherently more difficult to learn due to lack of information?

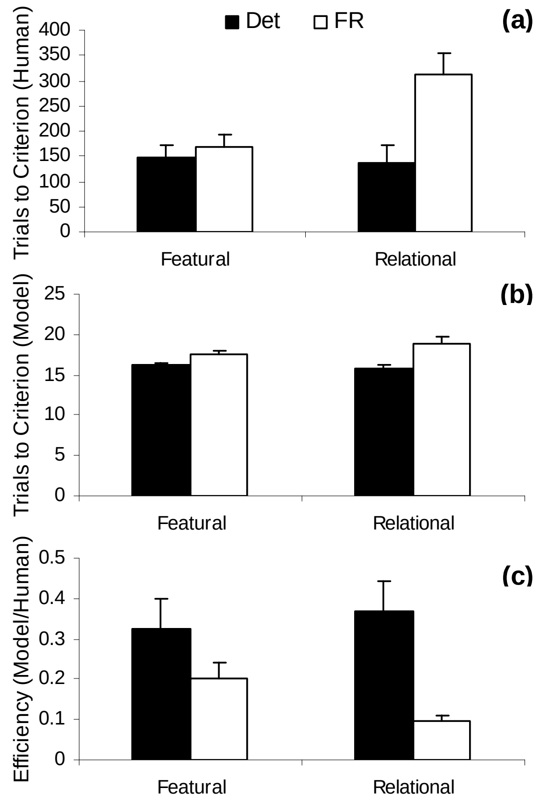

Figure 2. **(a)** Mean trials to criterion taken by participants to learn the categories. Det=Deterministic, FR=Family Resemblance. **(b)** Mean trials to criterion for the model to learn the categories given the same stimuli in the same order as each individual participant. **(c)** Efficiency of human performance compared to model performance measured by the ratio of the number of trials needed by the model to learn to criterion to the number of trials needed by human learners. (Note that efficiency is calculated on a per-subject basis, and so cannot be determined from panels (a) and (b) alone.)

To answer this question we adapted the ideal observer model described earlier to the current experimental task. The features and relations available to participants were coded as discrete values on separate dimensions and used as inputs to the model. For example, the model received as separate inputs the size of the square, the size of the octagon, and the relation of which was bigger. The same information was available to participants, who could use information about either the features or the relations on each trial. However, the relational and featural information on a dimension were not independent: in the example above, if the relation was “octagon bigger than square”, then knowing the size of the octagon provides information about the size of the square (which must be smaller; see Figure 3). To account for this dependency, relations were modeled as independent

[4] This criterion was chosen so that simple feature-tracking strategies (e.g., memorizing the associations of single features with categories) would lead to sub-criterion performance.

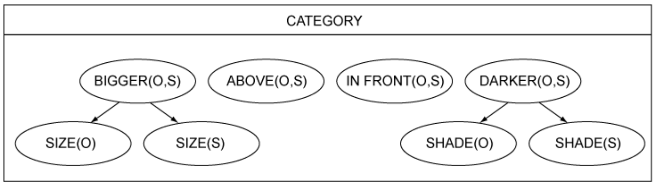

Figure 3. Relations and features involved in category generation. Arrows depict dependencies (i.e., constraints) between relations and features.

inputs, whereas feature inputs were conditional on their respective relation.[5]

For the ideal observer described here to be truly optimal, the generating model must meet certain assumptions. First, the distribution of category members must be sampled from independent multinomial distributions with Dirichlet-distributed parameters. This assumption holds true: category members were generated by sampling from independent multinomial distributions with an equal likelihood of each member appearing. Second, all dependencies that arise in generating feature values for each category member must be captured in the conditional probabilities (i.e., relations constraining features). This assumption is also valid: the dependencies shown in Figure 3 reflect how the exemplars were generated.

The ideal observer model was run on each participant's data, and the number of trials necessary to learn to criterion was measured[6] (see Figure 2b). We then computed each participant's efficiency according to Equation 6. These efficiencies are depicted in Figure 2c. Human performance as measured by statistical efficiency was much worse in the relational family resemblance condition than in any other condition. An ANOVA was performed on the efficiency measure following a log transformation to normalize the variances. The results demonstrated that the critical interaction was significant, $F(1, 95) = 3.93$, $p < .05$. Since efficiency takes into account differences in model as well as human performance, finding an interaction on this measure indicates that the human learning rates for these conditions were more different than would expected given the inherent difficulty of the conditions. That is, inherent informational differences between the conditions were insufficient to account for the disparities in human performance.

Discussion

The behavioral results revealed a clear impairment in acquisition for relational categories defined by a family resemblance structure, as compared to categories based on features, which are learned quickly whether they had family resemblance or deterministic structure. Relational categories with deterministic structure were learned as quickly as deterministic feature-based categories, suggesting that the effect is not merely due to the relational nature of the task. This interaction is inconsistent with the hypothesis that relational categories are learned in the same way as feature-based categories.

An ideal observer analysis was used to determine whether this impaired learning might be due to inherent informational differences between conditions. By comparing the efficiency of human performance to that of the ideal model, we were able to show that objective differences in difficulty between conditions did not account for the experimental data. Rather, it appears that relations and features are represented or processed differently in human category learning.

Identifying exactly how relations and features differ is an important subject for future research. One potentially useful approach is to determine what changes to the ideal observer could make it more closely match human data. For example, what happens when the model does not have perfect memory, or cannot perfectly update its prior? Or when its "working memory" is impaired so it cannot attend to all relations and features at once? Observing how the model degrades as additional constraints are added could provide valuable insights into human information processing.

Alternatively, it is possible that no processing-related changes in the ideal observer will capture the dissociation in the human data. Instead, it may be necessary to take into account the representational difference between simple features and relational predicates. It remains an open question how to incorporate structured predicates into a Bayesian framework; extant analyses of categorization using Bayesian inference treat relations as correlations or unstructured features rather than as structured predicates (e.g., Kemp et al., 2004). Indeed, one possible explanation of our results may be that the likelihood updating mechanism at work in featural categorization may not be used for relational categorization, resulting in impairment of relational family resemblance learning.

At first glance the present results appear counterintuitive: relational category learning is severely impaired if no elements are constant across all exemplars, yet people seem able to conceptualize family resemblance relational categories, such as Wittgenstein's (1953) classic example of "game". This paradox highlights the need for additional empirical studies. One approach to exploring this seeming inconsistency may be to examine prior knowledge and experience. While a single constant element may be necessary to learn novel relational categories, when prior knowledge and experience are brought into play this critical need may be reduced. It is possible that a coherent theory that explains a relational family resemblance structure might make learning easier (Rehder & Hastie, 2004; Rehder & Ross, 2001). In addition, repeated experience with the relevant relations may lead to low-level chunking of a stimulus, as in chess experts' memory for board positions. Thus both higher-order causal explanations and lower-order experience may facilitate relational learning.

[5] A natural question is: should the features be defining in the featural conditions, rather than the relations? No change is needed in the model because in the featural conditions the dimensions on which the features were considered dependent (relative size and shade) had the same relational value for both categories. Thus the features become effectively independent.

[6] A more statistically accurate phrasing of this would read: "the number of samples needed to learn the distribution to a certain degree of accuracy."

In summary, the dissociation between feature-based category learning, which is robust to family-resemblance structure, and relation-based category learning, which is not, suggests that current feature-based models of category learning may have limited applicability to relational categories. The difficulty for such models is not only that feature lists are inadequate to represent relations, but that the two kinds of categories are processed differently as well.

The present study also demonstrates how ideal observer methods can be applied in higher-order category learning. Here we used an ideal observer to provide an objective measure of the ease of learning in each condition. The model is easy to implement in category learning studies with discrete stimuli for which the generating model is known. We believe that the ideal-observer approach can have general applicability for studies of category learning in which different learning conditions may have different informational content.

Acknowledgments

Preparation of this paper was supported by NSF grant SES-0350920 to KH. Special thanks to Hongjing Lu, Zili Liu, and Barbara Spellman for helpful comments and discussion.

References

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review, 98*, 409-429.

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.

Ashby, F. G., & Waldron, E. M. (1999). On the nature of implicit categorization. *Psychonomic Bulletin & Review, 6*, 363-378.

Barsalou, L. W. (1983). Ad hoc categories. *Memory & Cognition, 11*, 211-227.

Domingos, P., & Pazzani, M. J. (1997). On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning, 29*, 103-130.

Geisler, W. S. (2003). Ideal Observer analysis. In L. Chalupa, and J. Werner (Eds.), *The visual neurosciences* (pp. 825-837). Boston: MIT Press.

Gentner, D., & Kurtz, K. J. (2005). Learning and using relational categories. In W. K. Ahn, Goldstone, R.L., Love, B.C., Markman, A.B., & Wolff, P.W. (Ed.), *Categorization inside and outside the lab*. Washington, D.C.: APA.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science, 7*, 155-170.

Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1986). *Induction: Processes of Inference, Learning, and Discovery*. Cambridge, MA, US: The MIT Press.

Holyoak, K. J., & Thagard, P. (1995). *Mental leaps: Analogy in creative thought*. Cambridge, MA: The MIT Press.

Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review, 110*, 220-264.

Keil, F. C. (1989). *Concepts, kinds, and cognitive development*. Cambridge: The MIT Press.

Kemp, C., Griffiths, T. L., & Tenenbaum, J. B. (2004). Discovering latent classes in relational data. MIT AI Memo 2004-019.

Kuehne, S., Forbus, K. D., Gentner, D., & Quinn, B. (2000). SEQL: Category learning as progressive abstraction using structure mapping. *Proceedings of the Twenty-second Annual Conference of the Cognitive Science Society*.

Legge, G. E., Klitz, T. S., & Tjan, B. S. (1997). Mr. Chips: An ideal-observer model of reading. *Psychological Review, 104*, 524-553.

Liu, Z., Knill, D. C., & Kersten, D. (1995). Object classification for human and ideal observers. *Vision Research, 35*, 549-568.

Medin, D. L., Goldstone, R. L., & Gentner, D. (1990). Similarity involving attributes and relations: Judgments of similarity and difference are not inverses. *Psychological Science, 1*, 64-69.

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

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

Murphy, G. L. (2002). *The big book of concepts*. Cambridge: The MIT Press

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

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

Rehder, B., & Ross, B. H. (2001). Abstract coherent categories. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 27*, 1261-1275.

Rips, L. J. (1989). Similarity, typicality, and categorization. In S. O. Vosniadou, Andrew (Ed.), *Similarity and analogical reasoning*. New York: Cambridge University Press.

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

Rosch, E., Simpson, C., & Miller, R. S. (1976). Structural bases of typicality effects. *Journal of Experimental Psychology: Human Perception & Performance, 2*, 491-502.

Ross, B. H., & Spalding, T. L. (1994). Concepts and categories. In R. J. Sternberg (Ed.), *Thinking and problem solving. Handbook of perception and cognition (2nd ed.)* (pp. 119-148). San Diego, CA: Academic Press, Inc.

Scholkopf, B., & Smola, A. J. (2002). *Learning with Kernels*: MIT Press.

Stankiewicz, B.J., Legge, G.E., Mansfield, J.S., & Schlicht, E.J. (in press). Lost in virtual space: Studies in human and ideal spatial navigation. *Journal of Experimental Psychology: Human Perception and Performance.*

Wittgenstein, L. (1953). *Philosophical investigations*. New York: Macmillan.