Conceptual Complexity and the Bias-Variance Tradeoff

Erica Briscoe (ebriscoe@rci.rutgers.edu)
Department of Psychology, Rutgers University - New Brunswick 152 Frelinghuysen Rd.
Piscataway, NJ 08854 USA

Jacob Feldman (jacob@ruccs.rutgers.edu)
Department of Psychology, Center for Cognitive Science, Rutgers University - New Brunswick, 152 Frelinghuysen Rd.
Piscataway, NJ 08854 USA

Abstract

In this paper we propose that the dichotomy between exemplar-based and prototype-based models of concept learning can be regarded as an instance of the tradeoff between complexity and data-fit, often referred to in the statistical learning literature as the *bias-variance tradeoff.* This continuum reflects differences in models' assumptions about the form of the concepts in their environments: models at one extreme, here exemplified by prototype models, assume a simple conceptual form, entailing high bias; models at the other extreme, exemplified by exemplar models, entertain more complex hypotheses, but tend to overfit the data, with a concomitant loss in generalization performance. To investigate human learners' place on this continuum, we had subjects learn concepts of varying levels of structural complexity. Concepts consisted of mixtures of Gaussian distributions, with the number of mixture components serving as the measure of complexity. We then fit subjects' responses to both a representative exemplar model and a representative prototype model. With moderately complex multimodal categories, the exemplar model generally fit subjects' performance better, due to the prototype models' overly narrow (high-bias) assumption of a unimodal concept. But with high-complexity concepts, the exemplar model's overly flexible (high-variance) assumptions made it *overfit* concepts relative to subjects, allowing it to outperform subjects on highly complex concepts. We conclude that neither strategy is uniformly optimal as a model of human performance.

Introduction

Modern research on categorization has centered around two general strategies for conceptual representation, termed *prototype theories* and *exemplar theories.* Prototype and exemplar theories make strikingly different assumptions about how learners use their past experience with category examples. Prototype theories (e.g. Smith & Minda, 1998; Nosofsky, 1987) assume that learners induce from observed category members a *central tendency*, called the prototype, used as a composite against which newly encountered items are compared. New items judged sufficiently similar to the prototype are judged to be members of the category. By contrast, in exemplar theories (e.g. Kruschke, 1992; Medin & Schaffer, 1978; Nosofsky, 1987), no central tendency is computed. Instead, the learner stores the attributes of observed examples along with category labels, called exemplars. Category decisions are then made by comparing newly observed items to these stored exemplars, and objects are classified via their similarity to these memorized examples.

Historically, prototype and exemplar strategies have been regarded as qualitatively different, and often competing, accounts of categorization. More recently, several authors (Ashby & Alfonso-Reese, 1995; Rosseel, 2002; Vanpaemel, Storms, & Ons, 2005) have pointed out how the two approaches may be seen as interrelated (see discussion below). In this paper, we suggest a new way in which prototype and exemplar models may be regarded as formally related, tapping a very basic continuum drawn from statistical learning literature, known as the *bias-variance tradeoff*. We then test this idea by teaching subjects concepts intended to favor one or the other end of the continuum—specifically, by varying the level of complexity of the set of objects to be learned. Our results suggest that human subjects tend to fall between the extremes represented by the two historically influential models, and more generally, that this is a fruitful new way of looking at categorization strategies.

The Bias-Variance Tradeoff

A key aspect of any learning model is the success with which it forms generalizations from training data, as measured by its accuracy in classifying further data from the same source. Somewhat counter-intuitively, this accuracy is not maximized by modeling training data as accurately as possible. An extremely close fit to training data tends to generalize poorly to future data, because such a fit inevitably entails fitting random (noise) aspects of the sample as well as regular, replicable trends, a phenomenon known as *overfitting*. On the other hand, a model that fits the training data too poorly will miss regular trends as well as noise, called *underfitting*. The critical variable modulating this phenomenon is the *complexity* of the hypotheses entertained by the learner (for example, the degree of the polynomial used in fitting a sequence of numeric data). More complex hypotheses (e.g. higher-degree polynomials) can, by definition, fit the training data more closely, while less complex ones may lack the flexibility to fit it well enough. The tradeoff is referred to as "bias vs. variance" because at one extreme (simple hypotheses), the model imposes a strong expectation or "bias" on the data, sacrificing fit, while at the other extreme (complex hypotheses), hypotheses are more flexible (i.e., exhibit greater variance), risking overfitting. The phenomenon can be seen schematically by plotting generalization accuracy as a function of model complexity; accuracy first rises to some optimum, but then declines (Fig. 1). This basic tradeoff arises in a wide variety of settings, as it seems to be fundamental to the vary nature of generalization of any data that involve an unknown mixture of regular and random elements (Dietterich, 2003; Hastie, Tibshirani, & Friedman, 2001).

The bias-variance tradeoff is a useful way to understand the

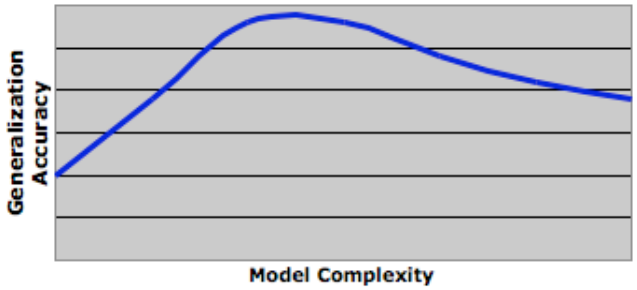

Figure 1: A schematic illustration of generalization accuracy as a function of model complexity, illustrating the bias-variance tradeoff. (*Note:* model complexity here is schematic and differs from the measure used in subsequent graphs.)

distinction between exemplar and prototype concept learning models. Along this continuum, prototype models exemplify the "bias" end of the spectrum. With a very simple, constrained schema for categories (the prototype), such models impose a strong bias on the observations, and thus will poorly fit data not conforming to this schema. Highly disjunctive or chaotic concepts, which are not well-described by a single central tendency, will generally be underfit by a prototype classifier. Data generated by a simple highly coherent source, by contrast, might be well fit by this strategy.

Exemplar models, on the other hand, exemplify the "variance" end of the spectrum. By using individual stored exemplars as classification standards, an exemplar model in effect entertains a very complex decision surface, with as many bends and wrinkles as there are learned exemplars. Such a system, by design, can learn arbitrarily complex collections of examples, as it imposes minimal expectations (bias) on the structure they are liable to exhibit. Such a strategy allows flexible learning but, of course, risks overfitting when the concepts are simple and the observed complexities are actually noise.

An ideal model of categorization would seek to balance bias and variance, finding an ideal level of hypothesis complexity, and thus optimizing its ability to generalize. But such an ideal balance cannot, unfortunately, be determined a priori, because it depends on the nature of the classes to be learned. Specifically, it depends on the proportion of regularity and noise in the source that generates the observations, i.e. on the *complexity* typical of the environment.

In this light it may be seen that prototype and exemplar models reflect different tacit assumptions about the nature of "natural" concepts in the human learner's environment. Prototype models will be effective in an environment in which natural classes tend to take the form of single, unimodal probability distributions in some feature space, and would be expected to perform poorly in more complex environments. Exemplar models, by contrast, tacitly presume a complex environment in which each object might, in principle, act as a "class of its own." Such a model represents an effective strategy in a complex environment, but in a simpler one may overfit spurious elements, literally committing to memory what are actually random events.

In the current paper we pursue a neutral, empirical approach to evaluating human generalization performance in a range of points on the bias-variance continuum by systematically manipulating the complexity of concepts presented to subjects. By confronting subjects with categories with varying levels of structural complexity, we may empirically examine how each of the two historically influential strategies fares as a model of human performance, ranging parametrically between the zone tacitly assumed by prototype models (simple concepts) to that tacitly assumed by exemplar models (complex concepts). We were interested in the effect of conceptual complexity on human performance, and also, more specifically, on the influence of complexity of the nature of subjects' learning strategies.

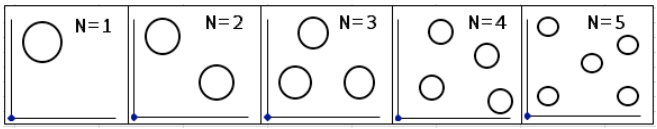

Figure 2: Schematic illustrations (contour plots) of the five concept types. Each concept consisted of a mixture of N Gaussian clouds in a two-feature space, with N ranging from 1 (left) to 5 (right).

Experiment: Manipulating Complexity

Accordingly, the strategy in our study was to confront subjects with concepts of varying levels of complexity, and then to fit a representative prototype model and a representative exemplar model to their classification responses. Our aim was also to adopt a simple, theoretically neutral measure of complexity. To this end we constructed concepts using two continuous features, with the positive examples drawn from a probability density function that was a mixture of N Gaussian sources, with N varying from 1 to 5 (Fig. 2). $N = 1$ concepts are simple, unimodal Gaussian clouds. At the other extreme, $N = 5$ concepts are highly heterogeneous mixtures with five distinct modes. The number N thus represents the complexity of the conceptual structure in a fairly transparent way.

A number of earlier studies (McKinley & Nosofsky, 1995; Smith & Minda, 1998) have varied the structure of concepts in attempts to probe subjects' learning strategies, and indeed concluded that the success of one strategy or the other depends on the "diffusion" or "differentiation" of the concepts on which performance was tested. Our study, we feel, takes a step forward by putting this intuition on a firmer footing, by varying conceptual structure in a more systematic way, and by tying the resulting variations in learning success to a fundamental spectrum of strategies in learning.

Subjects

Thirteen undergraduates at Rutgers University received class credit for participation.

Stimuli and Procedure

The objects observed by our subjects were parameterized by two dimensions loosely adapted from (Ashby & Maddox, 1990), who used semicircles with a spoke radiating from the center, with the two dimensions being the diameter of the circle and the orientation of the spoke. We embedded similar parameterized figures in depictions of flags flying from "ships,"

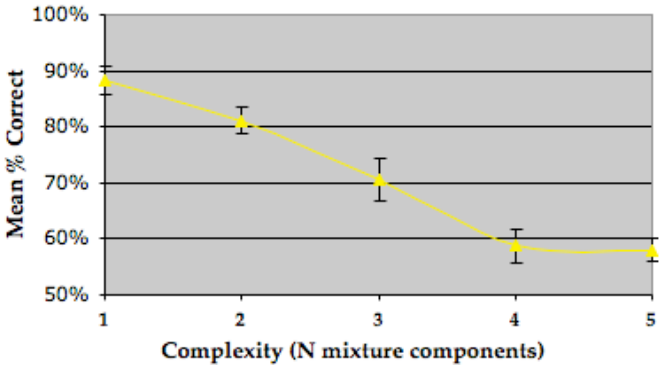

Figure 3: Subject performance as a function of conceptual complexity. Error bars indicated ± one standard error.

which the subjects were asked to classify as either hostile (pirate) or friendly (good guy) depending on the appearance of the flag. Each ship floated in from off-screen, with a flag containing a black rectangle and a white sword. The width of the rectangle (0 to 170 pixels) and the orientation of the sword (0-359°) served as the two quasi-continuous dimensions.

For each concept, the positive examples were generated from a probability density function in this two dimensional space. Each such distribution was a mixture of N bivariate circular Gaussians, with the number N of mixture components in serving as the manipulation of conceptual complexity (Fig. 2). Negative examples were generated from the corresponding inverse distribution. To ensure that subjects gave their primary attention to the positive set, whose structure we were manipulating, the total area of the positive set (i.e. the integral of the positive probability density) was held constant on all concepts at one fourth of the total. The means of the Gaussian components were separated using a criterion relative to their standard deviations, to prevent them from overlapping and thus obscuring our quantification of complexity. Half of the trials were drawn from the positive set and half from the negative, randomly intermixed within a concept, with a total of 150 items per concept. Each subject saw one concept from each of the five complexity levels, in random order, so all comparisons are within-subject.

Subjects were presented with instructions indicating that on each trial, a ship would move onto the screen whose flag he or she must look at in order to determine if the ship was a pirate or a "good guy." Feedback was provided after each classification, from which the subject gradually learned the correct classification. Each session consisted of 150 such trials, taking about ten minutes. Each subject ran one such session at each of the five complexity levels, in random order.

Results

The most conspicuous aspect of the results is the steady decline in subject performance as conceptual complexity increases (Fig. 3), mirroring similar findings with other types of stimuli and complexity measures (Feldman, 2000; Fass & Feldman, 2002). This trend, interesting in and of itself, reflects a simplicity bias that is apparently ubiquitous in human learning.

Details of Models

Our more detailed analyses concerned the relative performance of prototype and exemplar models in accounting for performance as complexity is varied. We first provide further details of the models used in our analysis.

Prototype Model

The multiplicative prototype model, first used by Nosofsky (Nosofsky, 1987) allows for psychological similarity to decrease exponentially with increasing distance. To compute similarity between a to-be-categorized item and a prototype, the values of the item and the prototype are compared along stimulus dimensions. The prototype is represented as the average of all exemplars seen from that category. Formally, the scaled psychological distance between the to-be-categorized item i and prototype P is given by

$$D_{iP} = \left(\sum_{m=1}^{d} w_m |x_{im} - P_m|^r \right)^{1/r} \quad (1)$$

The distance, D_{iP} , is most commonly computed using a simple Euclidean metric ($r = 2$), where x_{im} and P_m are the values of the to-be-categorized item and prototype on dimension m in d-dimensional space. A weighting variable for dimension m, represented as w_m, is used to account for the inequality of attention on each dimension. This variable allows for a magnification of the psychological space along more attended dimensions and shrinkage along less attended dimensions (Kruschke, 1992; Nosofsky, 1986).

Similarity is then measured as a monotonically decreasing function of the psychological distance between the point representation of the stimulus and the prototype given by

$$s_{iP} = e^{-cd_j} \quad (2)$$

where c is a freely estimated sensitivity parameter. Higher values of c "magnify" the psychological space, increasing the differentiation between the prototypes within the psychological space by increasing the steepness of the similarity gradient around them. In order to make a decision as to which category a particular item belongs, similarities are calculated between a to-be-categorized item and the prototype from each category. A guessing parameter, g, is used to represent the probability the observer chooses at random between the two categories, resulting in a probability $(1 - g)$ that the subject uses the similarities for his decision. If, for example, there are two categories, A and B, then the similarity to prototype A is divided by the sum of Prototype A and Prototype B similarity to generate the model's predicted probability of a Category A response to stimulus i:

$$p(R_A|s_i) = g/2 + (1 - g) \left(\frac{s_{iP_A}}{s_{iP_A} + s_{iP_B}} \right) \quad (3)$$

Exemplar Model

The generalized context model (GCM) (Nosofsky, 1986) assumes that the evidence for a particular category is found by summing the similarity of a presented object to all category exemplars stored in memory. Items are represented as points in multidimensional psychological space, with the similarity

between objects i and j measured as a decreasing function of their distance in that space,

$$s_{ij} = e^{-cd_{ij}} \tag{4}$$

(Shepard, 1987). Here, as in the prototype model, c is a sensitivity parameter that describes how similarity scales with distance. With large values of c, similarity decreases rapidly with distance; with smaller values of c, similarity decreases more slowly with distance. Distances are calculated similar to that in the prototype model, here summed from the to-be-categorized item and every exemplar, where x_{im} is the value of the to-be-categorized item i on dimension m and y_{jm} is the value of a category exemplar j on dimension m. As with the prototype model, w_m is used as an attentional weight granted dimension m.

$$d_{ij} = \left(\sum_{m=1}^{d} w_m |x_{im} - y_{jm}|^r \right)^{1/r} \tag{5}$$

To make a classification decision, similarities are calculated and summed between the to-be-categorized item and the exemplars from each category. If, for example, there are two categories, A and B, then summing across the category A exemplars and category B exemplars results in the total similarity of the item to category A members and category B members. For category A, C_A represents all exemplars in category A and C_B represents all the exemplars in category B. Using the similarity choice rule (Luce, 1963), the probability of category A response for stimulus i depends on the ratio of i's similarity to A to its total similarity to A and B,

$$p(R_A|s_i) = g/2 + (1-g) \left(\frac{\sum_{j \in C_A} s_{ij}}{\sum_{j \in C_A} s_{ij} + \sum_{j \in C_B} s_{ij}} \right) \tag{6}$$

where again g is a guessing parameter previously described.

Model Fit to Subject Data

To evaluate model performance, we first used parameter values that optimized each models' log likelihood fit to subjects' responses. Both models generally decrease in fit as complexity increases (Fig. 4). This simply reflects subjects' poorer performance with increasing complexity, meaning that their responses become progressively more random and thus more unpredictable as complexity increases. At very high complexity ($N = 4$ and 5), subject performance is very poor (see Fig. 3), so the two models begin to converge in their ability to fit what are now substantially random responses.

But at lower complexity levels, especially $N = 2$ and 3, the fit of the exemplar model is substantially better than that of the prototype model. By design, the prototype model determines similarity based on each exemplar's distance from the prototype, an average of all previously seen exemplars. For $N = 1$, this assumption closely matches the actual generating distribution, where there is one true center of the positive examples about which positive exemplars radiate outward and the probability of a positive exemplar becomes exponentially less likely as its distance from the center increases. The exemplar model also fits the data at this level of complexity well,

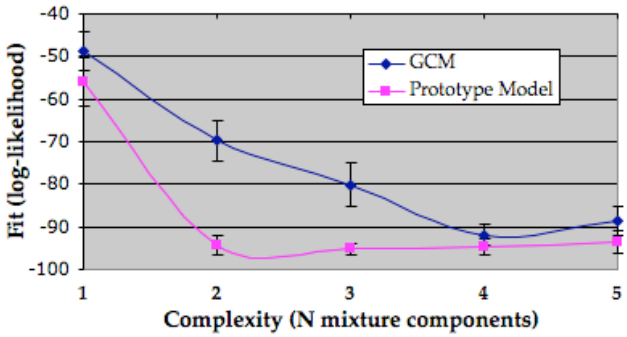

Figure 4: Model fit to subject concept learning data as a function of conceptual complexity.

with a summed log–likelihood of –48.7, resulting in a fit close to that of the prototype model, with a fit of –56.0.

At complexity $N = 2$, the category generative model is a mixture of two Gaussian clouds. Though subjects' performance worsens, they are still well above chance, averaging around 80 percent correct. Here the probability distribution in psychological space created from the prototype model, because it allows for only one central prototype, peaks in a region that falls between the two actual generative distributions. The prototype model cannot account as well for the data as can the exemplar model, which is able to represent the similarity space as a distribution with two modes.

The advantage provided to GCM at $N = 2$ reoccurs at $N = 3$, with GCM's fit at -80.2 and the prototype model's fit at -95.2 At complexity levels $N = 3$ and 4, however, exemplar and prototype performance begins to converge. At these high complexities, subject performance drops near chance. As their responses follow less of a discernible strategy, both the exemplar and prototype models are less able to approximate their responses.

Model Fit to Concepts

The analysis above involved optimizing each model to fit subjects' data as well as possible. While this method puts each model in the best possible light, it is undesirable in that it entails setting each models' parameters based on information that the model (and subject) could not, in principle, have access to, namely the performance of the ensemble of subjects on the task. As a second analysis, we refit each model to the data, this time setting the parameters in order to maximize the log likelihood of the *training examples* observed so far at each point in the experiment—that is, simply allowing each model to learn the concepts as well as possible. This method inevitably results in worse fit to the subject data, but illustrates more accurately how each model would perform were it "left to its own devices" to learn the training data as presented to the subject.

Fig. 6 shows the fit of each model to subject data using the settings optimized to the concept. These results are quite different from Fig. 4, and bring out in a very clear way the different places that the two models occupy along the bias-variance continuum. As before, both models generally decrease in fit as complexity increases, reflecting the generally poorer performance of subjects at high complexities. But the exem-

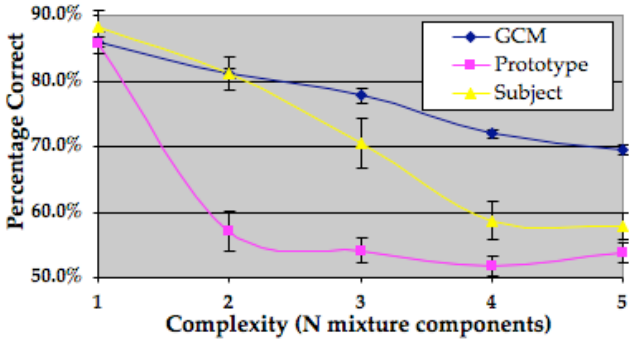

Figure 5: Model and subject performance as a function of conceptual complexity.

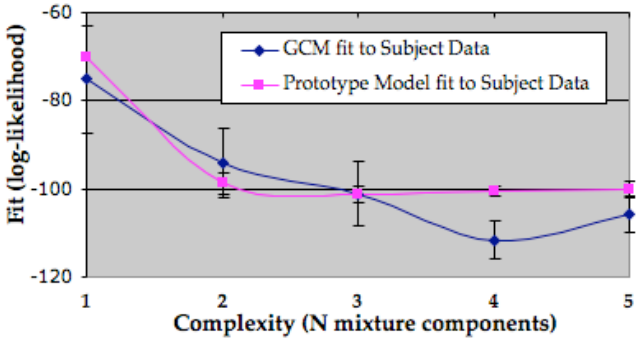

Figure 6: Fit of the two models to subject data using parameters fit to the training data.

plar model, with its inherent capacity to learn complex concepts, does not diminish in "learning" performance nearly as quickly as the prototype model (Fig. 5). As a result, the exemplar model is able to "out-learn" both the prototype model and subjects on concepts at the higher levels of complexity—and in this sense makes a poor model of subject performance at high levels of complexity. This failure is noticeable in Fig. 6, where now the exemplar model fits *worse* than the prototype model. But in an exactly complementary way, the prototype model does not learn moderately complex multimodal concepts ($N = 2$ and 3) as well as do subjects—its bias towards unimodal concepts is too strong and insufficiently variant. In other words, the results show that as complexity increases, the fits cross from one end of the bias-variance continuum to the other. Subjects' "default" bias-variance setting is somewhere in the middle between the two models—more complex than the unimodal assumption made by prototype models, but not *as* complex as the super-multimodal assumption embodied by GCM.

It is worth noting that GCM, like other exemplar models, includes a sensitivity parameter, c, that allows it to, in effect, slide along the bias-variance continuum. Recall that c controls the rate of exponential decay of probability as a function of dissimilarity from each exemplar. High values of c entail very narrow, "spiky" modes, and a more punctate decision surface, which in effect means a higher variance. Lower values of c entail smoother, broader, and less numerous modes, a simpler decision boundary, and in effect more prototype-like performance. Hence sensitivity provides the potential for GCM to fall at various points along the bias-variance continuum. Prototype models, while they also have a sensitivity parameter, consistently possess a unimodal assumption, with c controlling only how quickly the category typicality falls off with distance from the prototype. In this sense, prototype models exhibit an inflexibly high bias that is definitely not representative of subject performance in our study. Exemplar models are not as firmly tied to a single fixed point in the bias-variance continuum. Nevertheless, as the above results show, exemplar models do not necessarily automatically find the identical location along the continuum as do subjects when fit only to the training samples they have observed. Endowed with the capacity to overfit the data, they may well do so.

Integrating the Two Approaches

Many researchers have previously recognized the need to combine the benefits of both prototype formation and exemplar memorization, leading to a number of "hybrid" models in the literature containing elements of both. In addition, as mentioned, several previous concept-learning models have represented exemplar and prototype models in a unified continuum in ways different from our approach. Ashby and Alfonso-Reese (1995) emphasized the dichotomy between parametric (prototype) and non-parametric (exemplar) models, seeing both as varieties of density estimation. More recently, Vanpaemel et al. (2005) developed a categorization model incorporating both prototype and exemplar type elements by placing them at the extremes of their varying abstraction model. In their model, the number of items to which a new item can be compared may vary, allowing the model to form representations that lie between pure prototype and exemplar type structures. Rosseel (2002) explicitly adopts a mixture model, assuming a generalization space that, like our concepts, is a finite mixture of multivariate probability distributions. By allowing the number of mixture components to vary, this model can mimic both parametric (prototype) and nonparametric (exemplar) performance. While we have not yet fit this model to our data, we expect excellent performance, as its assumptions closely match the actual conditions under which our concepts were generated.

All these approaches are clearly related to ours, but we would argue that bias-variance analysis sheds a particularly clear light on the historical dichotomy in the psychological literature, by connecting the various proposals to a tradeoff that is fundamental to statistical generalization regardless of the specific form of the learning mechanism.

Conclusion

A number of conclusions can be drawn from our study, some empirical, some more conceptual.

First, prototype and exemplar approaches to categorization may fruitfully be regarded as points along a basic continuum of model complexity, reflecting a spectrum of possible assumptions about the complexity of concepts in the environment.

Second, conceptual complexity influences the degree to which each strategy accurately accounts for human perfor-

mance. At low but multi-modal levels of complexity ($N = 2, 3$), prototype models underfit, but at high levels of complexity ($N = 4, 5$), exemplar models overfit relative to human learners. Human learners apparently make a more intermediate assumption about conceptual complexity than does either classical strategy, and possibly modulate between the two extremes in accordance with the observations. Considering a range of complexity levels, neither strategy is consistently superior as a model of subjects.

Speaking methodologically, when carrying out studies of human learning, it is imperative to consider concepts with a range of complexity values. Our $N = 1$ concepts, which resemble many concepts studied in the literature, elicited very similar performance from prototype and exemplar models, which might shed some light on the decades of uncertainty and controversy about the relative merits of the two approaches. Clear differences between the two strategies only emerged at complexity $N = 2$ and above. A more systematic approach to varying complexity is necessary to paint a complete picture of human generalization under arbitrary conditions.

Acknowledgments

Supported by NSF SBE-0339062.

References

Ashby, F. G., & Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology*, *39*, 216–233.

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

Dietterich, T. G. (2003). Machine learning. In *Nature encyclopedia of cognitive science.* London: Macmillan.

Fass, D., & Feldman, J. (2002). Categorization under complexity: a unified MDL account of human learning of regular and irregular categories. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in neural information processing 15.* Cambridge: MIT Press.

Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature*, *407*, 630–633.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning: Data mining, inference, and predicition*. New York: Springer.

Kruschke, J. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22–44.

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

McKinley, S. C., & Nosofsky, R. M. (1995). Investigations of exemplar and decision bound models in large, ill-defined category structures. *J. Experimental Psychology: Human perception and performance*, *21*(1), 128–148.

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

Nosofsky, R. (1986). Attention, similarity, and the identificationcategorization relationship. *Journal of Experimental Psychology: General*, *115*, 38–57.

Nosofsky, R. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*, 87–108.

Rosseel, Y. (2002). Mixture models of categorization. *Journal of Mathematical Psychology*, *46*, 178–210.

Shepard, R. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*, 1317–1323.

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

Vanpaemel, W., Storms, G., & Ons, B. (2005). A varying abstraction model for categorization. In L. B. B. Bara & M. Bucciarelli (Eds.), *Proceedings of the 27th annual conference of the cognitive science society.* Hillsdale, NJ: Lawrence Erlbaum Associates.