Cognition 118 (2011) 2–16

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

Conceptual complexity and the bias/variance tradeoff

Erica Briscoe [a], Jacob Feldman [b,*]

[a] Aerospace, Transportation & Advanced Systems Laboratory, Georgia Tech Research Institute, United States
[b] Department of Psychology, Center for Cognitive Science, Rutgers University – New Brunswick, United States

ARTICLE INFO

Article history:
Received 27 July 2009
Revised 6 October 2010
Accepted 8 October 2010

Keywords:
Concept learning
Complexity
Bias/variance

ABSTRACT

In this paper we propose that the conventional dichotomy between exemplar-based and prototype-based models of concept learning is helpfully viewed as an instance of what is known in the statistical learning literature as the *bias/variance tradeoff*. The bias/variance tradeoff can be thought of as a sliding scale that modulates how closely any learning procedure adheres to its training data. At one end of the scale (high variance), models can entertain very complex hypotheses, allowing them to fit a wide variety of data very closely—but as a result can generalize poorly, a phenomenon called *overfitting*. At the other end of the scale (high bias), models make relatively simple and inflexible assumptions, and as a result may fit the data poorly, called *underfitting*. Exemplar and prototype models of category formation are at opposite ends of this scale: prototype models are highly biased, in that they assume a simple, standard conceptual form (the prototype), while exemplar models have very little bias but high variance, allowing them to fit virtually any combination of training data. We investigated human learners' position on this spectrum by confronting them with category structures at *variable levels* of intrinsic complexity, ranging from simple prototype-like categories to much more complex multimodal ones. The results show that human learners adopt an intermediate point on the bias/variance continuum, inconsistent with either of the poles occupied by most conventional approaches. We present a simple model that adjusts (*regularizes*) the complexity of its hypotheses in order to suit the training data, which fits the experimental data better than representative exemplar and prototype models.

1. Introduction

For about four decades, most research in human categorization has assumed that mental categories are "fuzzy" structures exhibiting varying degrees of similarity among category members (Posner, Goldsmith, & Welton, 1967; Posner & Keele, 1968; Rosch, 1978; Rosch & Mervis, 1975). More specifically, most contemporary models understand categories in terms of variations in typicality exhibited by category members (Love, Medin, & Gureckis, 2004; Ashby & Alfonso-Reese, 1995; Nosofsky, 1986). But researchers have been divided about how the distribution of typicality is acquired and represented, and exactly how new objects are evaluated.

The two dominant approaches, known as *prototype theories* and *exemplar theories*, make critically different assumptions about how people learn from 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, and use it as a composite against which to compare newly encountered items. New items judged sufficiently similar to the prototype are judged to be members of the category. Exactly how the prototype is computed, and what information it retains, differs from model to model, but all prototype models share this central process of abstraction. By contrast, in exemplar theories

* Corresponding author. Address: Department of Psychology, Center for Cognitive Science, Rutgers University – New Brunswick, 152 Frelinghuysen Rd., Piscataway, NJ 08854, United States. Fax: +1 732 445 6715.
E-mail address: jacob@ruccs.rutgers.edu (J. Feldman).

0010-0277/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.cognition.2010.10.004

(e.g. Kruschke, 1992; Medin & Schaffer, 1978; Nosofsky, 1987), no central tendency is computed. Instead, the learner stores the attributes of observed examples, called exemplars, along with their category labels. New objects are classified by placing them into the category to whose exemplars they bear the greatest similarity.

Prototype and exemplar strategies have traditionally been presented as qualitatively different and competing accounts of categorization, the former involving abstraction, the latter not (Smith & Medin, 1981, though cf. Hahn & Chater, 1998 for a more nuanced view). A number of authors have argued in various ways for a continuum connecting the two approaches (Ashby & Alfonso-Reese, 1995; Gentner & Medina, 1998; Rosseel, 2002, Vanpaemel, Storms, & Ons, 2005), or have proposed overtly hybrid models incorporating the benefits of both approaches (e.g. Anderson & Betz, 2001; Nosofsky, Palmeri, & McKinley, 1994; Nosofsky & Palmeri, 1997), as will be discussed more extensively below. But most commonly the two approaches are regarded as fundamentally disparate, even to the point of involving anatomically distinct brain systems (Ashby, Alfonso-Reese, Turken, & Waldron, 1998; Smith, Patalano, & Jonides, 1998).

But notwithstanding their neural implementation, prototype and exemplar strategies each represent solutions to the fundamental computational problem of category formation, and their similarities and differences can be most fully appreciated when they are considered in terms of the abstract learning procedures they embody—that is, in the spirit of Marr's "theory of the computation" (Marr, 1982). In this paper we suggest, as have others before us (Love et al., 2004; Jäkel, Schölkopf, & Wichmann, 2007), that prototype and exemplar strategies are helpfully seen as points along a very basic continuum of inference often discussed in the statistical learning literature, known as the *bias/variance tradeoff* (or sometimes the tradeoff between *data-fit* and *complexity*). As we argue in more detail below, this viewpoint not only adds a useful theoretical perspective, but also suggests a critical empirical test that does not seem to have been carried out before. Below, we present an experiment along these lines, the results of which show that neither prototype nor exemplar models are entirely consistent with human learning learners' position along the bias/variance continuum.

2. The bias/variance tradeoff

The bias/variance tradeoff relates to the *granularity* at which abstractions are made from training data—the coarseness with which examples are integrated to form generalizations. The key insight, first articulated in these terms by Geman, Bienenstock, and Doursat (1992) (see Hastie, Tibshirani, & Friedman, 2001 for an introduction), is that this scale profoundly influences the effectiveness of generalization in essentially *any* situation in which inductions must be made from data.

To see why, consider the computational problem faced by an agent attempting to form generalizations from training data. Naturally, such an agent seeks to make generalizations in such a way as to maximize the accuracy of future classifications from the same data source. Perhaps counter-intuitively, this accuracy is not maximized by learning the training data as precisely as possible. An extremely close fit to training data tends to generalize poorly to future data, because such a fit inevitably entails fitting random aspects of the sample (i.e., noise) as well as regular components. Any model that learns every quantitative detail of the training data—inevitably including many that will never be repeated—misses the broader regularities in the data. (A student who memorizes every last detail of the teacher's lecture inevitably retains details of the delivery at the expense of the broader ideas in the lesson.) Fitting training data too closely in this sense—fitting noise as well a real trends—is often referred to as *overfitting*, while fitting it too loosely—missing real trends as well as noise—is called *underfitting*. This basic tradeoff arises in a wide variety of settings, and seems to be fundamental to the vary nature of generalization. Every real data source involves a mixture of regular and stochastic elements, and effective generalization requires finding the right balance between them—so that the regular (repeatable) components may be learned and the noise disregarded. Of course, it is impossible to know a priori what is noise and what is not. So any learner must guess how broadly to generalize, which means adopting a point on the continuum between bias (a strong prior model) and variance (a weak prior model). A learner can paint with a broad brush or a fine quill; bias/variance is a sliding scale that determines the size of the brush.

The critical variable modulating bias and variance 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, or more generally the number of degrees of freedom (fittable parameters) in any model. 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 very well. The phenomenon can be seen schematically by plotting generalization accuracy as a function of model complexity (Fig. 1). Accuracy first rises to some optimum, but then declines as overfitting sets in. Exactly where along the abscissa the peak (optimal performance) lies depends on the nature of the patterns to be learned. 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 question then is how close a fit is desirable. There is no "right answer" a priori, because the optimal decision depends on how much of the observed data is actually due to stochastic processes as opposed to regular ones, which varies depending on the nature of the data source. Hence the optimal point on the continuum necessarily reflects an assumption about the nature of the environment—in the case of categorization, the statistical properties of the categories human learners are likely to encounter. For this reason, considering models of human categorization from this point of view helps shed light on the assumptions about the world they tacitly embody.

As mentioned above, it has occasionally been noted in the literature (Love et al., 2004; Jäkel et al., 2007) that

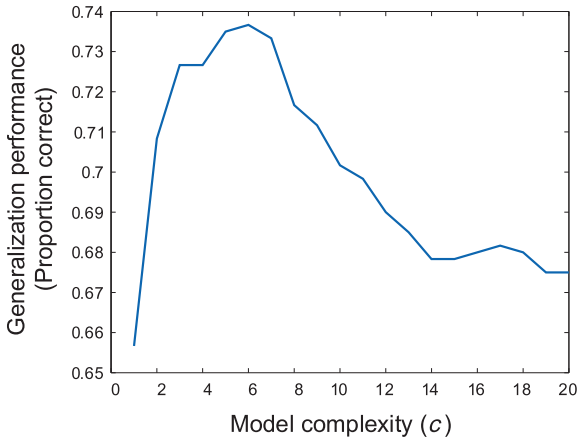

Fig. 1. Illustration of the tradeoff between bias and variance (complexity and data-fit). The plot shows generalization performance (proportion correct) of a model as a function of its complexity. As the model grows more complex (left to right) generalization improves until a peak, then declines. At higher complexities, the model "over-fits" the training data and performance suffers. The abscissa here is actually the parameter c from GCM (Nosofsky, 1986), which acts as a bias-modulating parameter (see text).

among traditional accounts of human categorization, prototype models exemplify the "bias" end of the spectrum, while exemplar models exemplify the "variance" end. Prototype models assume a very simple, constrained form for categories: the prototype, often realized as a unimodal distribution of feature values. Such a model imposes a strong bias on the observations, and thus will fit data not conforming to this schema relatively poorly. Highly disjunctive or unstructured 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, would be fit relatively well.

Exemplar models, on the other hand, exemplify the variance end of the spectrum, because in using individual stored exemplars for comparison judgments, an exemplar model may entertain a very complex decision surface, with each exemplar contributing a potentially distinct "wrinkle" to the decision surface. Whereas a prototype model assumes one central distribution, an exemplar model in principle assumes n of them, corresponding to the n stored exemplars. (Exactly how many peaks the learned density will actually have—possibly far less than n—depends on how narrow or broad each of these distributions is assumed to be, an important point to which we will return later.) Such a model, by design, can represent arbitrarily complex collections of examples, as it imposes minimal expectations (bias) on the structure they are liable to exhibit. In this sense what exemplar models do is less like *generalization* and more like *memorization* (Blair & Homa, 2003). Such a strategy allows flexible learning, but risks overfitting when the concepts are simple and the observed complexities are actually noise.

That prototype and exemplar models represent opposite ends of the bias/variance continuum will, we suspect, seem obvious to experienced readers of the statistical learning literature. Nevertheless, this perspective appears not to have been comprehensively articulated in the psychological literature on concept learning, and bears working out carefully. In particular, it suggests a natural class of empirical experiments, in which the complexity of training data is manipulated in order to evaluate human learners' position on this continuum, which does not appear to have been systematically attempted before. One key question is where human learners fit along this continuum, which (as explained below) can only be ascertained by evaluating their performance on training data of various levels of intrinsic complexity—an experiment that has not been carried out systematically before.

Hence in the experiments below we confront human subjects with conceptual structures whose intrinsic complexity we have systematically manipulated, making them in some cases more closely resemble the assumptions (bias) underlying prototype models, and in other cases the assumptions underlying exemplar models, and in other cases lie somewhere in between. This allows us to ask where human learners fall in the bias/variance continuum, which is, as we have argued, a very basic aspect of human categorization. More broadly, we see this as a beneficial way of looking at human categorization, because it allows us to step beyond the algorithms we identify, and shed light on the premises and assumptions that underlie them.

As mentioned, a number of prior approaches have intermixed prototype and exemplar strategies. Among these are some that parameterize the distinction between prototype and exemplar approaches in various ways, some related to our approach (e.g. Ashby & Alfonso-Reese, 1995; Rosseel, 2002). Some authors (Ashby & Alfonso-Reese, 1995; Griffiths, Sanborn, Canini, & Navarro, 2008; see also Shepard, 1957), model categorization explicitly as probability density estimation, in which case bias/variance issues inevitably arise, as they do in any statistical estimation problem. Vanpaemel et al. (2005) and Vanpaemel and Storms (2008) overtly vary the degree of abstraction imposed by the model, and thus substantially share our perspective, although the thrust of their model differs in most other respects. But none of these approaches explicitly conceptualizes the prototype-exemplar distinction in terms of bias and variance, and thus none of them allows us to pose this very basic question about human categorization: where do humans fit along this continuum? How much bias do human learners impose on the patterns they observe? More concretely, our perspective suggests an experimental approach in which we vary the intrinsic complexity of the training data presented to subjects, allowing us to evaluate human performance—and its fit by various models—as a function of the complexity of the training data. The performance of each learning model on training data at different complexity levels, and how closely this performance matches that of human subjects, constitutes a natural testbed in which to challenge existing models.

2.1. Varying conceptual complexity

In statistical estimation, the most common measure of model complexity is the number of fittable parameters in the model (for example, the number of coefficients in a regression model) because in general this modulates how

closely the final model can fit the training data. As sketched above, the variation in model complexity between prototype and exemplar models is best understood in terms of the number of peaks or modes they "expect" in a natural concept, i.e. 1 in the case of prototype models and up to n in the case of exemplar models. Hence in the context of this experiment the most natural way to vary the complexity of the training data is, similarly, in terms of the number of modes or components in the training data source. Of course this is also a modulation in the number of fittable parameters, because each stored exemplar in d-dimensional feature space requires d parameters to store it, resulting in nd total parameters to store n exemplars. A prototype model, on the other hand, needs only to store the mean (or other central tendency) of the training examples, a far smaller number which is independent of n.

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 as mentioned above such a perfect balance cannot be determined a priori, because it depends on the nature of the classes to be learned. In this sense, prototype and exemplar models reflect different tacit assumptions about the nature of concepts in the human learner's environment. Prototype models presume, and thus are in principle most effective in, an environment in which natural classes tend to take the form of single, unimodal probability distributions. By the same token, they would be expected to perform poorly in more complex environments. Exemplar models, by contrast, presume a complex environment in which each object could, in principle, act as a single category with itself as the only member (see Feldman, 2003). Such a model represents an effective strategy in a complex environment, but may overfit spurious elements, retaining in memory what are actually random events. Viewing prototype and exemplar models in this way begs for the test of a model that makes only *mid-level* assumptions about the complexity of the world, and thus, we also analyze the performance of a third model that represents a point on the bias/variance continuum between purely exemplar and prototype models.

In the following experiment, then, we aim to evaluate human concept learning at 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 popular strategies, as well as our "mid-level" model, performs as a model of human performance, ranging parametrically between the part of the continuum tacitly assumed by prototype models (simple concepts) to that tacitly assumed by exemplar models (complex concepts). Of interest is the effect of conceptual complexity on human performance, and also, more specifically, the influence of complexity of the nature of subjects' learning strategies.

2.2. Experimental approach

The strategy in this study was to confront subjects with concepts of varying levels of complexity, and then to fit a variety of models to their classification responses, including a representative prototype model, a representative exemplar model, and later a new model. Our aim was to manipulate the complexity of the concepts in a simple and theoretically neutral manner. To accomplish this, we defined each concept as a probability density function over two continuous features, with the positive examples drawn from a probability density function defined over the two features. To vary the complexity of the concept, we used *mixtures* with varying numbers of components. A mixture (see McLachlan & Basford, 1988) is a probability density function $p(X)$ that is the sum of some number of distinct components or sources $g_i(X)$, i.e.

$$p(X) = \sum_{i=1}^{K} w_i g(X). \quad (1)$$

Here K is the number of components, and the w_i are the mixing proportions, meaning the probabilities with which the respective components are drawn, which must sum to one ($\sum_{i=1}^{K} w_i = 1$). (In what follows we always set these equal at $1/K$.) Mixtures are a simple way of capturing the idea of a data source with a complex internal structure of sub-sources. For our purposes, they allow a simple way of modulating the complexity of the concept, namely by varying the number of components, K. A source with $K = 1$ consists of a single component such as a simple multivariate Gaussian. A source with a large value of K is heterogeneous, with multiple types intermixed. Thus the number K serves as a simple, theoretically neutral measure of conceptual complexity, which we vary in what follows in order to study the effect of complexity on subjects' learning strategies.

In the experiment below, each concept was a mixture of K Gaussian sources, with K ranging from 1 to 5. Fig. 2 depicts example distributions for the five levels of complexity as isoprobability contour plots. At the $K = 1$ level, each concept is a simple, unimodal Gaussian cloud (corresponding in the plots to a single circle, Fig. 2, left). At the other extreme, each $K = 5$ concept is a highly heterogeneous mixture with five distinct modes (corresponding to five circles, Fig. 2, right). The number K thus quantifies the complexity of the conceptual structure in a straightforward way. Fig. 3 illustrates the generating probability distribution for a $K = 3$ concept as well as a sample drawn from it.

A number of earlier studies have varied the structure of concepts in attempts to probe subjects' learning strategies. McKinley and Nosofsky (1995) also used concepts composed of mixtures of Gaussians, arguing that many naturally occurring categories are mixtures. Smith and Minda

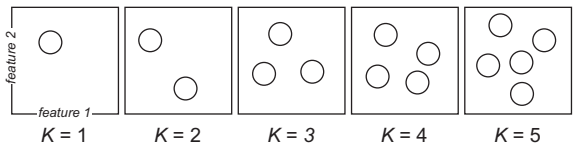

Fig. 2. Schematic illustrations (isoprobability contour plots) of the five concept types. Each concept consisted of a mixture of K Gaussian clouds in a two-dimensional feature space. K modulates the complexity of the resulting concept, ranging from 1 (left) to 5 (right).

(1998) also included a variety of conceptual forms in their studies, concluding (very much in the spirit of our analysis) that the relative success of various models seems to relate to the "diffusion" or "differentiation" of the concepts on which the models are tested. More recently Minda and Smith (2001) overtly varied the complexity of concepts (though in a different way than we do), again noting that different types of models seemed to excel with different types of concepts. Our study takes a step forward 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.

3. Experiment

3.1. Subjects

Thirteen undergraduate students at Rutgers University received class credit for participation. Subjects were naive to the purposes of the experiment.

3.2. Stimuli

The objects observed by our subjects were parameterized by two dimensions loosely adapted from Ashby and 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. Similarly parameterized figures were incorporated into depictions of flags flying from "ships," which the subjects were asked to classify as either hostile (pirate) or friendly (good guy) depending on the appearance of the flag (Fig. 4). Each ship floated in from off-screen, with a flag containing a black rectangle and a white sword. The width of the rectangle (0–170 pixels) and the orientation of the sword (0–359°) served as the two quasi-continuous dimensions.

3.3. Design

In the experiment, subjects were asked to learn a series of concepts, each consisting of a set of positive examples and a set of negative examples. For each concept, the subject was shown a series of unlabeled objects, both positives and negatives randomly interspersed and all in random order. The subject's task was to attempt to classify each object and indicate their guess with a keypress. Feedback (correct or incorrect) was provided after each response, allowing the subject to gradually learn the concept over the series of examples. The main dependent variable was the subject's classification accuracy as a function of the structure of the concept.

As sketched above, for each concept, positive examples were drawn from a probability density function defined over the two quasi-continuous features (flag width and sword orientation, see above). For each concept, the positive distribution was constructed from a mixture of K circular bivariate Gaussians, with K ranging from 1 to 5, so that K served as a modulation of the complexity of the concept (Figs. 2, 3). Technically, the positive distribution $p(f_1,f_2)$ was a mixture of circular Gaussians $p(f_1,f_2) \propto \sum_{i=1}^{K} \mathbf{N}(f_1,f_2)$, each with a random mean and equal standard deviations, resulting in a category structure with K distinct "modes." Fig. 3 shows an example with $K = 3$. The negative examples were drawn from the complementary density function $1 - p(f_1,f_2)$, so that the ensemble of examples (positive and negative) were uniformly distributed across the feature space. To ensure that subjects gave their primary attention to the positive set, which was what 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 quarter of the total. To prevent overlap and thus

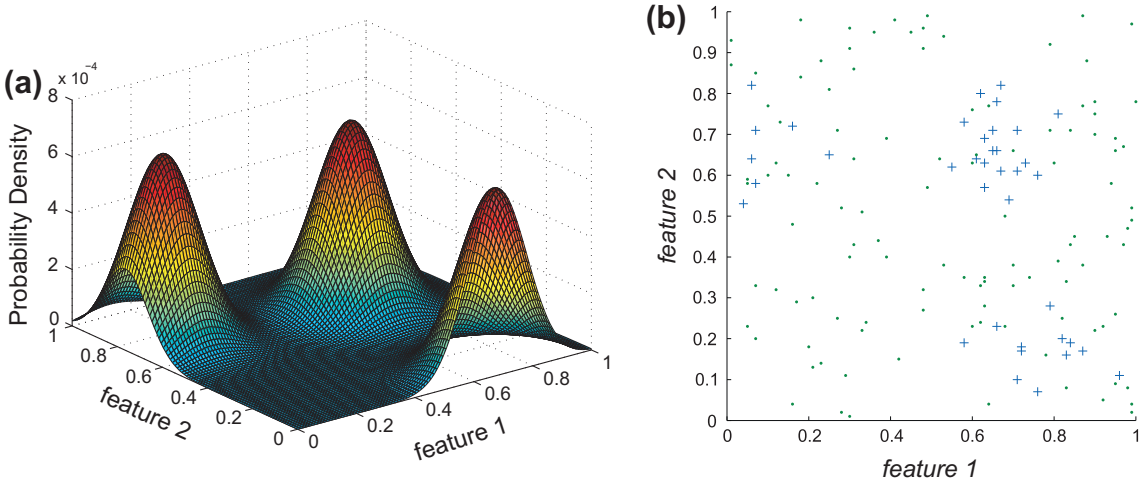

Fig. 3. (a) An example of a probability density function (positive only) with $K = 3$, and (b) sample training data drawn from it, now including both positive samples (indicated by +) drawn from the illustrated density $p(f_1,f_2)$ and negatives (indicated by ·) drawn from the complementary density $1 - p(f_1,f_2)$ (normalized). As in the experiments, the illustrated sample comprises 150 examples, about one fourth positive and the rest negative. Three clusters of positives are plainly visible, corresponding to the three modes in the density function.

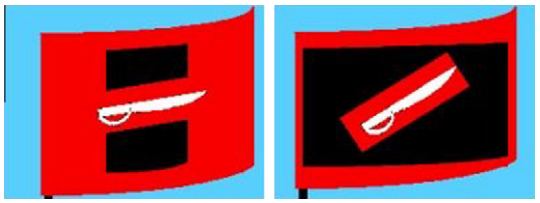

Fig. 4. Examples of the pirate flag stimuli shown to subjects.

the obscuring of the quantification of complexity, the means of the Gaussian components of the positive distribution were separated by at least five standard deviations. 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.

3.4. Procedure

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, with short breaks in between blocks.

3.5. Results and discussion

The most salient trend in the results is the steady decline in performance as conceptual complexity increases (Fig. 5), mirroring similar findings with other types of stimuli and complexity measures (Aitkin & Feldman, 2006; Fass & Feldman, 2002; Feldman, 2000; Pothos & Chater, 2002). Performance is generally good (nearly 90%) with single-component $K = 1$ concepts, but declines to nearly chance levels (50%) levels by $K = 5$. This trend reflects a simplicity bias that is apparently ubiquitous in human learning (Chater, 1997; Feldman, 2003): human learners perform progressively worse as conceptual complexity increases. As will be seen below, the performance of the various theoretical models tends to decrease with complexity as well, but not necessarily at the same rate or in the same way as human performance does.

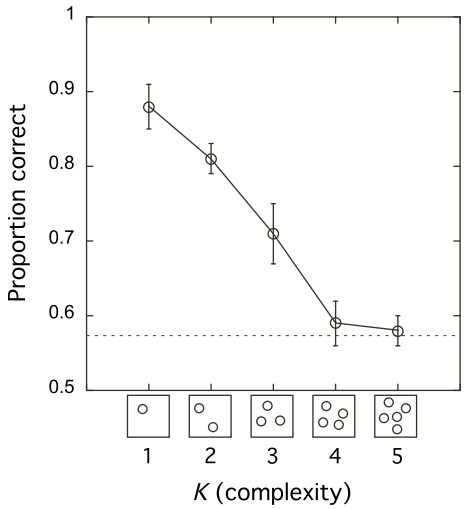

Fig. 5. Subject performance (proportion correct) as a function of conceptual complexity K. Error bars indicate ±1 s.e. Chance performance is 50%; performance below dotted line is not significantly different from chance at $p = .05$.

Beyond this general simplicity bias, our main interest is in the relative performance of various models in accounting for human performance as conceptual complexity varies. Because of our focus on variations in performance as a function of bias/variance, we chose fairly generic examples of prototype and exemplar models that are very similar to each other except for the primary difference in their number of modes they assume, i.e. in their bias. Before presenting fitting results we give details of the models.

3.6. Prototype model

The multiplicative prototype model proposed by 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 the average of all exemplars seen from a given category. Formally, the scaled psychological distance between the to-be-categorized item i and prototype P is given by

$$D_{iP} = \left(\sum_{m=1}^{d} w_m |x_{im} - P_m|^r \right)^{1/r}. \quad (2)$$

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

$$s_{iP} = e^{-cd_j}, \quad (3)$$

(Shepard, 1987) 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 cate-

gory. A guessing parameter g is used to represent the probability the observer chooses at random between the two categories, with a complementary probability $(1-g)$ that the subject uses the similarities to make a decision. Similarities are normalized to the summed similarity over categories. Putting all this together, the probability of response R_A to stimulus s_i is

$$p(R_A|s_i) = g/2 + (1-g)\left(\frac{s_{iP_A}}{s_{iP_A} + s_{iP_B}}\right). \quad (4)$$

3.7. 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. The similarity function is assumed to be the same for every exemplar. 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,

$$s_{ij} = e^{-cd_{ij}} \quad (5)$$

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.

$$d_{ij} = \left(\sum_{m=1}^{d} w_m |x_{im} - y_{jm}|^r\right)^{1/r} \quad (6)$$

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, E_A represents all exemplars in category A and E_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,

$$p(R_A|s_i) = g/2 + (1-g)\left(\frac{\sum_{j\in E_A} s_{ij}}{\sum_{j\in E_A} s_{ij} + \sum_{j\in E_B} s_{ij}}\right) \quad (7)$$

where again g is a guessing parameter previously described.

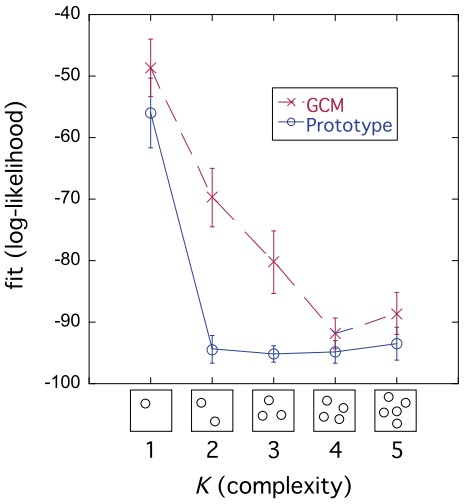

Fig. 6. Model fit to subject data as a function of conceptual complexity, using parameter values chosen to optimize fit to the subject data. Higher log-likelihood indicates better fit. Error bars indicate ±1 s.e.

4. Analysis: model performance

4.1. Model fits to subject data

To analyze the performance of each model (GCM and Prototype), we fit each model on each individual subjects' responses (maximizing log likelihood). This means that the fitted parameters are optimized to fit the ensemble of each subject's responses. (We first report the results aggregating over subjects; later, after introducing a new model for comparison, we give a detailed analysis of fits for individual subjects.) These parameters reflect the settings of the model that make the subjects' responses most likely, and thus cast each model in the best possible light.

As can be seen in Fig. 6, both models' fit to subject data generally decreased as complexity increased. This presumably reflects subjects' poorer performance with increasing complexity, meaning that their responses become progressively more random and thus more unpredictable as complexity increased. At very high complexity ($K = 4$ and 5), subject performance is very poor (see Fig. 5), so the two models begin to converge in their ability to fit what are now substantially random responses.

But at lower complexity levels, especially $K = 2$ 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 $K = 1$, 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, 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 $K = 2$, the category generative model is a mixture of two Gaussian clouds. Though subjects'

performance worsens, they are still well above chance, averaging around 80% 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. At this level of complexity, the exemplar model is able to fit the subject's data with a summed log-likelihood value of −69.7, substantially better than the prototype model's fit of −94.4.

The advantage provided to the exemplar model at complexity level two reoccurs at $K = 3$, with GCM's fit at −80.2 and the prototype model's fit at −95.2. At complexity levels $K = 4$ and 5, however, exemplar and prototype performance begins to converge. At these high complexities, subject performance drops near, but still slightly above, chance. As their responses follow less of a discernible strategy, both the exemplar and prototype models are less able to approximate their responses. At $K = 4$ and 5, the models' fits are similar at −94.8 and −93.5 for the prototype and −91.9 and −88.6 for the exemplar model.

4.2. 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, for our purposes 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. That is, these "optimal" parameters values are based on an analysis of the subjects' data after the experiment is complete, and attempt to bring the model into closest possible agreement with this corpus of responses. It is obviously not reasonable to suppose that subjects in the midst of the experiment could know what these optimal parameters would be, nor is it the subjects' primary goal to set any particular scientific model in the best possible light. Instead, the subject's goal is simply to learn the examples presented as well as possible. As we have argued, these same parameters materially influence the success of learning, in part because they modulate the degree of generalization. Hence from the subject's point of view it makes sense to optimize these parameters for *learning* instead of for model evaluation. So it is only reasonable to ask how each model performs when its parameters are set so as to maximally achieve this goal instead.

Hence 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.

We acknowledge that the first analysis (optimized to the responses) is the conventional approach, and we do not propose the second one (optimized to the training data) to replace it, but rather to complement it. The two analyses are designed to reveal different aspects of the situation. Parameter estimation with learning models is a subtle problem (Lamberts, 1997) and we present both analyses to give what we see as the most complete picture.

Fig. 7 shows the performance of the prototype and exemplar models, using parameters fit to the training data, compared to the performance of subjects (replotted from Fig. 5). Like subjects, both models decline in performance with increasing complexity. However, the three curves do not decline in the same way or at the same rate. At $K = 1$, when the concept is a single Gaussian cloud—like the concepts most often studied in the literature—performance by both models are similar, and both closely match that of subjects. But as complexity (K) increases, increasingly diverging from the unimodal assumption inherent in the prototype model, performance of the prototype model falls off much more rapidly than subjects. But performance of the exemplar model does not fall off rapidly *enough* to match subjects; by $K = 5$ its performance is far superior to that of human learners. (Indeed, at $K = 5$ the subjects are essentially at chance, but GCM is still effective.) With its inherent capacity to learn complex concepts, the exemplar model is able to learn complexities of the training data that humans cannot, and in this sense overfits relative to subjects.

This pattern can also be seen when we consider the fit of the models. Fig. 8 shows the fit of each model to subject data using parameters optimized to the training data. As before, both models generally decrease in fit as complexity increases, reflecting the generally poorer (more random) performance of subjects at high complexities. But at higher values of complexity ($K = 4$ and 5) the prototype model begins to fit subjects' data *better* than the exemplar model, reflecting the fact that at these complexities the exemplar

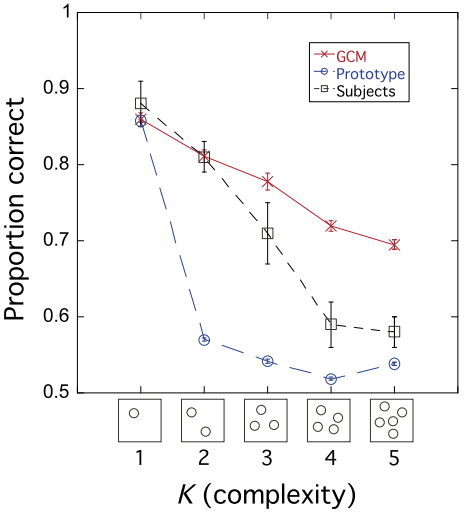

Fig. 7. Subject performance compared to the performance of the exemplar model (GCM) and the prototype model when their parameters are fit to the training data. Chance performance is 50%; performance below dotted line is not significantly different from chance at $p = .05$.

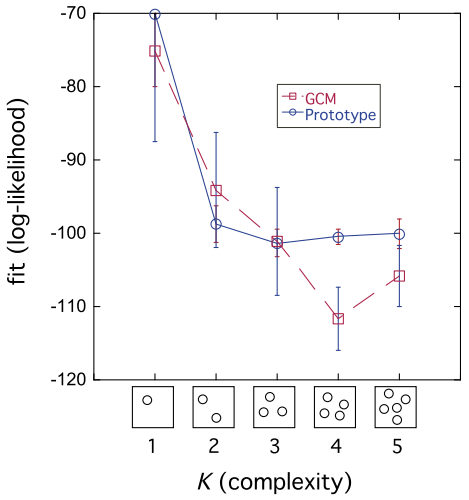

Fig. 8. Model fit to subject data as a function of conceptual complexity, using parameter values chosen to optimize fit to the training data. Higher log-likelihood indicates better fit. Error bars indicate ±1 s.e.

model can learn the concepts well, but subjects generally cannot. With relatively simple multimodal concepts ($K = 2$ and 3) prototype models cannot perform well (because they assume unimodality) but subjects can, showing that subjects' conceptual models are flexible enough to accommodate some multimodalty. But with more complex multimodal concepts ($K = 4$ and 5), subjects become essentially unable to learn the concepts, but exemplar models still can—leading to an increasingly poor fit between the exemplar model and human performance.

In other words, the results show that as complexity increases, human learners go from fitting one class of models better to fitting the other class of models better. Prototype models assume a unimodal concept, and their performance degrades rapidly as K increases—more so than do subjects. Exemplar models assume a multimodal concept, and begin to *out-perform* subjects as K increases. As we modulate the environment from the bias of prototype models (unimodality, $K = 1$) to the variance of exemplar models (multimodality, $K = 5$), subjects go from outperforming the prototype models to *being* outperformed by the exemplar model. On the spectrum from bias to variance, human subjects lie somewhere in the middle, intermediate 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.

As we have emphasized, our goal is not to pit exemplar and prototype approaches against each other, but rather to understand how they relate to each other, and in particular to understand how human learning relates to the spectrum along which they lie—bias/variance. The above analysis helps illuminate human concept formation by making it clear that it lies somewhere in between the two poles.

4.3. The sensitivity parameter (c) as a modulator of bias/variance

Fitting parameters to subject data allows them to "float" in a way that maximally accommodates the data (see Nosofsky & Johansen, 2000). But tuning the parameters this way begs the question of *why* they take the values they do—especially when one considers that alternative values (e.g. those that optimally fit the examples) would have improved performance from the subject's point of view (that is, would have allowed them to learn the training examples better). Why would subjects set their own parameters in an apparently suboptimal way? Our argument is that they do so in order to optimize their position along the bias/variance continuum—reducing overfitting and thus leading to more effective learning in the long run.

As mentioned, GCM (and many similar models) includes a sensitivity parameter, c, that modulates how steeply similarity falls off with distance from each example (Eq. (5)). The ability of exemplar models to fit learning data can very dramatically as a function of c, which has prompted a number of discussions of how this parameter is best interpreted (Smith, 2005; Verguts & Storms, 2004). We argue that c has the effect of modulating the learner's position along the bias/variance spectrum. High values of c entail very narrow, "spiky" modes, while lower values of entail smoother, broader, and less numerous modes. High values mean that exemplars can have more "local" effect, influencing generalization only nearby in the stimulus space, which results in decision boundaries that more closely resemble the arbitrarily complicated arrangement of exemplars. Low values mean that exemplars influence generalization more broadly, resulting in a simpler decision boundary with smoother boundaries and less abrupt transitions between positive and negative regions of the space.

In this sense high values of c make GCM more "exemplar-like," and low values more "prototype-like." In the language of statistical learning theory, we would say that c *regularizes* the model; that is, it modulates the complexity of hypotheses that it entertains. (Prototype models also have an analogous sensitivity parameter c (Eq. (3)). But because of the fixed assumption of a single prototype per class, this parameter does not have a similar effect, instead simply controlling how rapidly membership in the class degrades with distance from the induced prototype.) Above, we spoke of exemplar models as tacitly assuming many modes—in principle, as many as modes as the number of training examples. But in light of the variation in c, exemplar models can be better thought of as assuming a variable number of *effective* modes, with low values of c blurring regions of the space together to form a set of implicit modes that may be far less numerous than the training examples.

In other words, varying c allows GCM to fall at various points along the bias/variance continuum, with high values entailing high variance, and low values entailing high bias (Fig. 1). Exactly where GCM fits on the bias/variance continuum thus depends on exactly how c is set. While prototype models exhibit an inflexibly high bias that (our data show) does not account well for human performance, GCM is more flexible, and not as firmly tied to a single fixed point in the bias/variance continuum. When c is set very high, exemplar models are capable of grossly overfitting examples; the decision surface can grow as complex as the training examples are numerous. In this sense

exemplar models can (and our data show, often do) find a poor point along the bias/variance spectrum. This is an inevitable result of subjects' attempting to learn the training examples as well as possible (which, after all, is what they have been instructed to do). Without constraining c, the model has the capacity to overfit the data, and nothing preventing it from doing so (Schölkopf & Smola, 2002).

4.4. The locally regularized context model

There are, of course, many ways of modulating bias/variance, as the continuum is an abstraction not tied to any particular parameter or class of models. (There are an infinite number of ways of being biased, but only one way of being totally unbiased—which unfortunately leads to chronic overfitting and very poor learning.) In our setting, perhaps the most obvious way to vary bias/variance between prototype and exemplar models is simply by adopting a mixture model with a varying number of components, much as in Rosseel (2002), and fitting the number of components to the training data. (This would essentially make K itself the bias/variance parameter, but one must be careful to distinguish the *actual* number of mixture components in the generating distribution K from the *estimated* number of components derived from the sample $\widehat{K}$.) In practice, explicitly varying the number of mixture components and varying the sensitivity parameter c within an exemplar framework have extremely similar effects, because (as explained above) the value of c determines the number of effective modes in the decision surface. (A similar idea is implemented in density estimators with adaptive [i.e., optimized] kernel sizes; see e.g. Jones, McKay, & Hu, 1994.) Hence below we include in our analysis a comparison with GCM with c fit to the training examples, which allows it to optimize the number of effective components to fit the concept.

However a number of considerations argue for a slightly more flexible way of varying bias/variance, which we adopt in our own model. In general there is no reason to assume that the gradation of typicality would be globally uniform over the entirety of feature space, as implicitly assumed when one adopts a single level of c or a simple variation in the number of mixture components. If, for example, we adjust c to match the entire ensemble of training examples (as we do below), we may find that it overfits in one region of the space and underfits in another. We were inspired by an observation by Op de Beeck, Wagemans, and Vogels (2008), who in modifying ALCOVE (Kruschke, 1992) found that they could not achieve good fits to their subjects' data unless different nodes in the network were allowed to take on different values of c, meaning that each value of c would reign *locally* over one region of the space. In natural settings, it is entirely possible for the data-generating source (the "natural concept") to be complex and variegated in one region of the space and smooth and regular in another. In the randomly generated multimodal concepts used in our experiments, some modes happen to fall relatively near each other, creating broader positive peaks and thus simpler decision boundaries; while other modes are relatively distant from one another, creating narrower and more isolated peaks and thus more complex decision boundaries. Such differences in the typical rate of change of category membership are essentially what c modulates.

These considerations led us to formulate a new model, a simple modification of GCM, in which sensitivity c is modulated *locally*, or in other word, the model is*locally regularized*. We sought to achieve this in the simplest possible way. In the Locally Regularized Generalized Context Model (LRGCM), we simply *partition* the space into distinct regions, with boundaries estimated from the training data, and set c separately and independently for each region. The resulting model is similar to GCM in that it sums the distance to previously seen exemplars, but unlike GCM, it calculates similarity using parameters that are optimal for the particular region in which the to-be-categorized exemplar occurs. Again, we emphasize that our interest is in the principle of local regularization, not to the implementation, and we attach little importance to the fact that the mechanism here is exemplar-matching; the issue is how the decision surface is shaped and why, and whether the data support the claim that human category formation resembles it.

The model is simple. For each dimension, we used a standard kernel density estimator (Duda, Hart, & Stork, 2000) to estimate the probability density along that dimension. We then placed boundaries at local minima of this estimated density function, dividing the dimension into distinct bins (not necessarily of equal width). The bins in all dimensions are then crossed to create a set of rectangular cells in D dimensions, restricting the total number of cells to 9. (The exact number of cells in the grid is obviously *ad hoc*. For our purposes we simply needed a number sufficient to resolve all the modes we knew our own subjects would encounter. Again, our focus is on the principle, in this case local regularization, not the implementation, which is as simple as possible.) Each cell is then endowed with its own local sensitivity parameter c_j, whose value is optimized to the training data over the course of learning. This allows c_j to serve as a local learning parameter rather than a uniformly set global one.

Our approach has some similarity to other categorization models, several of which were mentioned above. SUSTAIN (Love et al., 2004) attempts to add new items to known clusters, but creates new clusters when needed, allowing it to fit the number of clusters to the number of modes in the training data, thus implicitly modulating bias. Motivated by findings in neuroscience, ITCOVE (Op de Beeck et al., 2008) allows local modifications in the granularity of abstraction, modulating its distance metric in response to training data. Verbeemen, Storms, and Verguts (2003) also use a similar approach, instead using K-means clustering to determine multiple prototypes, similarly allowing for multiple abstractions within the stimulus space. Similarly Aha and Goldstone (1992)'s GCM-ISW model allowed the parameters of an exemplar model to be determined by an exemplar's neighborhood in psychological space. However, unlike the Aha & Goldstone model, our approach varies only the parameter c, the sensitivity, within each partition in the stimilus space, while the attentional weighting w and the guessing parameter g are constant over the space. This flexibility allows for

prototype-like abstraction within each partition, while exhibiting a multimodal structure structure similar to, but less complex then, a straight exemplar model. The main difference between these models and ours is in not in the details but in the motivation, with our modification of GCM being driven by the aim of appropriately regularizing hypothesis complexity, and in particular local hypothesis complexity, in order to accomodate the learning of heterogenous concepts such as those encountered by our subjects.

We emphasize that the LRGCM is not intended as a full-featured categorization model, and is not suitable for modeling the many categorization phenomena encompassed by more complex models. We propose it simply as a concrete proof-of-concept for its main novel feature, local regularization. The idea is simply to show that the performance of a standard model can be substantially improved by adding local modulation of bias/variance, as demonstration that this aspect is an important component of a more complete model of learning. Indeed, local regularization is an abstract computational idea that could easily be incorporated into more elaborate processing models. It also has a number of theoretical benefits that we discuss below (see Section 5.2).

Next, we set out to compare performance of the three models: GCM Prototype, and our LRGCM, by fitting them each subject data. (All fits are to individual subjects, though we sometimes report aggregates of the individual fits.) Because LRGCM has a variable number of parameters, we needed a flexible way of balancing fit to the data with the number of parameters. A standard approach to this problem is the Akaike information criterion (AIC; see Akaike, 1974), which provides a measure of fit that compensates for extra parameters according to the following equation

$$\mathrm{AIC} = -2ln\zeta + 2\rho \tag{8}$$

where ζ denotes the likelihood of the model, and ρ the number of free parameters it has. In our implementations, both GCM and the prototype model have three fittable parameters (attentional weight, the guessing parameter, and sensitivity), while our model uses a variable number that depends on how many regions the space is partitioned into (at most 9, making a total of 11 parameters). The AIC is intended to compensate for such differences in dimensionality in model comparisons, providing a level playing field in which to compare models with unequal numbers of free parameters.[1]

Fig. 9 plots AIC for each of the three models, with parameters optimized to fit the subject data. (Lower values of AIC indicate better fit.) The plot shows that our LRGCM consistently fits subjects' data better than the other two models, even when the difference in the number of parameters is taken into account.

[1] There is a vigorous debate in the literature over the relative merits of the AIC vs. the Bayesian information criterion (BIC) (see Burnham & Anderson, 2004). Without entering into technical details, we note that BIC has often been criticized for depending on the assumption that the "true" model lies within the model class under consideration, an assumption we cannot regard as reasonable in our context, which certainly includes an open-ended set of potentially superior models. Hence here we adopt the AIC which is, in any case, the more conventional choice.

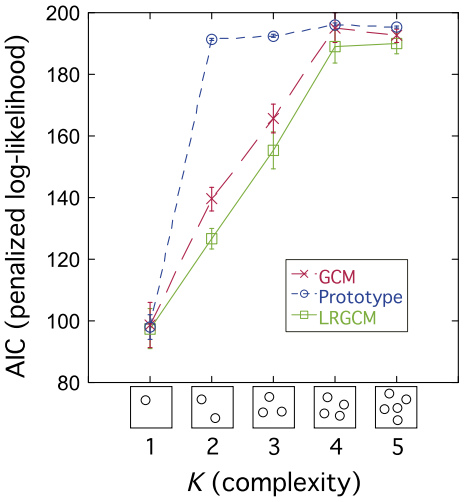

Fig. 9. Fit (AIC) of the three models (GCM, prototype, and LRGCM) to subject data. Lower values indicate better fit.

With unimodal concepts ($K = 1$), all three of the models demonstrate approximately equal levels of performance, with AIC values of 97.4, 98.7, and 98 for the locally regularized, exemplar, and prototype models, respectively. A *t*-test shows no significant differences among the means of the AIC values for the three models ($p = .185$ for prototype vs. LRGCM, $p = .67$ for GCM vs. LRGCM, $p = .06$ for GCM vs. prototype). But at higher complexity levels ($K \geqslant 2$) differences among models become evident. While GCM performs better than the prototype model at these higher levels (AIC values of 139.5 for GCM vs. 191.4 for the prototype model), LRGCM is still better (AIC = 126.7). By allowing the sensitivity parameter to vary locally, the LRGCM is able to capture the variation across the space in the concentrations of examples, setting c high where the training example require it and low where they do not. The subjects apparently share LRGCM's flexibility, as suggested by its superior fits at $K = 2$, 3 and 4 (vs. GCM, respectively $p = .005$, .014, .012; vs. prototype, respectively $p = .001$, .007, .036). There were no significant differences at $K = 5$ where, as previously mentioned, performance approached chance performance and thus all fits were poor.

These results suggest that the broad properties of the LRGCM, in particular its ability to modulate the effective number of modes locally, allow it to fit subject data more closely than competing models. Not only do human learners adopt a moderate position along the bias/variance continuum, they do so in a way that varies from place to place in feature space. This allows the learner to be appropriately "tuned" to the concept even when the concept is of non-uniform complexity.

4.4.1. Fit to individual subjects

We wanted to ensure that the pattern of performance we observed was reasonably stable over subjects (Smith & Minda, 1998). (Recall that although above we report aggregate properties of the fitting, e.g. the mean AIC over subjects, the fitting itself was always to individual subjects, never to aggregate data.) Here we break down the

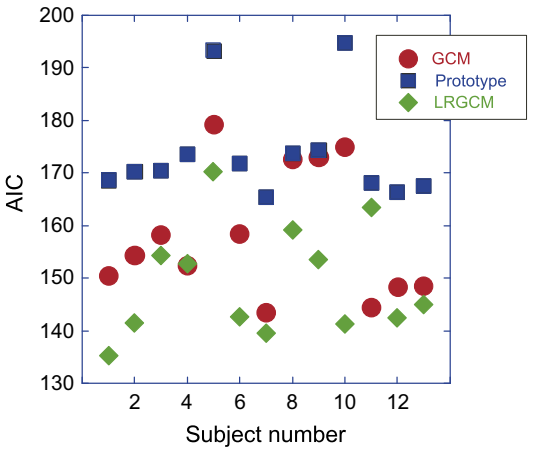

Fig. 10. Fit (AIC) of individual subjects to the three models (GCM, prototype, and LRGCM). Lower values indicate better fit. LRGCM is lowest (best) in 11/13 subjects.

individual fits in more detail. We found that the LRGCM was consistently superior across individual subjects, although of course not perfectly so. Fig. 10 shows the fit (AIC) for each subject for each model (collapsing over complexity). As can be seen in the figure, our LRGCM model fit the data better than the prototype model in all 13 subjects, and better than GCM in 11 out of 13 (worse in 1, and approximately tied in 1). We conclude that the superiority of the LRGCM is reasonably consistent over subjects and is not an artifact of aggregation.

5. Discussion

Many researchers have previously recognized the need to combine the benefits of prototype formation and exemplar memorization. The combination of different approaches has led to a number of "hybrid" models in the literature exhibiting various elements of both. Most of these overtly divide categorization into two distinct components or phases, sometimes thought of as verbal vs. implicit (Ashby et al., 1998). The RULEX model of Nosofsky et al. (1994) has one stage involving simple logical rules, and another involving the storage of occasional exceptions to these rules. Erickson and Kruschke (1998) proposed a hybrid connectionist model for categorization that consists of single-dimensional decision boundaries, an exemplar module for differentiating exemplars and categories, and a gating mechanism to link the two. Their model's behavior predicts that the exemplar module will primarily contribute to classification judgments for stimuli similar to learned exceptions and the rule module will dominate in other cases. Anderson and Betz (2001) likewise formed a hybrid connectionist model by combining both exemplar and rule-based processes into Anderson's ACT-R architecture (Anderson, 1993). Using this framework, they added a theory of strategy selection between an exemplar and rule-based strategy, using the exemplar-based random walk (EBRW) model (Nosofsky & Palmeri, 1997) and the rule-plus-exception (RULEX) model of Nosofsky et al. (1994). While not overtly a hybrid model, the SUSTAIN model of Love et al. (2004), discussed above, allows new clusters to form to accomodate examples that do not fit well into existing clusters, thus allowing the learner to effectively represent multimodal categories similar to those in our experiments.

Apart from the details of particular models, though, our main focus has been the broader theoretical issue of how various models fit on the bias/variance continuum, and again our own LRGCM model is primarily intended as a way of understanding where human learners fit along it. In this spirit, other researchers have also treated prototype and exemplar models as points along some kind of continuum. Smith and Minda (1998) suggested that the balance between prototype or exemplar strategies depends on the stage of concept learning, with prototype strategies prevailing in the early stages, and exemplar models in the late stages. As discussed in Nosofsky (1991), the well-known adaptive model of Anderson (1991) can be seen as equivalent to GCM in the presence of "poor category structure," which in our context means high complexity. The model of Vanpaemel et al. (2005) incorporates both prototype and exemplar aspects by placing them at the extremes of a "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. Implicit in this conceptualization is a core idea we wish to make more explicit, namely that these two extremes differ in how "abstract" they are, i.e. in our terms where they lie along the bias/variance continuum. Ashby and Alfonso-Reese (1995) conceptualized this spectrum in a somewhat different way, emphasizing the more traditional statistical dichotomy between parametric (prototype) and nonparametric (exemplar) models, seeing both as varieties of density estimation. The distinction between parametric and nonparametric statistics is very well-established in traditional statistics, and is indeed related to the bias/variance distinction which arises in machine learning and statistical estimation theory. But the latter differs in that it explicitly connects the estimation procedure to the effectiveness of generalization with further samples from the same data-generating source. In a related vein, Rosseel (2002) has proposed an explicit mixture model, assuming a generalization space that, like the concepts used in our experiment, 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. Finally, the recent models of Feldman (2006) and Goodman et al. (2008) place probabilities over rule-like structures, deriving similarity-like relations from probability calculations in a setting of logic-like rules—again, though in a very different way, placing both approaches in a common formal framework.

In the context of the argument we have made about bias and variance, we would argue that the benefit of a hybrid or mixed approach is to allow the learner some flexibility in setting the bias. "Natural" concepts take on a wide range of complexity levels—not just the two levels favored by conventional models—and learners need be able to

modulate their bias to the examples they encounter. In this light, pure prototype and pure exemplar systems might be seen as laboratory artifacts, which only emerge in pure form in the context of unnaturally simple—or, respectively, unnaturally complex—artifactual concepts.

5.1. Finding the right balance between bias and variance

In this connection, it is natural to ask: what is the "true" level of complexity of naturally-occurring concepts? We would have to know the answer to this question in order to determine what level of abstraction by a learner was, in fact, optimal—because, by definition, only a level appropriately matched to the data-generating source will generalize optimally. Some authors have occasionally touched on this topic, but rarely in a very quantitatively precise way. Rosch (1978) famously suggested that natural concepts tended to comprise subtypes, in a branching typology involving a hierarchy of superordinate and subordinate classes. Augmenting this idea, Keil (1981) proposed that human typologies obey a strict branching-only cladistic structure. More recently, several papers given more quantitative accounts of the inference of such hierarchies (Feldman, 2006; Navarro, 2008; Roy, Kemp, Mansinghka, & Tenenbaum, 2006). All these proposals assume that natural concepts contain a certain amount of substructure, but none quantify exactly *how much*—e.g. how many subdivisions a natural concept typically contains. Without this number, or more precisely without a probability distribution over this parameter, bias and variance cannot be completely optimized for nature.

We argue that this question is impossible to answer definitively, however, for several reasons. First, exactly how many subtypes a concept contains is essentially a subjective question, depending on exactly how a learner chooses to partition the data, which is precisely what we are trying to optimize; thus it cannot be answered definitely without a vicious circle. Second, even if the complexity of concepts could be objectively quantified, it would surely be context-dependent. Classes of fish, for example, could not be assumed to contain the same degree of internal homogeneity and subtypology as do classes of rocks. Nature is heterogenous, and it is heterogeneously heterogeneous.

What we can say is what the human learning mechanism *assumes* about the natural complexity of the environment. Our results suggests that human learners assume an *intermediate* number of components or subtypes in a typical concept (in our data, about two or three, though the exact number presumably depends on the number of training examples)—not one, as assumed by prototype models, nor many, as assumed by pure exemplar models, but somewhere in between.

Of course, the limit on the number of subtypes per concept might—like all forms of bias—be viewed as a performance limitation (in this case a memory limitation) rather than an inductive assumption. That is, perhaps human learners simply can't keep track of too many subtypes per concept. But our argument in effect places a *functional interpretation* on such a limitation. Limiting the number of modes or exemplars that can be stored, rather than hampering learning, might actually enhance it, by adjusting the bias in a way that better suits the environment and thus improves generalization.

5.2. Compositionality

Local regularization as exemplified by the LRGCM places human category formation at an intermediate point on the bias/variance spectrum, specifically allowing the learner to "split" concepts into some small number (in our data, typically 2 or 3) of component concepts. This aspect bears an important relationship to historical arguments about the nature of human concepts, which we briefly mention.

In the philosophical literature on concepts, prototype models have occasionally been criticized for not handling the problem of *compositionality,* namely that concepts are built from, and derive their meaning from, constituent concepts (Fodor, 1996, though see e.g. Kamp & Partee, 1995; Smith & Osherson, 1984). (The philosophical community generally uses the term *prototypes* to refer to fuzzy concepts built on gradations in typicality, thus including both prototype and exemplar models in the psychologist's terminology. The philosophical literature has taken little notice of exemplar models as such, and thus has not seriously considered the debate between prototype and exemplar models that is so prominent in psychology.) The central problem from this perspective is that prototypes are thought to lack internal compositional structure. In a paradigmatic example, the concept *pet fish* has a typicality gradient that seems unrelated to the typicality gradients of its apparent constituents *pet* and *fish*. Notwithstanding this criticism, Goodman, Tenenbaum, Feldman, and Griffiths (2008) have proposed a system for computing Bayesian belief over logical forms that can capture both the compositional structure of concepts as well as observed variations in typicality, apparently accommodating both perspectives. The concept of local regularization brings out another perspective on how prototype-like and compositional aspects of category structure can be reconciled.

Our data show that human learners are happy to treat concepts as, in effect, *mixtures* of several constituent components (cf. Rosseel, 2002), each of which has a separate typicality gradient and thus granularity (operationalized in the model as a locally prevailing value of c). Indeed, "prototypes" in the psychologist's sense—meaning, roughly, unitary statistical summaries of observed instances—are

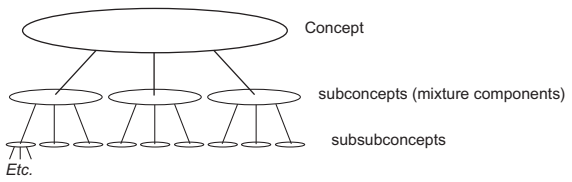

Fig. 11. Hierarchical structure arising in a composite concept with $K = 3$. Unimodal ($K = 1$) prototypes have no composite structure, but higher values of K can induce recursive subtypologies, giving rise to a recursive hierarchy as depicted here.

not compositional, in that they do not comprise constituent subtypes. But more complex mixtures such as those readily learned by our subjects—no longer prototypes in the psychologists' sense, but still prototypes in the philosopher's sense—*are*, by definition, composites of simpler subconcepts. Thus the subconceptual fission introduced by mixtures induces compositionality. In principle this idea can be extended hierarchically to create tree structures, with concepts built from several constituent subconcepts, which in turn are built from several sub-subconcepts, and so forth (Fig. 11). The resulting picture recapitulates the well-established hierarchical structure of human concepts mentioned above (Rosch, 1978; Keil, 1981), and allows typicality gradients to coexist harmoniously with a complex compositional structure. While prototype models (in the psychological sense of the term) assume that categories comprise a single undifferentiated "type," and exemplar models assume they comprise a large (potentially unbounded) number of individual instances, a locally regularized model assumes that categories are composed of a small number of subtypes—which opens the door to a recursively hierarchical constituent structure not available in either conventional approach.

6. Conclusion

Several conclusions, both empirical and conceptual, can be drawn from this study.

First, subjects are proficient at learning $K = 1$ concepts, and though their performance declines as complexity increases, are also fairly competent at learning $K = 2$ and $K = 3$ concepts. At complexity levels $K = 4$ and 5, subjects are less successful, approaching chance performance. All models tested also decrease in performance as complexity increases, though not all at the same rate, and not all in a way that is equally consistent with the human data: the prototype model we tested fell off too quickly to match human performance, and the exemplar model not quickly enough. Indeed the exemplar model was able to learn very complex ($K = 5$) concepts on which subjects were at chance.

Second, more broadly, conceptual complexity influences the degree to which each strategy accurately accounts for human performance. At low but multimodal levels of complexity ($K = 2$ and 3), where the subjects are still able to learn the concepts fairly effectively, prototype models underfit the training data and perform poorly. Their heavy bias towards unimodality means they cannot modulate their decision surfaces so as to accomodate such complex concepts—in contrast to the subjects, who apparently can. At high levels of complexity ($K = 4$ and 5), subjects are no longer able to keep up with the complexity of the concept. But the exemplar model, left to its own devices (i.e. unregularized) can, thus demonstrating an overfit relative to human learners. Our model, the LRGCM, with greater flexibility to regularize where necessary, fits the subject data better. This suggests that human learners have a position along the bias/variance continuum—that is, a set of assumptions about conceptual complexity—that is both more intermediate and more flexible.

Third, prototype and exemplar approaches to categorization may effectively 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. Models that can vary their complexity locally, such as the locally-regularized model we presented here, can therefore subsume the desirable aspects of both prototype an exemplar approaches.

Finally, we draw an important methodological conclusion. When carrying out studies of human learning, concepts with a range of complexity values must be included. The $K = 1$ concepts, which are unimodal Gaussian clouds and thus resemble many concepts studied in the literature, elicited very similar performance from prototype and exemplar models—which, in light of their very divergent results at other complexity values, must be considered misleading. Indeed, this result alone sheds some light on the many decades of controversy about the relative merits of the two approaches, which are indeed difficult to distinguish if experimental concepts are poorly chosen. Because clear differences between the two strategies only emerged at complexity $K = 2$ and higher, it seems imperative to include such concepts in any study of human concept learning. Indeed, it seems important that complexity should be systematically varied over a range of levels, as for all the reasons detailed above it has the power to modulate the relative merits of competing models. Only this type of comprehensive investigation can guarantee a sufficiently broad view of human generalization under arbitrary conditions.

Acknowledgments

This research was supported by NSF SBE-0339062 to J.F. E.B. was supported in part by the Rutgers NSF IGERT program in Perceptual Science, NSF DGE 0549115. We are grateful to Cordelia Aitkin, David Fass, Fabien Mathy, and three anonymous reviewers for helpful comments and discussions. An earlier report on this research appeared in the Proceedings of the Cognitive Science Society, 2006.

References

D. Aha, & R.L. Goldstone (1992). Concept learning and flexible weighting. In *Proceedings of the fourteenth annual conference of cognitive science society* (pp. 534–539).

Aitkin, C.D. & Feldman, J. (2006). Subjective complexity of categories defined over three-valued features. In *Proceedings of the 28th conference of the cognitive science society* (pp. 961–966).

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review, 98*(3), 409–429.

Anderson, J. R. (1993). *Rules of the mind*. New Jersey: Hillsdale.

Anderson, J. R., & Betz, J. (2001). A hybrid model of categorization. *Psychonomic Bulletin and Review, 8*, 629–647.

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

Ashby, F. G., Alfonso-Reese, L. A., Turken, U., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review, 105*(3), 442–481.

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

Blair, M., & Homa, D. (2003). As easy to memorize as they are to classify: The 5–4 categories and the category advantage. *Memory & Cognition, 31*(8), 1293–1301.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research, 33*(2), 261–304.

Chater, N. (1997). Simplicity and the mind. *Psychologist, 10*(11), 495–498.

Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern classification*. New York: Wiley.

Erickson, M., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General, 127*, 107–140.

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* (Vol. 15). Cambridge: MIT Press.

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

Feldman, J. (2003). The simplicity principle in human concept learning. *Current Directions in Psychological Science, 12*(6), 227–232.

Feldman, J. (2006). An algebra of human concept learning. *Journal of Mathematical Psychology, 50*, 339–368.

Fodor, J. (1996). The pet fish and the red herring: Why concepts aren't prototypes. *Cognition, 58*, 243–276.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural computation, 4*, 1–58.

Gentner, D., & Medina, J. (1998). Similarity and the development of rules. *Cognition, 65*(2/3), 263–297.

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science, 32*(1), 108–154.

Griffiths, T. L., Sanborn, A. N., Canini, K. R., & Navarro, D. J. (2008). Categorization as nonparametric bayesian density estimation. In M. Oaksford & N. Chater (Eds.), *The probabilistic mind: Prospects for rational models of cognition*. Oxford: Oxford University Press.

Hahn, U., & Chater, N. (1998). Similarity and rules: Distinct? exhaustive? empirically distinguishable? *Cognition, 65*, 197–230.

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

Jäkel, F., Schölkopf, B., & Wichmann, F. A. (2007). A tutorial on kernel methods for categorization. *Journal of Mathematical Psychology, 51*(6), 343–358.

Jones, M. C., McKay, I. J., & Hu, T.-C. (1994). Variable location and scale kernel density estimation. *Annunal of the Institute of Statistical Estimation, 46*(3), 521–535.

Kamp, H., & Partee, B. (1995). Prototype theory and compositionality. *Cognition, 57*(2), 129–191.

Keil, F. (1981). Constraints on knowledge and cognitive development. *Psychological Review, 88*(3), 197–227.

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

Lamberts, K. (1997). Process models of categorization. In K. Lamberts & D. Shanks (Eds.), *Knowledge, concepts and categories* (pp. 371–403). Cambridge: MIT Press.

Love, B., Medin, D. L., & Gureckis, T. (2004). SUSTAIN: A network model of category learning. *Psychological Review, 111*(2), 309–332.

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

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco: Freeman.

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

McLachlan, G. J., & Basford, K. E. (1988). *Mixture models: Inference and applications to clustering*. New York: Marcel Dekker.

Medin, D. L., & Schaffer, M. M. (1978). Context model 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, 27*(3), 775–799.

Navarro, D.J. (2008). From natural kinds to complex categories. In *Proceedings of the conference of the cognitive science society* (pp. 621–626).

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

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

Nosofsky, R. M. (1991). Relation between the rational model and the context model of categorization. *Psychological Science, 2*(6), 417–421.

Nosofsky, R. M., & Johansen, M. (2000). Exemplar-based accounts of mulitple-system phenomena in perceptual categorization. *Psychonomic Bulletin and Review, 7*, 375–402.

Nosofsky, R. M., Palmeri, T., & McKinley, S. (1994). Rule-plus-exception model of classification learning. *Psychonomic Bulletin and Review, 101*, 53–79.

Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review, 104*, 266–300.

Op de Beeck, H. P., Wagemans, J., & Vogels, R. (2008). The representation of perceived shape similarity and its role for category learning in monkeys: A modeling study. *Vision Research, 48*(4), 598–610.

Posner, M. I., Goldsmith, R., & Welton, K. E. (1967). Perceived distance and the classification of distorted patterns. *Journal of Experimental Psychology, 73*(1), 28–38.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77*(3), 353–363.

Pothos, E. M., & Chater, N. (2002). A simplicity principle in unsupervised human categorization. *Cognitive Science, 26*, 303–343.

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. Lloyd (Eds.), *Cognition and categorization*. New York: Lawrence Erlbaum Associates.

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

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

Roy, D.M., Kemp, C., Mansinghka, V.K., & Tenenbaum, J.B. (2006). Bayesian modeling of human concept learning. In *Advances in neural information processing systems*.

Schölkopf, B., & Smola, A. J. (2002). *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. Cambridge, MA, USA: MIT Press.

Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika, 22*(4), 325–345.

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

Smith, E. E., & Medin, D. (1981). *Categories and concepts*. Cambridge, MA: Harvard University Press.

Smith, E. E., & Osherson, D. N. (1984). Conceptual combination with prototype concepts. *Cognitive Science, 8*(4), 337–361.

Smith, E. E., Patalano, A. L., & Jonides, J. (1998). Alternative strategies of categorization. *Cognition, 65*, 167–196.

Smith, J. D. (2005). Wanted: A new psychology of exemplars. *Canadian Journal of Experimental Psychology, 59*, 47–53.

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. (2008). In search of abstraction: The varying abstraction model of categorization. *Psychonomic Bulletin & Review, 15*, 732–749.

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.

Verbeemen, T., Storms, G., & Verguts, T. (2003). Varying abstraction in categorization: A *k*-means approach. In *Proceedings of the 27th annual conference of cognitive science society*.

Verguts, T., & Storms, G. (2004). Assessing the informational value of parameter estimates in cognitive models. *Behavior Research Methods, Instruments, & Computers*.