Cognitive Science 32 (2008) 1403–1424
Copyright © 2008 Cognitive Science Society, Inc. All rights reserved.
ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1080/03640210802073697

Exemplars, Prototypes, Similarities, and Rules in Category Representation: An Example of Hierarchical Bayesian Analysis

Michael D. Lee[a], Wolf Vanpaemel[b]

[a]*Department of Cognitive Sciences, University of California, Irvine*
[b]*Department of Psychology, University of Leuven*

Received 12 July 2007; received in revised form 14 January 2007; accepted 15 January 2008

Abstract

This article demonstrates the potential of using hierarchical Bayesian methods to relate models and data in the cognitive sciences. This is done using a worked example that considers an existing model of category representation, the Varying Abstraction Model (VAM), which attempts to infer the representations people use from their behavior in category learning tasks. The VAM allows for a wide variety of category representations to be inferred, but this article shows how a hierarchical Bayesian analysis can provide a unifying explanation of the representational possibilities using 2 parameters. One parameter controls the emphasis on abstraction in category representations, and the other controls the emphasis on similarity. Using 30 previously published data sets, this work shows how inferences about these parameters, and about the category representations they generate, can be used to evaluate data in terms of the ongoing exemplar versus prototype and similarity versus rules debates in the literature. Using this concrete example, this article emphasizes the advantages of hierarchical Bayesian models in converting model selection problems to parameter estimation problems, and providing one way of specifying theoretically based priors for competing models.

Keywords: Varying Abstraction Model; Hierarchical Bayesian models; Generalized Context Model; Category learning

1. Introduction

For a cognitive scientist interested in category learning, there are two ways Bayesian statistics might make a contribution. The first way is to use Bayesian methods as a theoretician would, as a metaphor or working assumption about how the mind solves the inference problems it faces. Anderson's (1991) rational model of categorization, and its recent developments and

Correspondence should be sent to Michael D. Lee, Department of Cognitive Sciences, University of California, Irvine, Irvine, CA 92697-5100. E-mail: mdlee@uci.edu

extensions (e.g., Griffiths, Canini, Sanborn, & Navarro, 2007), are good examples of this approach. Consistent with other “rational” or “computational-level Bayesian” models (e.g., Chater, Tenenbaum, & Yuille, 2006), these models provide accounts of what sorts of inferences people make if they aim to categorize objects according to Bayesian ideals.

The second way of using Bayesian methods in category learning is as a statistician would, as a framework for making justified inferences from the available models and data. In this application, there are no constraints on the nature of the category learning models that can be considered, and existing process models that do not necessarily have any Bayesian basis (e.g., Kruschke, 1992; Love, Medin, & Gureckis, 2004; Nosofsky, 1986) can be considered. The role of Bayesian methods is to improve the analysis of the models and data, consistent with the push to adopt Bayesian statistical inference throughout cognitive modeling (e.g., Lee & Wagenmakers, 2005; Pitt, Myung, & Zhang, 2002).

This article considers how Bayesian methods—especially in their hierarchical form—can help in modeling category learning when used in the second, statistical sense. Hierarchical Bayesian methods are standard and powerful ways of analyzing models and drawing inferences about parameters from data, and are widely used in statistics, machine learning, and throughout the empirical sciences. The hierarchical Bayesian approach employs the basic machinery of Bayesian statistical inference, with all the advantages it entails (e.g., Jaynes, 2003; Sivia, 1996), but is designed to work with richly structured hierarchical models. Introductions to hierarchical Bayesian methods can be gained from textbook accounts in statistics and machine learning (e.g., Gelman, Carlin, Stern, & Rubin, 2004; Mackay, 2003) or from recent expositions aimed at psychologists (e.g., Griffiths, Kemp, & Tenenbaum, 2008; Lee, 2008; Shiffrin, Lee, Wagenmakers, & Kim, this issue).

To make our case that hierarchical Bayesian methods can contribute to understanding category learning, we tackle a specific example in some detail. We focus on two fundamental debates in the category representation literature. One debate—comparing exemplar and prototype representations—asks to what extent abstraction is involved in representing categories (e.g., Komatsu, 1992; Nosofsky, 1987, 1992; Smith & Minda, 2000). Another debate—comparing the use of similarities and rules—asks on what basis categories cohere into a single meaningful representation of a class of stimuli (e.g., Nosofsky, Clark, & Shin, 1989; Nosofsky & Palmeri, 1998). Both debates can usefully be regarded as model selection or evaluation problems. Progress has most often been sought by developing models that adopt the various theoretical positions, and seeking evidence for or against these models on the basis of experimental data.

Although this approach is a sensible starting point, it has a number of difficulties and limitations. One is that specifying separate models for different theoretical positions can obscure underlying compatibilities and forces model evaluation to become a matter of choosing one model to the exclusion of the other, rather than searching for the good and bad elements of each model. We show how hierarchical Bayesian methods can address this problem by converting a model selection problem to a parameter estimation problem.

A second difficulty with comparing distinct models is that any complete approach to evaluating models, including particularly Bayesian model selection, requires the specification of their prior probabilities (Pitt et al., 2002). Most current evaluations make the working assumption of equal prior probabilities (i.e., that there is no reason to believe one model is better than another until data have been collected), even when there is existing knowledge that makes this assumption problematic. The assumption of equal prior probabilities comes primarily from a lack of formal methods for determining priors. We show that hierarchical Bayesian methods provide one avenue for determining the required prior probabilities formally, using existing theories and knowledge.

To demonstrate these properties of hierarchical Bayesian analysis, we focus on the recently developed Varying Abstraction Model (VAM; Vanpaemel & Storms, 2008) of category representation. Our aim is to show how, using hierarchical Bayesian methods to do inference with the VAM, basic but otherwise difficult questions about category representation can be addressed.

This article is structured as follows. We begin by providing an intuitive introduction to the VAM account of category representation, before explaining our approach to its hierarchical Bayesian analysis. We then analyze 30 previously studied data sets from the category learning literature using hierarchical Bayesian methods. The results demonstrate the sorts of insights into category representation—particularly in terms of the exemplar versus prototype and similarity versus rules debates—that can emerge from the analysis. We conclude with a discussion of how the hierarchical Bayesian approach helped in providing these evaluations, and mention future possibilities for the approach in the cognitive sciences more generally.

2. The VAM

Fig. 1 shows 15 different representations of a category with four stimuli. Each stimulus is represented by a two-dimensional point in space. The different representations are

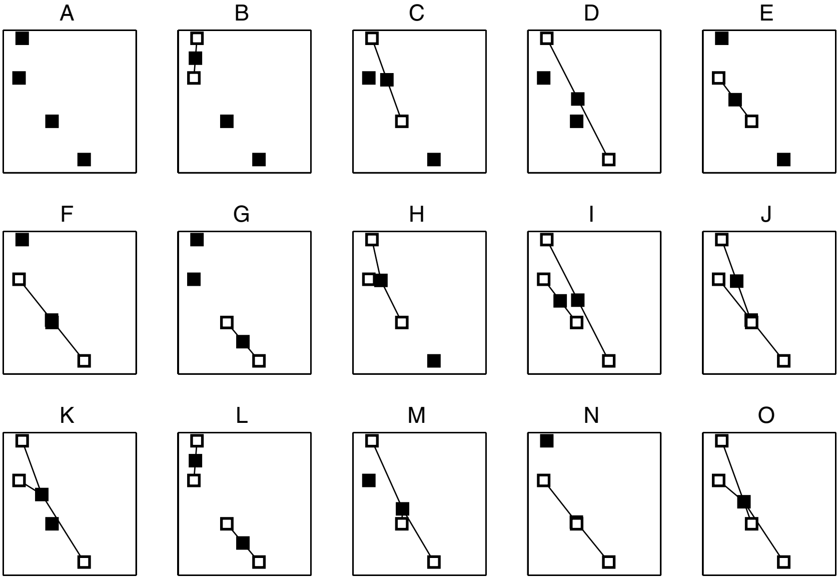

Fig. 1. The 15 possible Varying Abstraction Model representations for a four-stimulus category.

formed by merging one or more subsets of the stimuli. The exemplar representation in Panel A is the one where no stimuli are merged. Panel B shows the representation created when two of the stimuli are merged, with the original stimuli shown as white squares, joined by lines to their merged representation, which are black squares. The remaining panels in Fig. 1 show the representations resulting from averaging other stimuli. Panels B through G show the results of a single merge, whereas Panels H through N show the results of two merges. The representation in Panel O shows the prototype representation in which all four stimuli are merged into a single representation.

The category representations shown in Fig. 1 are exactly those described by the VAM. They naturally encompass both exemplar and prototype accounts, as well as allowing for various intermediate accounts, including ones that use multiple prototypes. In addition, which stimuli are merged in the representations spans a range of possibilities in the similarity versus rules debate. For example, if similarity plays a central role, the category representations B and E are more likely than C and D. In this way, the class of VAM representations generalizes and unifies the exemplar versus prototype and similarity versus rules distinctions. It includes the extreme versions of these theoretical positions as special cases, but introduces a set of intermediate representations that facilitate asking estimation rather than selection questions. It was for these reasons we chose to focus on the VAM in a hierarchical Bayesian setting, although there are other category learning models, including SUSTAIN (Love et al., 2004) and various mixture-models (e.g., Rosseel, 2002), which are also worth considering in the future.

Original applications of the VAM (Vanpaemel & Navarro, 2007; Vanpaemel & Storms, 2008; Vanpaemel, Storms, & Ons, 2005) inferred which of the VAM class of representations was being used from existing behavioral data. Choosing between the representations amounted to a model selection problem, which was tackled using maximum likelihood methods. Although the results of these analyses are informative, they have at least three important limitations.

The most obvious limitation is that maximum likelihood evaluation is insensitive to the different inherent complexities of the various representations. The importance of balancing goodness-of-fit with complexity in model evaluations has been emphasized in psychology in recent years (e.g., Myung, Forster, & Browne, 2000; Pitt et al., 2002), and model evaluation therefore requires better techniques than maximum likelihood methods.

A second important limitation of previous VAM modeling is that it does not capture important relationships between the representations. For example, inferring from data that representation A or B in Fig. 1 are the ones likely being used demands a very different interpretation from inferring that representations A or O are being used. In the first case, both possibilities are consistent with exemplar representation, in the sense that little abstraction is being used. In the second case, however, it is not clear whether exemplar or prototype representation in being used. Formally, however, the VAM modeling in both cases just reports that two representations are likely. The close theoretical links between representations A and B, and the important theoretical differences between representations A and O, are not captured. In this sense, the model is incomplete as a mechanism for estimating from data the level of abstraction used in category representation. The same is true for the problem of inferring the importance of similarity.

A final limitation of previous VAM modeling is that representational models are evaluated without incorporating prior probabilities. In relation to Fig. 1, for example, it has been assumed that each of the 15 representations is equally likely *a priori*. It seems reasonable, however, to assert that some of the representations are more likely than others. In particular, there exists evidence for both the pure exemplar and pure prototype representations A and O; and representations including B, E, G, and L seem more likely than many of the others. It is desirable to have a method for incorporating priors with these sorts of properties into the evaluation process. The goal of specifying priors is not, of course, to try to override the evidence provided by data. Rather, the goal is to provide an *inductive bias*, preferring some representations to others on some principled basis. Given sparse data, these biases will help guide more intuitive inferences. Given overwhelming data, they will be overturned where inappropriate.

3. Hierarchical Bayesian analysis for the VAM

3.1. The hierarchical analysis

We address all of these limitations using hierarchical Bayesian methods to extend and analyze the VAM, taking our lead from Vanpaemel and Lee (2007). Fig. 2 provides a schematic representation of our hierarchical analysis,[1] showing the three levels in the hierarchy. At the bottom level are the observed data, in the form of counts k of the number of times each stimulus is classified as belonging to Category A in a two-category task.[2]

In the middle level is the category representation. This includes the VAM representation x that is used (i.e., x is a counting number indicating which of the VAM family of representations—as, for example, in Fig. 1—is the one being used). It also includes the generalization gradient c and attention weight w that are applied to this representation. In the hierarchical VAM, as in the original VAM, it is assumed the categorization process follows

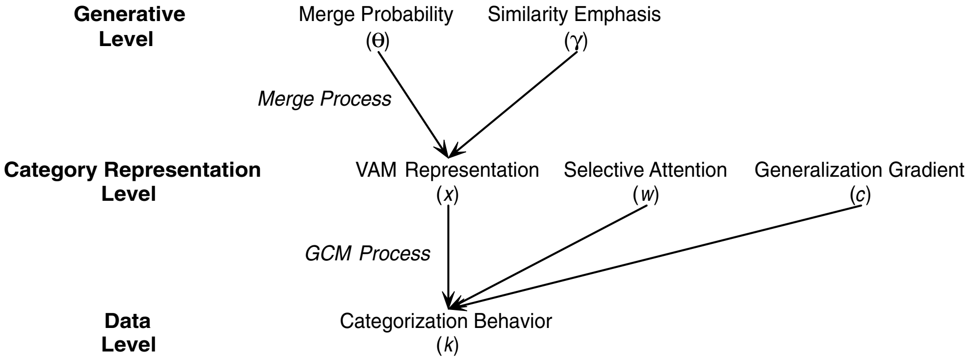

Fig. 2. Schematic representation of the hierarchical Bayesian analysis. VAM = Varying Abstraction Model; GCM = Generalized Context Model.

the Generalized Context Model (GCM; Nosofsky, 1986). We describe this category learning process in detail later.

It is the top level of the hierarchy in Fig. 2, however, that makes our analysis hierarchical and extends the original VAM. Rather than placing a prior distribution directly on the representations x—in effect, this is what the original VAM did, using a uniform prior, and so making each possible VAM representation equally likely—we assume that the VAM representations themselves are generated by an additional psychological process. This process is called the *merge* process, and is driven by two parameters, θ and γ. The parameter θ controls the merging probability, and so dictates whether exemplar-like or prototype-like VAM representations are generated. The parameter γ controls the emphasis on similarity, and so dictates the extent to which similar stimuli are merged.

3.1.1. The merge process

More formally, the merge process starts with the exemplar representation. The parameter $0 \leq \theta \leq 1$ then gives the probability that an additional merge will take place. This means, at any stage, there is a $1 - \theta$ probability that the current representation will be maintained as the final one. When an additional merge is undertaken, it is based on the similarities between all of the current representations (i.e., the original stimuli or their merged replacement). The similarity between the ith and jth representations is modeled as an exponentially decaying function of the distance between their points, according to a Minkowski r-metric:

$$s_{ij} = \exp\left\{ - \left[\sum_k (|v_{ik} - v_{jk}|^r) \right]^{1/r} \right\}, \tag{1}$$

where v_{ik} is the coordinate location on the kth dimension for the point that represents the ith stimulus. Given these similarities, across all pairs in the current representation, the probability, m_{ij}, of choosing to merge the pair (i, j) is given by an exponentiated Luce-choice rule:

$$m_{ij} = \frac{(\exp s_{ij})^\gamma}{\sum_x \sum_{y \geq x} (\exp s_{xy})^\gamma}. \tag{2}$$

The parameter $\gamma \geq 0$ controls the level of emphasis given to similarity in determining the pair to be merged. As γ increases, the maximally similar pair dominates the others, and will be chosen as the pair to be merged with probability approaching one. At the other extreme, when $\gamma = 0$, similarity is not taken into account. All choices of pairs to merge then are equally likely, and the merge is essentially chosen at random. Values of γ between these two extremes result in intermediate behavior.

Given a value for the θ and γ parameters, every VAM representation has some probability of being generated by the merging process. The top five rows in Fig. 3 give some examples for the 15 VAM representations in Fig. 1. In the top row $\theta = 0.99$, so merging is very likely; hence, the prototype representation almost always results. In the second row $\theta = 0.01$, so merging is very unlikely; hence, the exemplar representation is almost always retained. The third, fourth, and fifth rows show, for a fixed $\theta = 0.7$, the effect of the γ parameter. When $\gamma = 0$ in the third row, the exemplar and prototype representations are most likely, but all others are possible. In particular, any representation arising from a single merge is equally

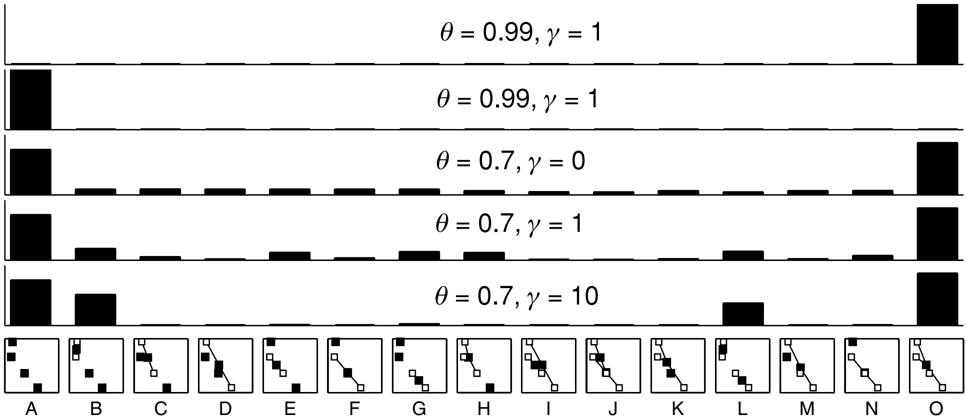

Fig. 3. The bottom row shows the 15 possible Varying Abstraction Model (VAM) representations for a four-stimulus category structure. The top five rows give the probability distribution over these 15 representations for specific combinations of the θ and γ parameters.

likely, and any representation arising from two merges is equally likely because the pair of stimuli to be merged is chosen at random. In the fourth row, when $\gamma = 1$, representations like B and L that involve merging similar stimuli become much more likely, although some other possibilities remain. Once $\gamma = 10$ in the fifth row, only the most similar stimuli are merged, and B and L are the only intermediate possibilities between exemplar and prototype representation with non-negligible probability.

3.1.2. The GCM process

In the categorization process, the attention-weighted distances between the original stimuli and representations are calculated, according to the Minkowski r-metric, so that

$$d_{ij} = (w|p_{i1}-v_{j1}|^r + (1-w)|p_{i2}-v_{j2}|^r)^{\frac{1}{r}}, \tag{3}$$

where p_{ik} is the value of the ith stimulus on the kth dimension, and w is the attention weight parameter measuring the relative emphasis given to the first stimulus dimension over the second.

From the distances, the generalization gradient with scale parameter c and shape α determines the similarities:

$$\eta_{ij} = \exp\big\{ - cd_{ij}^{\alpha}\big\}. \tag{4}$$

The assignment of the representations to the two categories is defined by the category structure of the task. The probability of the ith stimulus being chosen as a member of the Category A is determined by the sum of similarities between the ith stimulus to the N_x representations in each category, according to the choice rule:

$$r_i = \frac{\sum_j a_j \eta_{ij}}{\sum_j a_j \eta_{ij} + \sum_j (1 - a_j)\eta_{ij}}, \tag{5}$$

where a_j indicates the category to which the jth stimulus belongs.

Finally, the response probabilities are used to account for the observed data, which are the counts, k_i of the number of times the ith stimulus was chosen in Category A out of the t_i trials it was presented. The counts k_i follow a binomial distribution

$$k_i \sim \text{Binomial}(t_i, r_i). \tag{6}$$

The only way in which this categorization process differs from the GCM is that the category similarity of the stimuli presented to participants is formed from their similarities to the VAM category representation (i.e., the possibly abstracted v_{ik} coordinates), rather than to an assumed exemplar representation (i.e., the original p_{ik} coordinates).

3.1.3. Priors

The final part of the hierarchical Bayesian analysis involves specifying priors for the θ and γ parameters of the merge process, and the c and w parameters of the categorization process. For the generating level parameters, we use priors:

$$\begin{aligned} \theta &\sim \text{Uniform } (0, 1), \\ \gamma &\sim \text{Erlang } (1). \end{aligned} \tag{7}$$

The uniform prior for the rate θ is an obvious choice. The Erlang prior for γ gives support to all positive values, but has most density around the modal value one, corresponding to our prior expectations. As an alternative prior for γ we considered a Uniform distribution on the range (0,10). All of our results were qualitatively identical, and quantitatively extremely similar, with this alternative prior. It is common for hierarchical analyses to have this property, where conclusions become less sensitive to the choice of priors, as more levels are included in a hierarchical model (Gelman et al., 2004).

For the category representation level parameters, we use priors:

$$\begin{aligned} w &\sim \text{Uniform } (0, 1), \\ c^2 &\sim \text{Gamma } (\varepsilon, \varepsilon). \end{aligned} \tag{8}$$

The uniform distribution for w is again an obvious choice. The c parameter functions as an inverse scale (i.e., $1/c$ scales the distances), implying c^2 functions as a precision, and so is given the standard near non-informative Gamma prior with $\varepsilon = .001$ set near zero.

3.2. Inference methods

The hierarchical Bayesian analysis of the VAM defines a precise and complete probabilistic relationship between the parameters of the model and the observed data; that is, it specifies a likelihood function giving the probability of observed data for a given parameterization. The four parameters are the θ and γ parameters that control the representation, and the c and w parameters that control the categorization. Bayesian inference uses the relationship between parameters and data to update what is known about the parameters, converting their joint prior distribution to a joint posterior distribution, using the evidence provided by data. One of the great advantages of the Bayesian approach is that these inferences are conceptually straightforward. Once behavioral data are observed, inference just involves reversing the generative process and working out what parameter combinations are the ones likely to have produced the data. The posterior probability distribution represents this information, specifying the relative probability of each possible combination of θ, γ, c, and w being the ones that generated the data. We give some details on how we calculated the posterior probability statements of interest for this application in the Appendix.

3.3. Features of the hierarchical analysis

The hierarchical Bayesian analysis of the VAM has many attractive properties. In particular, it addresses the three shortcomings of the original application of the VAM—balancing goodness-of-fit with model complexity, relating and interpreting the VAM representations, and having sensible prior probabilities for the representations—we identified earlier.

3.3.1. Principled statistical inference

The posterior distributions obtained from Bayesian analysis are complete and coherent. Once a model is built (i.e., the likelihood function that relates parameters to models is specified), and the priors are given, the observed data automatically dictate what the posterior inferences must be, governed by the laws of probability theory (e.g., Cox, 1961). These posterior distributions represent everything that is know and unknown about the parameters based on the model, the prior assumptions, and the data. This clarity and generality contrasts favorably with the *ad hoc* set of methods for model analysis that currently dominate practice in the cognitive sciences.

Perhaps most important, however, hierarchical Bayesian methods implement full Bayesian model selection, and so automatically balance goodness of fit with complexity. As has been pointed out a number of times in the context of cognitive modeling (e.g., Lee, 2004, 2008; Pitt et al., 2002), the key feature of fully Bayesian analysis is that it evaluates how well a model fits data on average (i.e., the marginal likelihood), rather than how well it fits data in the best-case scenario (i.e., the maximum likelihood). More complicated models, by definition, are those that are able to predict more data patterns by varying their parameters (Myung, Balasubramanian, & Pitt, 2000). This means that, by considering the average fit of a model, Bayesian model selection penalizes more complicated models because it includes those predictions that poorly fit the observed data in the average.

3.3.2. Interpreting VAM representations

The original VAM, by introducing a range of category representations between prototypes and exemplars, conceived of the debate between these two possibilities as a problem of estimating the level of abstraction, rather than choosing one extreme over the other. The hierarchical analysis provides the extra benefit of allowing the representations to be interpreted and related to one another. In particular, the parameters of the merge process quantify interpretable properties of the VAM representations relating to the level of abstraction and the role of stimulus similarity in abstraction.

This means the posterior distributions over θ and γ inferred from category learning data directly convey the conclusions that can be drawn about abstraction and similarity. Returning to our earlier example, if representations A and B are the most likely, the posterior for θ will give most probabilities to low values, showing the consistent (near) exemplar nature of the representations. If representations A and O are most likely, the posterior for θ will show the divergent nature of the conclusions, giving density to both low and high values. In this way, the hierarchical introduction of the parameterized merge process captures the relationships between VAM representations.

3.3.3. Priors using inductive bias

Fig. 4 shows the overall inductive bias over the 15 VAM representations in Fig. 1 imposed by the merge process. This is simply the average of all of the distributions, like those shown in Fig. 3, which result from some combination of θ and γ, weighted by the priors on θ and γ, as given formally by Equation 11 in the Appendix. It can be seen in Fig. 4 that the bias is strongly toward pure exemplar or pure prototype representations A and O, and with some greater emphasis on the intuitively more likely representations like B, E, G, H, and L.

The inductive bias distribution in Fig. 4 is naturally and appropriately interpreted as a prior distribution for the VAM category representations. The distribution follows directly from the way the merge process works and the priors on the θ and γ parameters that control the process. This means it directly follows from theoretical assumptions about how the VAM category representations are generated, and is fully specified before any data have been observed or inferences have been made. In this sense, our hierarchical Bayesian analysis defines a theoretically principled prior distribution over the category representations.

We emphasize that the inductive bias corresponds to priors at the category representation level, not at the level of the θ and γ parameters. The goal of the current modeling is to place sensible priors on the category representations because they do not all seem equally plausible, and this is achieved by introducing the higher level generating process. A worthy, but different, goal would be to determine priors from first principles for the θ and γ parameters of the generating process itself (using, perhaps, the transformation invariance ideas advocated by Jaynes, 2003, chap. 12).

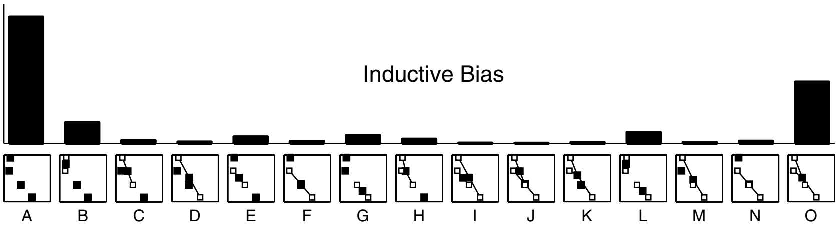

Fig. 4. The 15 possible Varying Abstraction Model representations for a four-stimulus category structure and the inductive bias over these representations defined by the hierarchical analysis.

4. Data for empirical evaluation

To apply our hierarchical Bayesian approach to the VAM, we considered 30 previously published data sets, taken from six articles presented by Nosofsky and colleagues (Nosofsky, 1986, 1987, 1989; Nosofsky et al., 1989; Nosofsky & Palmeri, 1997, 1998). These data sets are detailed in Table 1, and have previously been analyzed by Vanpaemel and Storms (2008) using the original VAM and maximum likelihood methods.

The experiments all involve learning various two-category structures over a small number of stimuli, all of which vary on two continuous dimensions. In each experiment, a subset of training stimuli are assigned to Categories A and B, and the remaining stimuli are retained as transfer stimuli. In most of the experiments, a training-test procedure is used, which consists of a training (or category learning) phase followed by a test phase. During the training phase, only the training stimuli are presented, with corrective feedback. The relevant data for modeling are from the test phase, recording how people categorized all stimuli, including those for which they had not received training.

Collectively, the data sets span a range of possibilities for category learning tasks. Some report single participant data, and others report aggregated data over groups of participants. Some also report data at different stages in the learning sequence. The stimuli vary from simple geometric shapes, to colors, to combinations of visual and auditory stimuli. The category structures that must be learned vary widely and include some that require selective attention for effective learning. Several of the tasks involve a variety of instruction conditions, including people being told to follow rules. And the data sets vary significantly in the number of test trials used.

5. Results of hierarchical Bayesian analysis

Using the observed behavioral data, we drew inferences about the model parameters for all 30 of the data sets listed in Table 1. Our primary interest is on two posterior distributions: the representation posteriors $p(x \mid D)$, which describe the inferences made by the model about what VAM representation is being used; and the marginal parameter posteriors $p(\theta \mid D)$ and $p(\gamma \mid D)$, which describe inferences about what process people used to generate that representation, informing us about the use of abstraction and the importance of similarity, respectively.

We divide the results presented here into three parts. First, we examine how the results can inform the exemplar versus prototype issue. Second, we turn to the similarity versus rules issue. Finally, we present a series of other findings that show how our approach can move beyond the motivating research questions, and suggest further theoretical and modeling developments.

5.1. Exemplar versus prototype representations

Fig. 5 shows in the upper panels the maximum *a posteriori* VAM representations for 14 selected data sets. All of these single representations accounted for more than one half of the

Table 1
Details of the 30 data sets

Data Set	Reference	Experiment	Stimuli	r	α	Categories	Condition	s	t
1	Nosofsky (1986)	—	Shepard circles	2	2	Criss-cross	Participant 1	1	219^{a}
2	Nosofsky (1986)	—	Shepard circles	2	2	Criss-cross	Participant 2	1	225^{a}
3	Nosofsky (1986)	—	Shepard circles	2	2	Diagonal	Participant 1	1	250^{a}
4	Nosofsky (1986)	—	Shepard circles	2	2	Diagonal	Participant 2	1	225^{a}
5	Nosofsky (1986)	—	Shepard circles	2	2	Dimensional	Participant 1	1	225^{a}
6	Nosofsky (1986)	—	Shepard circles	2	2	Dimensional	Participant 2	1	200^{a}
7	Nosofsky (1986)	—	Shepard circles	2	2	Interior–exterior	Participant 1	1	256^{a}
8	Nosofsky (1986)	—	Shepard circles	2	2	Interior–exterior	Participant 2	1	144^{a}
9	Nosofsky (1987)	2	Munsell colors	2	1	Brightness	—	49	10
10	Nosofsky (1987)	2	Munsell colors	2	1	Criss-cross	—	24	10
11	Nosofsky (1987)	2	Munsell colors	2	1	Saturation A	—	24	7.5^{a}
12	Nosofsky (1987)	2	Munsell colors	2	1	Saturation B	—	40	10
13	Nosofsky (1989)	—	Shepard circles	2	2	Angle	—	41	4^{a}
14	Nosofsky (1989)	—	Shepard circles	2	2	Criss-cross	—	37	4^{a}
15	Nosofsky (1989)	—	Shepard circles	2	2	Diagonal	—	43	4^{a}
16	Nosofsky (1989)	—	Shepard circles	2	2	Size	—	37	4^{a}
17	Nosofsky et al. (1989)	1	Shepard circles	1	1	Interior–exterior	Free	122	5
18	Nosofsky et al. (1989)	2	Shepard circles	1	1	Interior–exterior	Rule 1	30	5
19	Nosofsky et al. (1989)	2	Shepard circles	1	1	Interior–exterior	Rule 2	28	5
20	Nosofsky et al. (1989)	3	Pitch and line length	1	1	Conjunctive rule	Free 1	30	5
21	Nosofsky et al. (1989)	3	Pitch and line length	1	1	Conjunctive rule	Free 2	30	5
22	Nosofsky et al. (1989)	3	Pitch and line length	1	1	Conjunctive rule	Rule	30	5
23	Nosofsky and Palmeri (1997)	2	Munsell colors	2	1	U7	—	31	8
24	Nosofsky and Palmeri (1997)	2	Munsell colors	2	1	U8	—	31	8
25	Nosofsky and Palmeri (1998)	1	Shape and brightness	1	1	Interior–exterior	Beginning blocks	164	3
26	Nosofsky and Palmeri (1998)	1	Shape and brightness	1	1	Interior–exterior	Middle blocks	164	3
27	Nosofsky and Palmeri (1998)	1	Shape and brightness	1	1	Interior–exterior	End blocks	164	3
28	Nosofsky and Palmeri (1998)	2	Shape and brightness	1	1	Interior–exterior	Beginning blocks	120	1
29	Nosofsky and Palmeri (1998)	2	Shape and brightness	1	1	Interior–exterior	Middle blocks	120	1
30	Nosofsky and Palmeri (1998)	2	Shape and brightness	1	1	Interior–exterior	End blocks	120	3

Note. r = metric; α = similarity function; s = number of participants; t = number of trials per test stimulus per participant.
[a] = approximate average value.

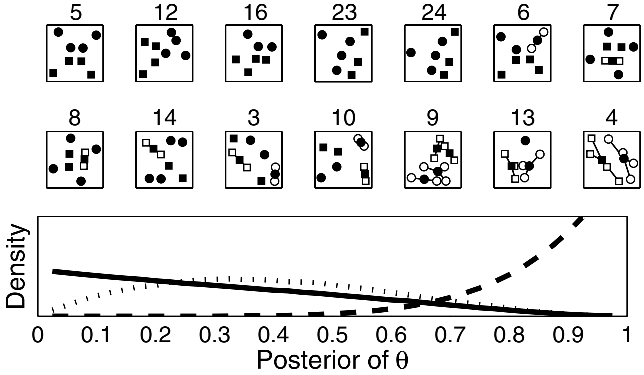

Fig. 5. The upper panels show the maximum *a posteriori* representations for 14 of the data sets, labeled according to Table 1. Black symbols correspond to the category representation and are connected to the stimuli they merge, which are shown as white symbols. The two categories are shown by circles and squares. The bottom panel shows the posterior distribution over θ for data sets 5 (solid line), 7 (dotted line), and 4 (dashed line).

posterior mass, and most accounted for almost all of the mass. The data sets were selected because they demonstrated the ability of the model to infer a range of outcomes, starting with pure exemplar representations, and finishing with pure prototype representations. Between these extremes, there is evidence for partial abstraction in the form of a single merge of two similar stimuli (e.g., data sets 6, 7, 8, and 14), or in a form that more closely resembles multiple prototypes (e.g., data sets 9 and 13). The bottom panel of Fig. 5 shows the posterior marginal distribution for θ for data sets 5, 7, and 4 corresponding to exemplar, intermediate, and prototype representations. The posterior distributions clearly reflect these differences, giving more density to lower values of θ when there is less abstraction and more density to higher values of θ when there is more abstraction.

Still considering the exemplar versus prototype debate, Fig. 6 shows interesting results for data sets 11 and 15. Unlike the data sets shown in Fig. 5, no single VAM representation dominated the posterior for these data sets. Rather, a posterior distribution over possible representations was observed. The four most likely representations for data set 11 are shown,

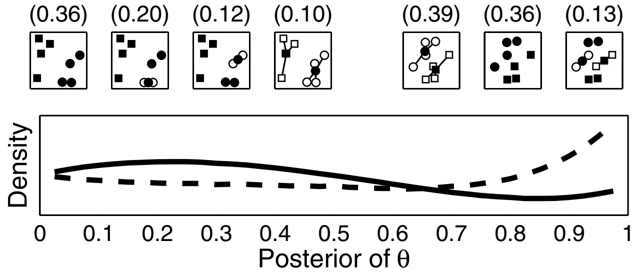

Fig. 6. The upper panels show the posterior distribution over category representations for data sets 11 (left 4 panels) and 15 (right 3 panels). The posterior mass for each representation is shown in brackets. The bottom panel shows the posterior distribution over θ for the data sets (11 = solid line, 15 = dashed line).

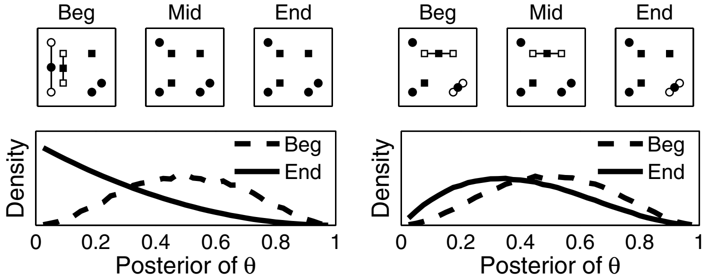

Fig. 7. The upper panels show the maximum *a posteriori* representation for two groups of three related data sets. Data sets 25, 26, and 27 are shown on the left, corresponding to the beginning (Beg), middle (Mid), and end blocks in the category learning task. Data sets 28, 29, and 30 are shown on the right. The bottom panels show the posterior distribution over θ for the beginning and end blocks in each case.

together with their posterior mass, on the left of Fig. 6, and the three most likely for data set 15 are shown on the right. It is clear that, for both data sets, there is uncertainty about the level of abstraction. Both the pure exemplar and pure prototype representations have significant mass in both cases, along with some intermediate possibilities. This uncertainty is represented in the posterior marginal distributions for θ, which have some multimodality, giving significant density to both low and high values.

Fig. 7 examines the issue of representational change over time (e.g., Johansen & Palmeri, 2002; Smith & Minda, 1998) by showing the representations and posterior distributions of θ inferred by the hierarchical VAM for two groups of related data sets. Data sets 25, 26, and 27 correspond to test performance from the beginning, middle, and end of one category learning task. Data sets 28, 29, and 30 correspond to test performance from the beginning, middle, and end of another category learning task. For each data set, the maximum a posteriori representation is shown and accounts for almost all of the posterior mass. It is suggestive that, for both groups, there is some evidence of a loss of abstraction in the form of a shift toward pure or near exemplar representation as testing progresses. The posterior marginal distributions for θ show this change, giving more density to low values of θ in the final data than for the beginning data.

5.2. *Similarity versus rule-based representations*

Data sets 17, 18, and 19 relate to the same category learning task, but involve different instructions given to three groups of participants. The first group was given no special instructions, and so was expected to learn the category structure using the similarity-based principles that underlie the GCM. The remaining two groups were instructed to use one of two simple rules accurately describing the category structure. Similarly, data sets 20, 21, and 22 involve one rule instruction condition, and two free conditions, for the same category learning task.

Fig. 8 shows the maximum a posteriori representations for each instruction condition for both groups of experiments. Once again, these representations accounted for almost all of the posterior mass. For the category learning task on the left, the free group have a VAM representation that does follow stimulus similarity, collapsing the similar stimuli in the

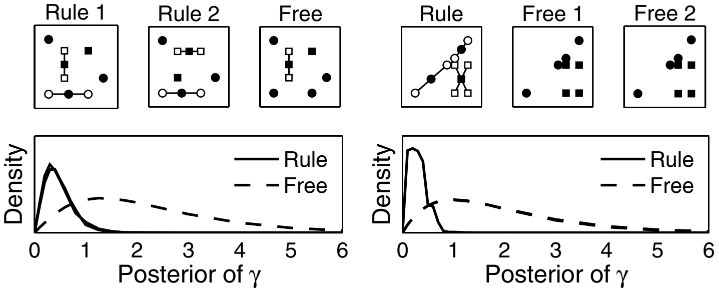

Fig. 8. The upper panels show the maximum *a posteriori* representation for two groups of three related data sets. Data sets 17, 18, and 19 are shown on the left, corresponding to the rule and free instructions conditions of the category learning task. Data sets 20, 21, and 22 are shown on the right. The bottom panels show the posterior distribution over δ for each data set (the distributions for the same instruction sets are so similar in each panel as to be indistinguishable).

interior category to a prototype, and largely preserving the less similar stimuli as exemplars in the exterior category. The groups given the rule instructions, however, do not follow stimulus similarity closely, especially through their merging of the same two dissimilar exterior stimuli. An examination of the rules used in instruction reveals that both had in common a logical proposition that directly corresponds to these two dissimilar stimuli, and so encouraged this merging. The γ parameter shows a lack of emphasis on stimulus similarity. Although this does not directly represent "rule-like" behavior, in this specific context it can appropriately be interpreted as corresponding to following rules.

The same analysis applies to the category learning task on the right, with the free instruction groups using exemplar representations, but the rule group merging stimuli. The posterior marginal distribution for γ, shown in the lower panel of Fig. 8 neatly distinguishes whether representations were similarity based, giving more density to larger values for the free groups and more density to values less than one for the rule group.

5.3. Other findings

One of the most important goals of model evaluation is not to provide definitive answers, but to sharpen the questions, and suggest new theoretical and empirical directions for future research. Our application of the hierarchical VAM suggested a number of such directions including modeling individual differences, unifying disparate representations, and suggesting different types of category representations. The relevant results motivating these ideas are summarized by the representations shown in Fig. 9, and we discuss them in turn.

The two left panels in Fig. 9, showing the inferred category representations for data sets 3 and 4, relate to two different participants performing exactly the same task. The representations strongly suggest the presence of individual differences, with one participant using a near-exemplar representation and the other using a pure prototype representation. This finding suggests that it may not be appropriate to aggregate category learning data over participants

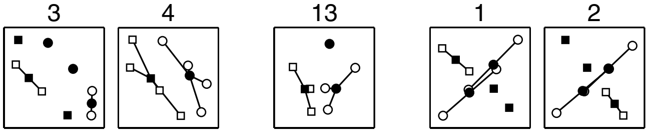

Fig. 9. Five Varying Abstraction Model representations, labeled by their data sets, all of which suggest extensions to our modeling approach. See text for details.

(Estes, 1956), and encourages the application of models, including hierarchical Bayesian ones, that attempt to account for individual differences in cognitive processes (Lee & Webb, 2005)

The final three panels in Fig. 9 show representations that suggest limitations in the VAM, and the hierarchical approach to its analysis, that we have used. The inferred representation for data set 13 seems naturally interpreted as a prototype-plus-outlier approach. The VAM class of representations can express this type of representation, but a more focused analysis is probably required to give it theoretical attention. In particular, it is likely that the joint posterior of θ and γ needs to be analyzed because prototype-plus-outlier representations will require simultaneously high values of both θ to generate the prototype and γ to leave dissimilar stimuli as exemplar outliers.

Finally, the inferred representations for data sets 1 and 2 in Fig. 9 suggest a more basic deficiency in the class of VAM representations. For the category shown by circles, there are several unusual merges involved in these representations, combining pairs of very dissimilar stimuli. The result of these merges, however, is highly interpretable, as forming prototypes in the interior of the categories and ignoring the exterior stimuli altogether. In other words, the final category representations are ones that would follow from deleting the external stimuli and retaining the internal ones (i.e., the dissimilar stimuli that are merged result in a prototype that is essentially the same as one of the original exemplars). The merging process that generates VAM representations does not allow deletions, but these analyses present strong suggestive evidence in favor of such a process for forming category representations. It is only a fortunate coincidence of the geometry of the original stimuli, and the category structures that had to be learned, that the VAM representations based on merging are able to approximate the result of a deletion process in a different, and far less interpretable, way. The net result of this analysis is to suggest it will be useful to consider a new and different approach to forming category representations, which will result in using something other than the VAM class of representations as the basic units of analysis.

6. Discussion

It was not our goal in this article to reach definitive conclusions regarding whether people use exemplars or prototypes to represent categories, or whether they rely on similarities or rules to form their representations. These questions almost certainly do not have single answers, and even attempts to address the issues more generally will produce results that depend heavily on the theoretical assumptions made and the nature of the empirical evidence used. Our goal in this article was to provide a detailed demonstration of how hierarchical Bayesian methods can be used to extend and analyze cognitive models, and play a constructive and powerful role in investigating a basic question for cognitive science like, "How do people represent categories?"

One important strength of the hierarchical Bayesian approach was that it allows inferences about model parameters to correspond directly to information about the motivating psychological questions. The posterior distributions for θ are readily interpreted as measures of the extent of abstraction, and the posterior distributions for γ are readily interpreted as measures of the reliance on similarity in forming category representations. A natural extension of this line of analysis would be to consider data from many category learning experiments simultaneously. Different experimental designs will require different sets of VAM representations because the number of stimuli and their spatial locations will vary, but the more abstract θ and γ parameters will remain commensurable across all possible sets of VAM representations. In this way, hierarchical Bayesian analysis provides a promising avenue for using new data to update existing evidence in a more formal way than is currently done.

The other contribution of the hierarchical Bayesian approach is that, by specifying a process for generating VAM representations, together with sensible priors on the parameters that control this process, an inductive bias is automatically imposed on the VAM class of representations. This corresponds to specifying theoretically grounded priors for the different representations that are being compared. One interesting consequence of the sorts of priors derived from the merge process is that they potentially allow the VAM approach to be scaled to large numbers of stimuli. The number of possible VAM representations grows very rapidly, according to what are known as Bell numbers. For 100 stimuli, the number of representations is greater than 10^{116}. Fig. 4 suggests, however, that the prior over these representations could be approximated well by considering just a small subset of the models with non-negligible prior probability, as in Bayesian model averaging (Hoeting, Madigan, Raftery, & Volinsky, 1999), making analysis computationally feasible.

Of course, the specific results we report would be different if we used a different generating process for the VAM representations, or put very different priors on the generating parameters. This is unavoidable, and should be seen as desirable. It is not possible to draw inferences without making assumptions. Our hierarchical approach at least forces the assumptions to derive from explicit and theoretically motivated assumptions about how category representations might be formed. We think any reasonable theory of this type, including our first attempt described here, will lead to better priors than the uniform ones (i.e., that all representations are equally likely) previously assumed.

More fundamentally, the whole hierarchical Bayesian exercise serves to shift theoretical attention to the basic question of how category representations are generated. We have found evidence for processes involving deletion that were not originally considered, and imply a class of category representations not captured by the VAM. If ones of the goals of modeling is to find gaps in existing theories, and suggest approaches for improvement, then our modeling has served us well.

We think the advantages of the hierarchical Bayesian approach evident in our application to category representation will be true for many areas of modeling in cognitive science. There are many current debates that would benefit from being re-cast as problems of estimation along a dimension, rather than choosing between extremes. One possible example is the tension between random-walk and accumulator versions of sequential sampling processes (e.g., Ratcliff, 1978; Ratcliff & Smith, 2004; Vickers, 1979). Another is the tension between "one-reason" and "rational" accounts of decision making (e.g., Lee & Cummins, 2004; Newell, 2005).

Most generally, we think the ability for models to operate at many levels, and relate these various levels all the way down to observed data, is a crucially important one, with implications that go beyond what we have been able to demonstrate. It drives theoretical questions to ever more fundamental levels, and demands that formal models be developed and evaluated at these deeper levels. It offers a possibility to model individual differences, rather than just observing and interpreting parametric variation in cognitive processes, by requiring models be developed at the level where parametric variation captures individual differences. Hierarchical modeling also allows the possibility of integrating relevant evidence across multiple psychological tasks, with the higher levels corresponding to a constant psychological construct of interest, but lower levels giving the details of how those constructs generate data for specific tasks. These sorts of applications of hierarchical Bayesian analysis represent important and exciting areas of future research.

Acknowledgments

We thank Art Markman and several anonymous reviewers for many helpful comments. Wolf Vanpaemel was sponsored by Grant OT/05/27 of the Leuven Research Council and by Grant G.0513.08 of the Belgian Science Foundation.

Notes

1. A more detailed version of the analysis, in the form of a graphical model representation, can be found in Vanpaemel and Lee (2007).
2. All of our analysis in this article relates to two-category tasks, with stimuli represented using two dimensions. Generalization to more categories and dimensions are both straightforward.

References

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

Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences, 10*, 287–291.

Cox, R. T. (1961). *The algebra of probable inference.* Baltimore: Johns Hopkins University Press.

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin, 53*, 134–140.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.

Griffiths, T. L., Canini, K. R., Sanborn, A. N., & Navarro, D. J. (2007). Unifying rational models of categorization via the hierarchical Dirichlet process. In D. J. McNamara & J. G. Trafton (Eds.), *Proceedings of the 29th annual conference of the cognitive science society* (pp. 323–328). Austin, TX: Cognitive Science Society.

Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (2008). Bayesian models of cognition. In R. Sun (Ed.), *Cambridge handbook of computational cognitive modeling* (pp. 59–100). Cambridge, MA: Cambridge University Press.

Hoeting, J., Madigan, D., Raftery, A., & Volinsky, C. (1999). Bayesian model averaging. *Statistical Science, 14*, 382–401.

Jaynes, E. T. (2003). *Probability theory: The logic of science.* New York: Cambridge University Press.

Johansen, M. K., & Palmeri, T. J. (2002). Are there representational shifts during category learning? *Cognitive Psychology, 45*, 482–553.

Komatsu, L. K. (1992). Recent views of conceptual structure. *Psychological Bulletin, 112*, 500–526.

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

Lee, M. D. (2004). An efficient method for the minimum description length evaluation of cognitive models. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the 26th annual conference of the Cognitive Science Society* (pp. 807–812). Mahwah, NJ: Lawrence Erlbaum Association, Inc.

Lee, M. D. (2008). Three case studies in the Bayesian analysis of cognitive models. *Psychonomic Bulletin & Review, 15*, 1–15.

Lee, M. D., & Cummins, T. D. R. (2004). Evidence accumulation in decision making: Unifying the "take the best" and "rational" models. *Psychonomic Bulletin & Review, 11*, 343–352.

Lee, M. D., & Wagenmakers, E. J. (2005). Bayesian statistical inference in psychology: Comment on Trafimow (2003). *Psychological Review, 112*, 662–668.

Lee, M. D., & Webb, M. R. (2005). Modeling individual differences in cognition. *Psychonomic Bulletin & Review, 12*, 605–621.

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

Mackay, D. J. C. (2003). *Information theory, inference, and learning algorithms.* Cambridge, England: Cambridge University Press.

Myung, I. J., Balasubramanian, V., & Pitt, M. A. (2000). Counting probability distributions: Differential geometry and model selection. *Proceedings of the National Academy of Sciences USA*, *97*, 11170–11175.

Myung, I. J., Forster, M., & Browne, M. W. (2000). A special issue on model selection. *Journal of Mathematical Psychology*, *44*, 1–2.

Newell, B. R. (2005). Re-visions of rationality. *Trends in Cognitive Sciences, 9*, 11–15.

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

Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *115*, 87–108.

Nosofsky, R. M. (1989). Further tests of an exemplar-similarity approach to relating identification and categorization. *Perception & Psychophysics*, *45*, 279–290.

Nosofsky, R. M. (1992). Exemplars, prototypes and similarity rules. In A. F. Healy, S. M. Kosslyn, & R. M. Shiffrin (Eds.), *From learning theory to connectionist theory: Essays in honour of William K. Estes* (Vol. 1, pp. 149–167). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Nosofsky, R. M., Clark, S. E., & Shin, H. J. (1989). Rules and exemplars in categorization, identification, and recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 282–304.

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

Nosofsky, R. M., & Palmeri, T. J. (1998). A rule-plus-exception model for classifying objects in continuous-dimension spaces. *Psychonomic Bulletin & Review*, *5*, 345–369.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review, 109*, 472–491.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*, 59–108.

Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review*, *111*, 333–367.

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

Shiffrin, R. M., Lee, M. D., Wagenmakers, E. J., & Kim, W. J. (this issue). A survey of evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science, 32*, 1248–1284.

Sivia, D. S. (1996). *Data analysis: A Bayesian tutorial.* Oxford, England: Clarendon.

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.

Smith, J. D., & Minda, J. P. (2000). Thirty categorization results in search of a model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 3–27.

Vanpaemel, W., & Lee, M. D. (2007). A model of building representations for category learning. In D. J. McNamara & J. G. Trafton (Eds.), *Proceedings of the 29th annual conference of the Cognitive Science Society* (pp. 1605–1610). Austin, TX: Cognitive Science Society.

Vanpaemel, W., & Navarro, D. J. (2007). Representational shifts during category learning. In D. J. McNamara & J. G. Trafton (Eds.), *Proceedings of the 29th annual conference of the Cognitive Science Society* (p. 1599–1604). Austin, TX: Cognitive Science Society.

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 B. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the 27th annual conference of the Cognitive Science Society* (pp. 2277–2282). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Vickers, D. (1979). *Decision processes in visual perception.* New York: Academic.

Appendix: Some details on the inference

As noted in the main text, our primary interest focuses on the inferences that can be drawn from two posterior distributions: the representation posteriors $p(x \mid D)$, which describe the inferences made by the model about what VAM representation is being used; and the marginalized parameter posteriors $p(\theta \mid D)$ and $p(\gamma \mid D)$, which describe inferences about what process people used to generate that representation, informing us about the use of abstraction and the importance of similarity, respectively.

Representation posteriors

The representation posterior $p(x \mid D)$ is the probability of the xth representation being true given data D. It is computed by combining the likelihood of the data under the representation with the probability of the representation being true before any data are collected (i.e., its inductive bias) using Bayes' rule:

$$p\left(x \mid D\right) = \frac{p\left(D \mid x\right) p\left(x\right)}{p\left(D\right)}. \tag{9}$$

We discuss the three individual components of the right-hand side in turn.

First, the marginal likelihood $p\,(D \mid x)$ is defined as

$$p(D \mid x) = \iint p(D \mid c, w, x)p(c, w)\mathrm{d}c\ \mathrm{d}w$$
$$= \iint p(D \mid c, w, x)p(c)p(w)\mathrm{d}c\ \mathrm{d}w, \tag{10}$$

where $p\,(D \mid c, w, x)$ is the likelihood, and $p\,(c)$ and $p\,(w)$ are the category representation level parameter priors given in the main text.

Second, the representation prior $p(x)$ is defined as

$$p(x) = \iint p(x \mid \theta, \gamma)p(\theta, \gamma)\ \mathrm{d}\theta\ \mathrm{d}\gamma$$
$$= \iint p(x \mid \theta, \gamma)p(\theta)\ p(\gamma)\ \mathrm{d}\theta\ \mathrm{d}\gamma, \tag{11}$$

where $p(x \mid \theta, \gamma)$ is the merge distribution, and $p(\theta)$ and $p(\gamma)$ are the generative level parameter priors given in the main text.

The merge distribution is defined by Monte Carlo estimates of $p(x \mid \theta, \gamma)$, found by simulating the iterative process over the stimuli and category structures used in the applications across the grid $\theta = (0.025, 0.05, \ldots, 0.975)$ and $\gamma = (0, 0.1, \ldots, 10)$.

Finally, the probability of the data $p(D)$ serves as a normalizing constant to ensure that $\sum_x p(x \mid D) = 1$. To compare models, the posterior mass or the relative posteriors are used:

$$h_x = \frac{p\,(x \mid D)}{\sum_y p\,(y \mid D)}$$
$$= \frac{p\,(D \mid x)\,p\,(x)}{\sum_y p\,(D \mid y)\,p\,(y)}. \tag{12}$$

Because $p\,(D)$ is the same for all models, it does not need to be evaluated.

Parameter posteriors

As for the representation posteriors, the parameter joint posterior $p\,(\theta, \gamma \mid D)$ is computed using Bayes' rule:

$$p\,(\theta, \gamma \mid D) = \frac{p\,(D \mid \theta, \gamma)\,p\,(\theta, \gamma)}{p\,(D)}$$
$$= \frac{p\,(D \mid \theta, \gamma)\,p\,(\theta)\,p\,(\gamma)}{p\,(D)}. \tag{13}$$

The parameter marginal $p\,(D \mid \theta, \gamma)$ is defined as

$$p\,(D \mid \theta, \gamma) = \sum_x p\,(D \mid x)\,p\,(x \mid \theta, \gamma), \tag{14}$$

where the marginal likelihood $p\,(D \mid x)$ is given by Equation 10 and $p\,(x \mid \theta, \gamma)$ is again the merge distribution. From the parameter joint posterior, the parameter marginal posteriors

$p\,(\theta \mid D)$ and $p\,(\gamma \mid D)$ are computed by marginalizing:

$$p\,(\theta \mid D) = \int p\,(\theta, \gamma \mid D)\ \mathrm{d}\gamma, \tag{15}$$

and

$$p(\gamma \mid D) = \int p\,(\theta, \gamma \mid D)\ \mathrm{d}\theta. \tag{16}$$