Behavior Research Methods
2010, 42 (2), 421-437
doi:10.3758/BRM.42.2.421

ARTICLES

Abstraction and model evaluation in category learning

Wolf Vanpaemel and Gert Storms
University of Leuven, Leuven, Belgium

Thirty previously published data sets, from seminal category learning tasks, are reanalyzed using the varying abstraction model (VAM). Unlike a prototype-versus-exemplar analysis, which focuses on extreme levels of abstraction only, a VAM analysis also considers the possibility of partial abstraction. Whereas most data sets support no abstraction when only the extreme possibilities are considered, we show that evidence for abstraction can be provided using the broader view on abstraction provided by the VAM. The present results generalize earlier demonstrations of partial abstraction (Vanpaemel & Storms, 2008), in which only a small number of data sets was analyzed. Following the dominant modus operandi in category learning research, Vanpaemel and Storms evaluated the models on their best fit, a practice known to ignore the complexity of the models under consideration. In the present study, in contrast, model evaluation not only relies on the maximal likelihood, but also on the marginal likelihood, which is sensitive to model complexity. Finally, using a large recovery study, it is demonstrated that, across the 30 data sets, complexity differences between the models in the VAM family are small. This indicates that a (computationally challenging) complexity-sensitive model evaluation method is uncalled for, and that the use of a (computationally straightforward) complexity-insensitive model evaluation method is justified.

Abstraction in Category Learning

A classic issue in cognitive science is the use of abstraction in category representations. Prototype theorists (e.g., Posner & Keele, 1968; Reed, 1972; Smith & Minda, 2000) argue that learning a category involves storing one summary item that is abstracted across all category examples. Exemplar theorists (e.g., Brooks, 1978; Medin & Schaffer, 1978; Nosofsky, 1986), in contrast, argue that learning a category involves learning all the individual category examples, which implies that no abstraction takes place when a category is learned.

In the debate between the prototype and exemplar views, there is a clear tension between informativeness and cognitive economy (Komatsu, 1992). A prototype representation has an appealing economy, but fails to provide the information people appear to use (e.g., Fried & Holyoak, 1984). In contrast, an exemplar representation provides detailed information, but its lack of any cognitive economy makes it somewhat counterintuitive.

In an attempt to resolve this tension, Vanpaemel and Storms (2008) introduced the varying abstraction model (VAM; see also Vanpaemel, Storms, & Ons, 2005; for related but different approaches, see Anderson, 1991; Griffiths, Canini, Sanborn, & Navarro, 2007; Love, Medin, & Gureckis, 2004). The VAM considers the prototype and exemplar representations as endpoints only on the continuum of abstraction and introduces a range of category representations between these two possibilities. These intermediate representations balance the opposing pressures of informativeness and economy, providing just enough representational information to describe the category structure in a sufficiently complete way.

Intermediate representations, despite their intuitive appeal, have been largely overlooked in previous research on abstraction in category learning. Focusing on four previously published data sets, which in their original analysis provided no evidence for abstraction, Vanpaemel and Storms (2008) reported preliminary explorations of partial abstraction. Using the VAM, no evidence for partial abstraction could be found in one data set, but based on the three remaining data sets, it could be concluded that some abstraction had occurred during category learning.

One objection that may be raised against the study of Vanpaemel and Storms (2008) is its lack of generality. Compared with the impressively large number of studies supporting exemplar theory (see Nosofsky, 1992, for an overview), the weight of the evidence for partial abstraction is somewhat meager, given that it has only been demonstrated in three different data sets. Therefore, one purpose of the present research was to continue the inves-

W. Vanpaemel, wolf.vanpaemel@psy.kuleuven.be

© 2010 The Psychonomic Society, Inc.

tigation of partial abstraction and to extend the generality of the findings of Vanpaemel and Storms by considering a larger number of data sets, spanning a broad variety of conditions. In particular, the present study focuses on 30 different data sets, all previously presented by Nosofsky and his collaborators, all of which have been previously interpreted as providing evidence in favor of the absence of abstraction.[1]

Model Evaluation in Category Learning

The present research involves two other extensions of the earlier work reported by Vanpaemel and Storms (2008). Both extensions concern a methodological objection that can be raised against the study by Vanpaemel and Storms: Following the dominant model evaluation strategy in the prototype versus exemplar debate, they evaluated the different models in the VAM family on their optimal fit to human data. Evaluating models on their optimal fit, however, ignores differences in model complexity, which is the inherent flexibility of a model. It has been argued that ignoring model complexity results in an unfair evaluation, since a more complex model generally provides a better optimal fit than a simple model does. Therefore, a model evaluation method sensitive to model complexity should be used (e.g., Cutting, Bruno, Brady, & Moore, 1992; Myung, Forster, & Browne, 2000; Myung & Pitt, 1997; Pitt & Myung, 2002; Roberts & Pashler, 2000; Shiffrin, Lee, Kim, & Wagenmakers, 2008; Waldorp & Wagenmakers, 2006).

The simplest way to measure model complexity is to count the number of parameters. This idea of complexity lies at the root of model evaluation methods such as the likelihood ratio test, the Akaike information criterion (AIC; Akaike, 1973) and the Bayesian information criterion (BIC; Schwarz, 1978). Although taking the number of parameters into account when evaluating a model presents an important advance in model evaluation, it is sensitive to only one aspect of complexity. In particular, the range of the parameters, the parameter prior, and the functional form (the way the parameters are combined) also play a role in model complexity (e.g., Myung & Pitt, 1997; Vanpaemel, 2009b).

One model evaluation method sensitive to all these aspects of complexity is the marginal likelihood, which is at the heart of the Bayesian approach to model evaluation (Kass & Raftery, 1995; Lee & Wagenmakers, 2005; Myung & Pitt, 1997). Rather than the optimal fit, the marginal likelihood measures a model's average fit. Unfortunately, more than a decade after its introduction in psychology (but see Edwards, Lindman, & Savage, 1963, for an early example), there are only a handful of studies relying on the marginal likelihood for model evaluation (e.g., Lee, 2004; Massaro, Cohen, Campbell, & Rodriguez, 2001; Navarro & Lee, 2003; Steyvers, Lee, & Wagenmakers, 2009). Furthermore, we know of no study that has used the marginal likelihood to evaluate prototype and exemplar models against empirical data, let alone the intermediate models in the broader VAM family. Instead, these studies have relied on best fit statistics, such as the maximal likelihood and the percentage of variance accounted for. The present investigation is, to our knowledge, the first to use the marginal likelihood to evaluate the prototype model, the exemplar model, and the other models in the VAM family against empirical data.

Probably the most important reason the marginal likelihood is so often talked about but rarely applied is the technical sophistication required to calculate it, especially in contrast to the rather straightforward calculation of the maximal likelihood. Therefore, a question of interest is whether the marginal likelihood is needed to reach trustable conclusions. Obviously, if complexity differences between models are small, a complexity-insensitive method (such as the maximal likelihood) will perform just as well as a complexity-sensitive method (such as the marginal likelihood). In this case, the method that is equally successful, but easier to compute, is the preferred one. A second purpose of the present research was to investigate the model complexity differences between the models in the VAM family, using a large recovery study. This affords insight into the relative utilities of complexity-sensitive versus -insensitive model evaluation methods in a VAM analysis.

Outline

We begin by describing the 30 data sets and reviewing the VAM. After the two model evaluation methods have been explained, the VAM is applied to the 30 data sets, using both the maximal and the marginal likelihood. The VAM analysis provides further evidence for partial abstraction in category representations, irrespective of the model evaluation method. Finally, the recovery study indicates that all members of the VAM family are, roughly, equally complex, thus justifying the use of complexity-insensitive model evaluation methods in a VAM analysis.

The 30 Data Sets

We consider 30 previously published data sets from different seminal category learning tasks, reported in six articles by Nosofsky and his collaborators (Nosofsky, 1986, 1987, 1989; Nosofsky, Clark, & Shin, 1989; Nosofsky & Palmeri, 1997, 1998). All tasks involve learning various two-category structures over a small number of stimuli varying on two dimensions. In each task, a subset of the stimuli is assigned to Categories A and B, and the remaining stimuli are left unassigned. In most tasks, the *training-test* procedure is used, consisting of a training phase followed by a test phase. During the training phase, the category structure is learned. Only the assigned stimuli are presented and the participant classifies each presented stimulus into either Category A or Category B. Following each response, corrective feedback is presented. During the test phase, both the assigned and the unassigned stimuli are presented. Generally, corrective trial-by-trial feedback continues to be provided on trials in which assigned stimuli are presented, but is withheld on trials in which unassigned stimuli are presented, because there are no correct or incorrect answers for these trials. The relevant data for modeling are the categorization decisions from the test phase.

Some properties of the 30 data sets are shown in Table 1. A more detailed description of the experimen-

Table 1
Details of the 30 Data Sets

Data Set	Reference	Experiment	Stimuli	r	α	Categories	k	Condition	N	n	s
1	Nosofsky (1986)	–	Shepard circles	2	2	Criss-cross	225	Participant 1	150[m]	219[a]	1
2	Nosofsky (1986)	–	Shepard circles	2	2	Criss-cross	225	Participant 2	150[m]	225[a]	1
3	Nosofsky (1986)	–	Shepard circles	2	2	Diagonal	225	Participant 1	150[m]	250[a]	1
4	Nosofsky (1986)	–	Shepard circles	2	2	Diagonal	225	Participant 2	150[m]	225[a]	1
5	Nosofsky (1986)	–	Shepard circles	2	2	Dimensional	225	Participant 1	150[m]	225[a]	1
6	Nosofsky (1986)	–	Shepard circles	2	2	Dimensional	225	Participant 2	150[m]	200[a]	1
7	Nosofsky (1986)	–	Shepard circles	2	2	Interior–exterior	225	Participant 1	150[m]	256[a]	1
8	Nosofsky (1986)	–	Shepard circles	2	2	Interior–exterior	225	Participant 2	150[m]	144[a]	1
9	Nosofsky (1987)	2	Munsell colors	2	1	Brightness	780	–	–	10	49
10	Nosofsky (1987)	2	Munsell colors	2	1	Criss-cross	225	–	–	10	24
11	Nosofsky (1987)	2	Munsell colors	2	1	Saturation A	225	–	–	7.5[a]	24
12	Nosofsky (1987)	2	Munsell colors	2	1	Saturation B	225	–	–	10	40
13	Nosofsky (1989)	–	Shepard circles	2	2	Angle	75	–	46[a]	4[a]	41
14	Nosofsky (1989)	–	Shepard circles	2	2	Criss-cross	225	–	78[a]	4[a]	37
15	Nosofsky (1989)	–	Shepard circles	2	2	Diagonal	225	–	78[a]	4[a]	43
16	Nosofsky (1989)	–	Shepard circles	2	2	Size	75	–	46[a]	4[a]	37
17	Nosofsky, Clark, & Shin (1989)	1	Shepard circles	1	1	Interior–exterior	75	Free	43[a]	5	122
18	Nosofsky et al. (1989)	2	Shepard circles	1	1	Interior–exterior	75	Rule 1	43[a]	5	30
19	Nosofsky et al. (1989)	2	Shepard circles	1	1	Interior–exterior	75	Rule 2	43[a]	5	28
20	Nosofsky et al. (1989)	3	Pitch and line length	1	1	Conjunctive rule	225	Free 1	31[a]	5	30
21	Nosofsky et al. (1989)	3	Pitch and line length	1	1	Conjunctive rule	225	Free 2	31[a]	5	30
22	Nosofsky et al. (1989)	3	Pitch and line length	1	1	Conjunctive rule	225	Rule	19[a]	5	30
23	Nosofsky & Palmeri (1997)	2	Munsell colors	2	1	U7	75	–	19	8	31
24	Nosofsky & Palmeri (1997)	2	Munsell colors	2	1	U8	75	–	19	8	31
25	Nosofsky and Palmeri (1998)	1	Shape and brightness	1	1	Interior–exterior	75	Beginning blocks	10	3	164
26	Nosofsky & Palmeri (1998)	1	Shape and brightness	1	1	Interior–exterior	75	Middle blocks	20	3	164
27	Nosofsky & Palmeri (1998)	1	Shape and brightness	1	1	Interior–exterior	75	End blocks	40	3	164
28	Nosofsky & Palmeri (1998)	2	Shape and brightness	1	1	Interior–exterior	75	Beginning blocks	5	1	120
29	Nosofsky & Palmeri (1998)	2	Shape and brightness	1	1	Interior–exterior	75	Middle blocks	10	1	120
30	Nosofsky & Palmeri (1998)	2	Shape and brightness	1	1	Interior–exterior	75	End blocks	20	3	120

Note—r, metric; α, similarity function; k, number of models in the VAM; N, number of trials per stimulus per participant in the training phase; n, number of trials per stimulus per participant in the test phase; s, number of participants; [m]minimal value (actual value unreported); [a]approximate average value.

tal procedures can be found in the original studies or in Vanpaemel (2007). Collectively, the 30 data sets span a range of possibilities for category learning tasks. They involve 16 different category structures to be learned, shown in Figure 1. Further, they involve different types of stimuli, varying from simple geometric shapes, to colors, to combinations of visual and auditory stimuli. The data sets also vary in the number of participants, training trials, and test trials. Several of the tasks involve a variety of instruction conditions, including participants being told to follow rules. Some also report data at different stages in the learning sequence. The actual data are provided in Tables A1, A2, and A3 of Appendix A, reproducing the stimulus coordinates, the Category A response frequencies, and the sample sizes, respectively.

Review of the Varying Abstraction Model

At the heart of the VAM is the observation that the prototype and exemplar representations correspond to maximal abstraction and minimal abstraction, respectively. These extreme possibilities suggest a set of intermediate representations with a partial level of abstraction. Such an intermediate representation consists of a set of subprototypes formed by abstracting over mutually exclusive clusters of category members. This is illustrated in Figure 2, which provides the full set of representations defined by the VAM for a category with four members in a two-dimensional space.

Based on this set of representations, the VAM defines a family of categorization models. Such a model is con-

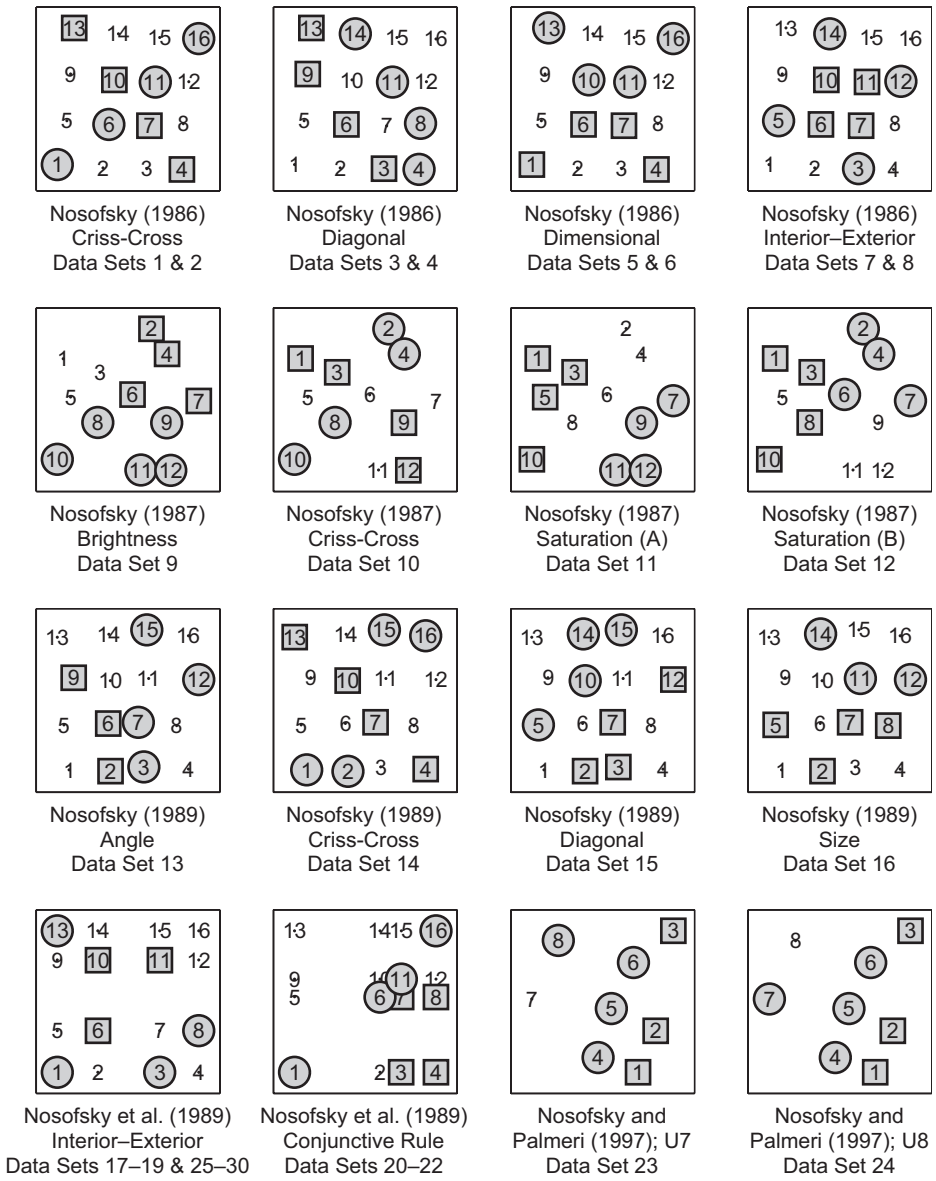


Figure 1. Category structures used in the 30 data sets. Squares denote stimuli assigned to Category A, circles denote stimuli assigned to Category B. The remaining stimuli are unassigned.

structed by choosing a representation for each of the relevant categories and combining these representations with the assumptions about similarity, selective attention, and response rule incorporated in Nosofsky's (1986) generalized context model (GCM), arguably the most successful exemplar model of category learning. By design then, all the models in the VAM contrast only in their representational assumptions.

In what follows, the VAM is formally outlined for stimuli varying on two dimensions in a category learning task with two categories, A and B. The more general formulation of the VAM can be found in Vanpaemel and Storms (2008).

Let $x_i = (x_{i1}, x_{i2})$ denote the coordinate locations of stimulus x_i in a two-dimensional space, as found by previous multidimensional scaling (MDS) analysis. Similarly, let $\mu_j = (\mu_{j1}, \mu_{j2})$ denote the coordinate locations of subprototype μ_j. These subprototypes are defined by partitioning the set of category members and abstracting the centroid for each cluster. Formally, if $\mathcal{Q} = \{Q_1, Q_2, \ldots, Q_q\}$ denotes a partition with q clusters of a category with n members and n_j ($j = 1, 2, \ldots, q$; $1 \leq n_j \leq n$) denotes the number of category members in cluster Q_j, subprototype μ_j is defined by

$$\mu_{jk} = \frac{1}{n_j} \sum_{x_l \in Q_j} x_{lk}. \tag{1}$$

Given the coordinates of the stimuli and the subprototypes, the distance between stimulus x_i and subprototype μ_j is computed according to the Minkowski r-metric:

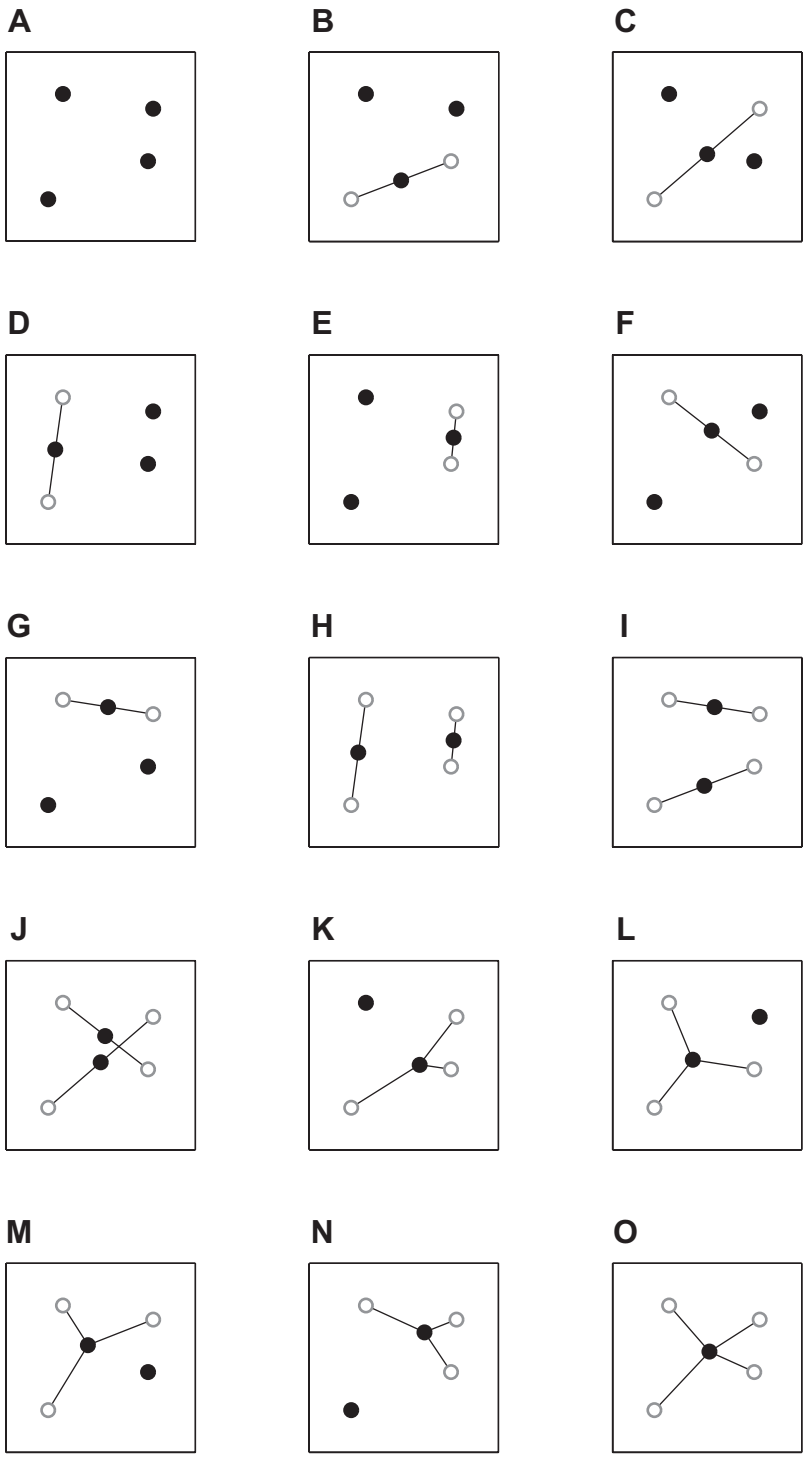


Figure 2. The 15 possible VAM representations for a category with four members in a two-dimensional space. The subprototypes representing the category are shown in black circles and are connected by lines to the original category members, shown in white circles. The first panel shows the exemplar representation, where no category members are merged; the final panel shows the prototype representation, where all four category members are merged into one single item. The remaining panels show the representations intermediate to the exemplar and prototype ones, in which a category is represented by three (panels B–G) or two (panels H–N) subprototypes.

$$d\left(x_i, \mu_j\right)^r = w\left(x_{i1} - \mu_{j1}\right)^r + \left(1 - w\right)\left(x_{i2} - \mu_{j2}\right)^r, \quad (2)$$

where w is the attention weight parameter measuring the relative emphasis given to the first dimension over the second, and r is the metric.

Similarity is then modeled as an exponentially decaying function of the distance between the stimuli and subprototypes:

$$s\left(x_i, \mu_j\right) = \exp\left\{-\left(cd\left(x_i, \mu_j\right)\right)^\alpha\right\}, \quad (3)$$

where c is the sensitivity parameter measuring the rate at which similarity declines with distance, and α determines the shape of the function.

The similarity between stimulus x_i and Category A is defined by the sum of the similarities between the stimulus and all the subprototypes representing the category:

$$s\left(x_i, \mathrm{A}\right) = \sum_{\mu_j \in S_\mathrm{A}} s\left(x_i, \mu_j\right), \quad (4)$$

where $S_\mathrm{A} = \{\mu_1, \mu_2, \ldots, \mu_{q_\mathrm{A}}\}$, the set of q_A subprototypes representing Category A. Finally, the response probability for stimulus x_i's being chosen as a member of Category A is computed according to the choice rule

$$p\left(\mathrm{A} \mid x_i\right) = \frac{s\left(x_i, \mathrm{A}\right)}{s\left(x_i, \mathrm{A}\right) + s\left(x_i, \mathrm{B}\right)}. \quad (5)$$

Typically, r and α depend on the nature of the stimuli, and are therefore not considered free parameters. Generally, the city-block metric ($r = 1$) is used when stimuli vary on separable dimensions, and the Euclidean metric ($r = 2$) is appropriate for stimuli varying on integral dimensions (see Shepard, 1991, for a review). Furthermore, the exponential decay function ($\alpha = 1$) is used when the stimuli are readily discriminable, whereas the Gaussian function ($\alpha = 2$) is typically preferred when the stimuli are highly confusable (Shepard, 1987). The remaining two parameters—the sensitivity c and the attention weight w—are free.

A model in the VAM is specified by $S_\mathrm{A} = \{\mu_1, \mu_2, \ldots, \mu_{q_\mathrm{A}}\}$ (i.e., the set of q_A subprototypes representing Category A) and $S_\mathrm{B} = \{\mu_1, \mu_2, \ldots, \mu_{q_\mathrm{B}}\}$ (i.e., the set of q_B subprototypes representing Category B). The VAM encompasses all the models that can be constructed by combining all the possible representations of each category. In particular, for Categories A and B having n_A and n_B category members, respectively, the VAM generates $bell(n_\mathrm{A}) \times bell(n_\mathrm{B})$ different models, where $bell(n)$ denotes the nth Bell number, which counts the number of partitions of a set of n members (Sloane, 2003). The numbers of models in the VAM generated by the designs considered in this article, being 75, 225, or 780, are shown in Table 1.

A critical aspect of the VAM is that, at extreme settings of q_A and q_B, it reduces exactly to the GCM and its theoretical rival, the MDS-based prototype model (MPM; Nosofsky, 1987, 1992; Reed, 1972). Consider a Category A with n_A members. At one extreme, when $q_\mathrm{A} = 1$, the category is partitioned in only one cluster of n_A members. Since $Q_1 = \mathrm{A}$ and $n_1 = n_\mathrm{A}$, Equation 1 reduces to $\mu_{1k} = (1/n_\mathrm{A}) \Sigma_{x_l \in A} x_{lk}$, which is the familiar definition of the Category A prototype π_A, so $\mu_1 = \pi_\mathrm{A}$, and thus $S_\mathrm{A} = \{\pi_\mathrm{A}\}$. Hence, Equation 4 reduces to

$$s\left(x_i, \mathrm{A}\right) = s\left(x_i, \pi_\mathrm{A}\right), \quad (6)$$

which is the key definition of the MPM.

At the other extreme, when $q_\mathrm{A} = n_\mathrm{A}$, the category is partitioned in n_A clusters of one member each. Since $Q_j = \{x_j\}$ and $n_j = 1$ for every j, Equation 1 reduces to $\mu_{jk} = x_{jk}$. This implies that $\mu_j = x_j$ for every j, and thus $S_\mathrm{A} = \mathrm{A}$. Hence, Equation 4 reduces to

$$s\left(x_i, \mathrm{A}\right) = \sum_{x_j \in \mathrm{A}} s\left(x_i, x_j\right), \quad (7)$$

which is the defining feature of the GCM.

Model Evaluation Methods

An intuitively attractive way to evaluate a model is to assess whether it is consistent with the empirical data. This consistency—or *match* or *fit*—between a model's predictions and the empirical data is often expressed using the likelihood function (e.g., Myung, 2003). Crucially, parameterized models, such as the GCM and the MPM, make different predictions at different parameter values, which implies that the likelihood function depends on the parameter values. To assess the fit between model and data, one therefore has to answer the question: the fit at which parameter value(s)? Although many different answers to this question have been advocated, we will focus on the two dominant ones.

One commonly used strategy is to measure the fit at one single parameter value. When evaluating a model at a single value only, it seems reasonable to pick the value at which the fit is highest. This is the idea behind evaluating a model on its maximal likelihood. By design, the maximal likelihood reflects the best possible fit of the model. If $p(d \mid \theta, M)$ denotes the likelihood under model M of having observed data d given parameter values θ, then $p(d \mid \hat{\theta}, M)$ denotes the maximal likelihood of the model, obtained by finding the value of θ that maximizes $p(d \mid \theta, M)$:

$$p\left(d \mid \hat{\theta}, M\right) = p\left(d \mid \theta = \hat{\theta}, M\right) = \max_\theta p\left(d \mid \theta, M\right), \quad (8)$$

where

$$\hat{\theta} = \arg\max_\theta p\left(d \mid \theta, M\right). \quad (9)$$

To find the parameter value at which to assess the model's fit under the maximum likelihood approach, one needs access to the data. So in a sense, the prediction the model makes, at the value $\hat{\theta}$, is not a prediction in the strict sense, but rather a postdiction, or a posterior prediction.

A second strategy is to measure the fit of the model across a set of different parameter values. Different proposals about what constitutes the appropriate set of values have been made. One attractive proposal is to assess the model's fit at the full set of values allowed by the model.

This is the idea behind evaluating a model on its marginal likelihood (Kass & Raftery, 1995; Myung & Pitt, 1997). The marginal likelihood does not reflect the best possible fit, but rather the average fit. Surely, the best fit is included in this evaluation, but so are all the other fits, some of which might be bad ones. The averaging process that results in the marginal likelihood, $p(d|M)$, is operationalized by integrating the likelihood, $p(d|\theta, M)$, as weighted by the prior probability of the parameters, $p(\theta|M)$, over the parameter range Ω:

$$p(d \mid M) = \int_{\Omega} p(d \mid \theta, M) p(\theta \mid M) d\theta. \qquad (10)$$

To find the parameter values at which to assess the model's fit under the marginal likelihood approach, no access to the data is required. Ideally, these values are given as part of the model definition, by specifying the range Ω and the prior $p(\theta|M)$. So the prediction the model makes, averaged across the values of θ, is a prediction in the strict sense. To stress that the prediction is made independent of any empirical data, it is called a prior prediction.

The maximizing-versus-averaging approach handles model complexity in different ways. A complex model can generate more data patterns than a simple model can, and thus has a higher chance of providing a good fit (i.e., making a prediction that matches the empirically observed data). The downside of the complex model's ability to generate more data patterns is that it also has a higher chance of providing a bad fit (i.e., making a prediction that poorly matches the empirically observed data). Increasing complexity increases the chance that the model's best fit is a good one, but it does not necessarily increase the model's average fit, since the additional fits might actually be poor ones. Hence, if only the best fit matters, a complex model has a clear advantage over a simple model. In contrast, if the average fit matters, a complex model is penalized when it provides bad fits, so the advantage disappears. In sum, the maximal likelihood is insensitive to complexity, since it ignores how many predictions a model can make. The marginal likelihood is sensitive to model complexity, and is therefore viewed as automatically implementing Occam's razor.

Abstraction in Category Learning

All of the 30 previously published data sets in our selection have been used in the original articles to evaluate the GCM or a closely related model. However, only the study by Nosofsky (1987) also reports the fit of the MPM to these data. To fill this gap, we first report the evaluation of both the GCM and the MPM against all 30 data sets. This MPM-versus-GCM analysis is a useful source of comparison to other studies that take an all-or-nothing view of abstraction, and contrast the prototype and exemplar models to empirical data (see, e.g., Nosofsky, 1992; Smith & Minda, 2000; Zaki, Nosofsky, Stanton, & Cohen, 2003). Unlike these studies, which invariably ignore model complexity, the present reanalysis relies on both complexity-insensitive and complexity-sensitive model evaluation methods.

Consistent with most previous findings, it will be seen that the MPM performs rather poorly relative to the GCM, indicating that no abstraction seems to be involved in category learning. Crucially, evidence against maximal abstraction should not be taken as evidence against the idea of any abstraction at all. In particular, the superiority of the GCM over the MPM does not rule out the possibility that people learning these categories might be relying on some intermediate level of abstraction. Such representations are provided by the VAM. To evaluate the possibility of partial abstraction, the 30 data sets are then analyzed using the VAM, again using both the maximal and marginal likelihood. Unlike the MPM-versus-GCM analysis, the VAM analysis provides evidence for the use of abstraction in category learning.

MPM-Versus-GCM Analysis

Table 2 provides the results of the MPM-versus-GCM analysis across our selection of the 30 data sets. It reports, for both models, the negative maximized log likelihood, $-\ln p(\hat{\theta})$, the percentage of variance accounted for, r^2 (based on the parameter values maximizing the likelihood), and the negative of the logarithm of the marginal likelihood, $-\ln p$. Note that the model minimizing the negative logarithm of the likelihood is to be preferred. Details on calculating the maximal and the marginal likelihoods are provided in Appendix B.[2]

Regardless of the model evaluation method, the same conclusion is supported. Table 2 provides overwhelming evidence in favor of exemplar-based category representations over prototype-based category representations. For only three data sets (Data Sets 4, 6, and 9), the MPM provides a better (optimal or average) fit than does the GCM. Moreover, the difference in fit between the MPM and the GCM is rather small in these cases. The superiority of the GCM over the MPM is consistent with many previous MPM-versus-GCM analyses (see Nosofsky, 1992, for an overview).

It is instructive to understand why the MPM performs so much worse than the GCM. The evidence in favor of the GCM and against the MPM is particularly compelling for the 4 data sets using a criss-cross category structure (Data Sets 1, 2, 10, and 14), and for the 11 data sets using an interior–exterior category structure (Data Sets 7, 8, 17, 18, 19, 25, 26, 27, 28, 29, and 30). Inspection of these category structures explains why the prototype representation is insufficient as a basis for categorization. For example, in Nosofsky's (1986) interior–exterior structure, the prototypes for Category A and for Category B virtually overlap; the same is true of the criss-cross structure used in Nosofsky (1986, 1987, 1989). In the interior–exterior structure used by Nosofsky et al. (1989) and Nosofsky and Palmeri (1998), as well, the two category prototypes are very close. Note that if the representation for Category A is identical to the representation for Category B, the similarity of a stimulus to Category A is identical to the similarity to Category B, and on account of Equation 5, performance at chance is predicted. Therefore, it is not surprising that the MPM fails to account for the data

Table 2
Best and Average Fits of the MPM and the GCM to the 30 Data Sets

Data Set	Reference	Experiment	Categories	Condition	MPM r^2	MPM $-\ln p(\hat{\theta})$	MPM $-\ln p$	GCM r^2	GCM $-\ln p(\hat{\theta})$	GCM $-\ln p$
1	Nosofsky (1986)	–	Criss-cross	Participant 1	0.10	959.75	964.86	96.29	86.82	100.16
2	Nosofsky (1986)	–	Criss-cross	Participant 2	5.88	615.79	626.82	87.69	131.35	145.12
3	Nosofsky (1986)	–	Diagonal	Participant 1	76.93	360.40	373.83	98.29	63.20	74.84
4	Nosofsky (1986)	–	Diagonal	Participant 2	99.26	49.94	63.72	98.07	62.80	74.76
5	Nosofsky (1986)	–	Dimensional	Participant 1	99.72	39.94	51.65	99.85	34.82	46.58
6	Nosofsky (1986)	–	Dimensional	Participant 2	98.24	58.39	69.73	97.86	63.62	74.41
7	Nosofsky (1986)	–	Interior–exterior	Participant 1	−19.23	600.19	610.89	70.16	188.09	200.91
8	Nosofsky (1986)	–	Interior–exterior	Participant 2	−22.86	360.10	364.70	60.29	147.83	160.05
9	Nosofsky (1987)	2	Brightness	–	99.65	54.79	67.06	99.64	60.34	72.56
10	Nosofsky (1987)	2	Criss-cross	–	1.78	933.65	943.14	98.66	50.36	61.12
11	Nosofsky (1987)	2	Saturation A	–	98.12	46.17	57.91	97.53	46.05	57.55
12	Nosofsky (1987)	2	Saturation B	–	95.58	131.52	143.84	98.69	58.82	70.90
13	Nosofsky (1989)	–	Angle	–	94.41	103.78	116.11	98.63	56.75	70.32
14	Nosofsky (1989)	–	Criss-cross	–	4.30	240.19	248.35	91.08	56.70	68.31
15	Nosofsky (1989)	–	Diagonal	–	97.82	49.44	62.58	98.55	49.08	61.45
16	Nosofsky (1989)	–	Size	–	99.14	44.27	55.20	99.38	40.77	51.92
17	Nosofsky, Clark, & Shin (1989)	1	Interior–exterior	Free	9.83	2,120.70	2,129.90	96.40	156.73	170.30
18	Nosofsky et al. (1989)	2	Interior–exterior	Rule 1	3.43	810.76	818.45	81.51	204.06	216.20
19	Nosofsky et al. (1989)	2	Interior–exterior	Rule 2	3.88	453.83	461.26	40.60	295.81	306.28
20	Nosofsky et al. (1989)	3	Conjunctive rule	Free 1	42.85	138.86	148.96	81.47	75.65	86.81
21	Nosofsky et al. (1989)	3	Conjunctive rule	Free 2	50.97	146.64	157.19	95.09	53.46	64.87
22	Nosofsky et al. (1989)	3	Conjunctive rule	Rule	59.78	256.50	268.09	94.37	87.14	98.87
23	Nosofsky & Palmeri (1997)	2	U7	–	93.18	171.05	182.73	99.66	32.18	42.79
24	Nosofsky & Palmeri (1997)	2	U8	–	88.07	185.71	200.84	99.42	41.27	52.15
25	Nosofsky & Palmeri (1998)	1	Interior–exterior	Beginning blocks	20.21	1,495.70	1,506.10	94.73	183.92	197.30
26	Nosofsky & Palmeri (1998)	1	Interior–exterior	Middle blocks	17.80	2,175.50	2,186.70	98.03	135.77	149.45
27	Nosofsky & Palmeri (1998)	1	Interior–exterior	End blocks	17.76	2,740.00	2,751.50	98.81	140.16	153.98
28	Nosofsky & Palmeri (1998)	2	Interior–exterior	Beginning blocks	16.10	265.34	274.36	76.57	106.40	117.53
29	Nosofsky & Palmeri (1998)	2	Interior–exterior	Middle blocks	14.09	509.37	517.71	89.50	112.46	124.60
30	Nosofsky & Palmeri (1998)	2	Interior–exterior	End blocks	13.17	2,004.00	2,014.10	90.02	365.13	378.70

Note—r^2, percentage of variance accounted for; $-\ln p(\hat{\theta})$, negative value of the maximized log likelihood; $-\ln p$, negative value of the logarithm of the marginal likelihood.

collected using category structures that imply (nearly) identical prototypes for both categories.

Interestingly, these structures might not support a totally abstract representation, but they naturally invite several partially abstract representations that seem highly plausible and intuitively appealing (an observation already made in Nosofsky, 1987). The viability of these representations is tested by analyzing the data using the VAM.

VAM Analysis

As in the MPM-versus-GCM analysis, the model evaluation method has hardly any effect on the conclusion. When the marginal rather than the maximal likelihood is used, only for Data Sets 5, 14, and 15 would a different model be preferred. The remarkable agreement between both methods will be discussed in more detail shortly. The results of the VAM analysis are shown in Figure 3 and Table 3. The figure shows, for each of the 30 data sets, the model in the VAM that provided the highest optimal fit to the data (i.e., with the highest maximal likelihood), using the graphical conventions adopted earlier (i.e., the subprototypes are shown in black and are connected by lines to the original category members, shown in white).

The table reports, for the model in the VAM shown in Figure 3, the negative maximized log likelihood, $-\ln p(\hat{\theta})$, the percentage of variance accounted for, r^2, and the negative of the logarithm of the marginal likelihood, $-\ln p$.[3]

Although the MPM-versus-GCM analysis of the 30 data sets fails to provide evidence for the use of abstraction, Figure 3 shows that the VAM analysis does reveal the use of representational abstraction in the majority of the 30 data sets. In some cases, abstraction seemed to be present to a

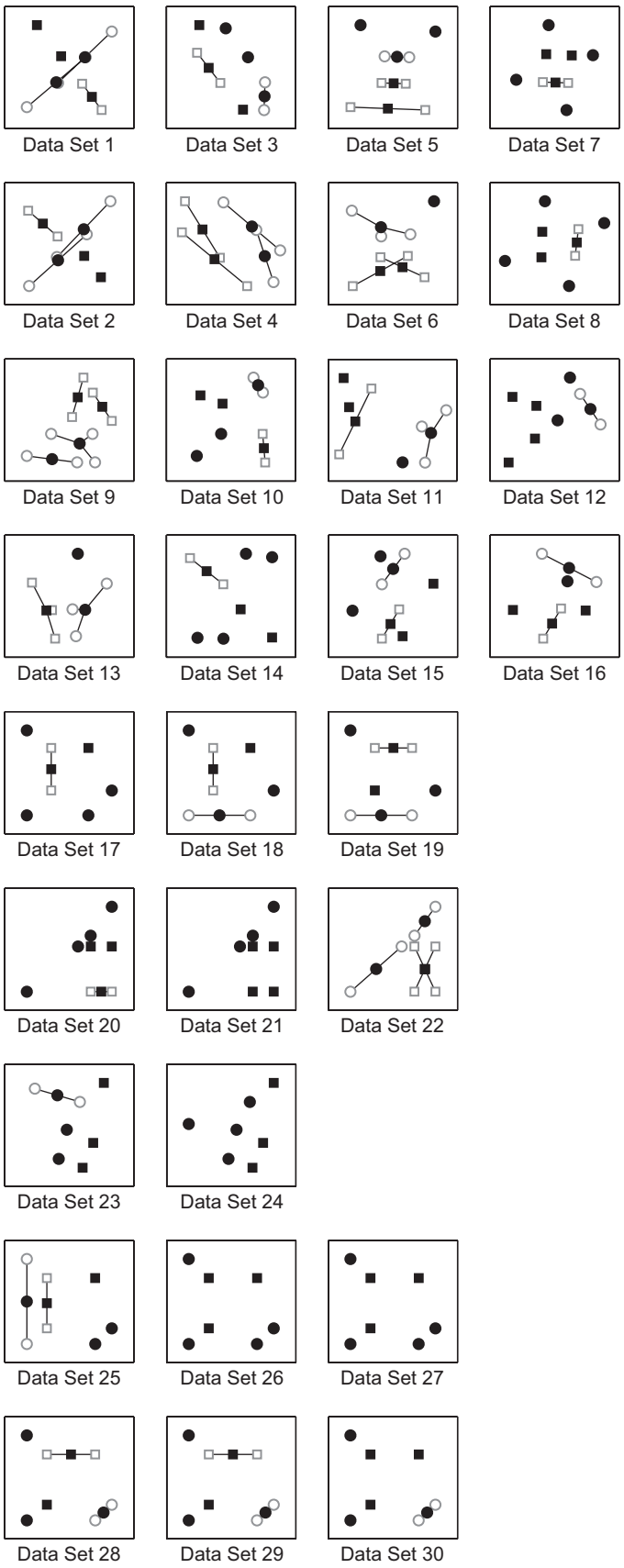


Figure 3. Representations with highest maximal likelihood.

Table 3
Best and Average Fits of the Best Model in the VAM to the 30 Data Sets

Data Set	Reference	Experiment	Categories	Condition	r^2	$-\ln p(\hat{\theta})$	$-\ln p$
1	Nosofsky (1986)	–	Criss-cross	Participant 1	97.03	70.49	83.26
2	Nosofsky (1986)	–	Criss-cross	Participant 2	94.60	82.89	95.13
3	Nosofsky (1986)	–	Diagonal	Participant 1	99.68	45.73	57.67
4	Nosofsky (1986)	–	Diagonal	Participant 2	99.47	48.02	61.52
5	Nosofsky (1986)	–	Dimensional	Participant 1	99.93	31.42	43.34
6	Nosofsky (1986)	–	Dimensional	Participant 2	99.40	45.42	56.25
7	Nosofsky (1986)	–	Interior–exterior	Participant 1	87.76	109.70	122.24
8	Nosofsky (1986)	–	Interior–exterior	Participant 2	83.50	89.45	101.61
9	Nosofsky (1987)	2	Brightness	–	99.85	42.31	54.19
10	Nosofsky (1987)	2	Criss-cross	–	99.09	44.65	55.73
11	Nosofsky (1987)	2	Saturation A	–	98.52	41.09	53.00
12	Nosofsky (1987)	2	Saturation B	–	99.04	56.45	68.76
13	Nosofsky (1989)	–	Angle	–	99.56	39.86	52.88
14	Nosofsky (1989)	–	Criss-cross	–	93.78	47.70	58.80
15	Nosofsky (1989)	–	Diagonal	–	98.93	46.81	58.89
16	Nosofsky (1989)	–	Size	–	99.62	36.47	47.64
17	Nosofsky, Clark, & Shin (1989)	1	Interior–exterior	Free	97.66	134.40	148.03
18	Nosofsky et al. (1989)	2	Interior–exterior	Rule 1	87.54	171.96	184.23
19	Nosofsky et al. (1989)	2	Interior–exterior	Rule 2	51.70	256.14	267.06
20	Nosofsky et al. (1989)	3	Conjunctive rule	Free 1	83.18	72.81	84.12
21	Nosofsky et al. (1989)	3	Conjunctive rule	Free 2	95.09	53.46	64.87
22	Nosofsky et al. (1989)	3	Conjunctive rule	Rule	96.74	57.24	69.08
23	Nosofsky & Palmeri (1997)	2	U7	–	99.63	30.58	42.29
24	Nosofsky & Palmeri (1997)	2	U8	–	99.42	41.27	52.15
25	Nosofsky & Palmeri (1998)	1	Interior–exterior	Beginning blocks	95.63	170.48	183.88
26	Nosofsky & Palmeri (1998)	1	Interior–exterior	Middle blocks	98.03	135.77	149.45
27	Nosofsky & Palmeri (1998)	1	Interior–exterior	End blocks	98.81	140.16	153.98
28	Nosofsky & Palmeri (1998)	2	Interior–exterior	Beginning blocks	85.71	82.84	94.15
29	Nosofsky & Palmeri (1998)	2	Interior–exterior	Middle blocks	90.52	107.92	120.01
30	Nosofsky & Palmeri (1998)	2	Interior–exterior	End blocks	93.22	327.59	341.21

Note—r^2, percentage of variance accounted for; $-\ln p(\hat{\theta})$, negative value of the maximized log likelihood; $-\ln p$, negative value of the logarithm of the marginal likelihood.

limited extent only, involving a single pair of stimuli being merged (e.g., Data Sets 7, 8, 12, 14, 17, 20, 23, and 30). Evidence for more abstraction was found in 9 other data sets (Data Sets 3, 10, 15, 16, 18, 19, 25, 28, and 29), where two separate pairs of category members are being merged. Some data sets supported highly—albeit not fully—abstract representations involving either three (Data Sets 1, 2, and 5) or four (Data Sets 4 and 22[4]) pairs of stimuli being merged, or the merging of a triplet as well as one or more pairs (Data Sets 6, 9, and 11). The representation best capturing Data Set 13 is the closest to a prototype one, although one exemplar remains unmerged. Importantly, for 4 data sets, there is no evidence for abstraction, since the exemplar representation is the representation with the highest optimal fit (Data Sets 21, 24, 26, and 27).

Model Evaluation in the VAM

In both the MPM-versus-GCM analysis and the VAM analysis, the maximal and marginal likelihoods implied virtually identical conclusions. This observation suggests that the complexity differences between the models in the VAM family are small and that the use of a complexity-insensitive model evaluation method is justified. The practical advantage of complexity-insensitive methods over complexity-sensitive methods is that the former are computationally mild, whereas the latter are computationally challenging. This difference in technical sophistication makes it worthwhile to consider whether a complexity sensitive method is needed in a VAM analysis. This question can be studied by means of a recovery study, in which both methods are evaluated on their ability to recover a known representation (Pitt, Kim, & Myung, 2003).

A separate recovery study was conducted for each of the 30 data sets. Each study involved generating artificial data sets from each model in the VAM and evaluating all models in the VAM to these sets using the maximal and the marginal likelihood. For each of the 30 data sets, and from each model in the VAM involved, 100 artificial data sets were generated, implying that the recovery study consisted of 535,500 artificial sets. Given that each model in the VAM was evaluated separately using each of these data sets, the recovery study involved close to 2×150 million model fits.

A simulated data set from a model was created from the parameter values obtained by fitting the model to the empirically observed data, using maximum likelihood estimation. From these parameter values, the response probabilities according to the model were computed using Equation 5. Sampling error was added by generating, for each of these probabilities, a set of binary-valued responses according to the binomial probability distribution. The number of hypothetical responses was the same as in the sample size in the published data set. Each simulated data set consisted of the number of times each stimulus was chosen in Category A. All models of the VAM were evaluated separately with re-

spect to the artificial data, under both the maximal likelihood and the marginal likelihood.

This, then, allows us to calculate the recovery rate (the percentage of correctly classified data sets) for each data set and under each model evaluation method. A data set is correctly classified if the generating model is chosen as the preferred model. Hence, the recovery rate measures the ability of the model evaluation method to recover the true model that generated the data. It turned out that the recovery rates under each model evaluation method were extremely close to each other, with the correlation between the recovery rates being as high as .9998. The recovery rate for each data set and each model evaluation method is shown in Table C1 of Appendix C. These nearly identical recovery rates indicate that the models in the VAM family are closely matched in terms of complexity. This finding implies that the use of a complexity-sensitive method is not required, and a complexity-insensitive method suffices.

Discussion

The small complexity differences between the models in the VAM family should not be overly surprising. It has been suggested that the following four dimensions contribute to model complexity: the number of parameters, the range of the parameters, the prior over the parameters, and their functional form (i.e., the way in which the parameters are combined; e.g., Myung & Pitt, 1997; Vanpaemel, 2009b). One of the main features of the VAM is that all its instantiations are equated on the first three of these dimensions: They have the same number of parameters with the same range and the same prior. The models do differ in their functional form, but the functional forms differ very slightly, only to the extent that a different S_A enters Equation 4.

The finding that the complexity differences between the different models in the VAM are negligible corroborates a previous finding by Pitt, Myung, and Zhang (2002), who conducted a similar recovery study. Their study was based on a single data set (taken from Shin & Nosofsky, 1992) and involved the GCM and the MPM only. As in the present study, their main finding was that a complexity-sensitive method did not perform better than a complexity-insensitive method did.[5] The present result generalizes Pitt et al.'s (2002) results in two important ways: The present study involved 75, 225, or 780, instead of only 2, models, and involved data from 30 experiments rather than from a single one.

A legitimate concern relates to the generality of this finding. A first point in this regard is that model complexity is not absolute, but rather depends on the experimental design, which includes the category structure and the sample size (e.g., Myung, Pitt, & Navarro, 2007; Pitt et al., 2002; Vanpaemel, 2009b). This implies that complexity differences may not be negligible across all possible experimental designs and sample sizes. It is therefore reassuring that the present recovery study involved 30 different experimental conditions, based on 16 different category structures, and spanning a large range of sample sizes, and that the equal performance of the maximal likelihood and the marginal likelihood is extremely robust across all these conditions.

A second point is that the version of the VAM used in the recovery study is a baseline version of the VAM, using as few parameters as needed. Two additional parameters have sometimes been used in the GCM, and could easily be incorporated in the VAM. Both of these parameters relate to the response rule of Equation 5. The first of these additional parameters is the bias parameter β, which reflects response bias to Category A. The second parameter is the response-scaling parameter γ, which reflects the amount of determinism in responding (Ashby & Maddox, 1993; Navarro, 2007). The modified response rule using both additional parameters becomes

$$p(\mathrm{A} \mid x_i) = \frac{\beta s(x_i, \mathrm{A})^{\gamma}}{\beta s(x_i, \mathrm{A})^{\gamma} + (1-\beta) s(x_i, \mathrm{B})^{\gamma}}. \quad (11)$$

A legitimate question is whether complexity differences are still small when this modified version of the VAM is used. When this modified response rule is incorporated in *all* the different models of the VAM family, all models still have identical parameters with identical ranges and priors, and still have extremely similar functional forms. Therefore, we suspect that adding these additional parameters to the VAM will not introduce complexity differences between the different models in the VAM.[6]

General Discussion

In an effort to escape from a limiting focus on extreme representational assumptions in category learning research, Vanpaemel and Storms (2008) introduced the VAM and applied it to a limited number of data sets. The present study extends the findings of Vanpaemel and Storms in three ways: It considers a much broader range of data; it relies on a complexity-sensitive model evaluation method; and it reports a large recovery study affording insight into the relative utility of complexity-sensitive versus -insensitive methods.

To understand the aim of a VAM analysis, it is useful to realize that formal modeling can serve at least two different purposes, depending on the status of the model (see, e.g., Vanpaemel, 2009a, for some examples). One is to regard the model as a theoretical model, implementing a particular theory about the nature of a cognitive process. The goal of formal modeling of psychological data is then to test the model against empirical data. To the extent that the model passes the empirical test, support is provided for the theory embodied by the model. A second possibility is that the model is used as a measurement model. The goal of modeling is then no longer to test the theory formalized by the model, but to test hypotheses about parameter values.

The two distinct purposes of formal modeling map onto a common distinction in statistics. In its first, theoretical, sense, modeling data relies on model evaluation, comparison, and selection, where some quantitative statement is made about the degree to which different competing

models, implementing different theories, are supported by the observed data. In contrast, modeling data in the second, measurement sense relies on parameter estimation, where some quantitative statement is made about the values of the unknown parameters of interest on the basis of observed data.

At face value, the VAM analysis presented here seems to rely on model evaluation. However, the models being evaluated are the individual models in the VAM family, rather than the VAM family as a whole. One can usefully think of all the different models in the VAM family as corresponding to different values of some abstraction parameter (albeit not the typical continuously varying quantitative one). From this perspective, the goal of a VAM analysis—inferring the best model in the VAM—corresponds to inferring the value of this abstraction parameter. In this sense, in a VAM analysis the VAM is applied as a measurement model and relies on parameter estimation to estimate the level of abstraction involved in people's category representation.

Model Evaluation

One goal of the present study was to evaluate whether a complexity-sensitive but technically challenging method such as the marginal likelihood is called for in a VAM analysis, or whether a complexity-insensitive but relatively straightforward method such as the maximal likelihood is sufficient. A first hint that both methods perform alike was found in the analysis of the empirical data; virtually identical conclusions were drawn, regardless of the method used. The equivalence of both methods was addressed in more detail using a recovery study, involving 30 different experimental setups. It was found that recovery did not improve when model complexity was incorporated in the model evaluation. The similar performance of both methods indicates that complexity differences between the different models in the VAM are rather small, so a VAM analysis can safely rely on a method insensitive to model complexity.

The maximum likelihood method is not free from limitations. In particular, it ignores two types of prior information, about the parameters and about the models. Unlike the marginal likelihood, it is insensitive to prior information about the parameter values. In the present study, it was assumed that no such information was available, which was reflected in the use of an uninformative prior. However, a researcher wishing to incorporate additional information, based on previous data or on theorizing, can choose to rely on informative priors (e.g., Vanpaemel & Lee, 2010). Since the maximal likelihood does not take priors into account, this additional information is wasted using this method. In contrast, the marginal likelihood naturally takes prior information into account.

A second limitation of the maximal likelihood, which holds for the marginal likelihood as well, is that it is insensitive to prior information about the models. Implicitly, each model is given equal prior status. However, intuitively, some of the VAM representations seem more likely than others, even before any data have been collected. In particular, both the exemplar and prototype representations, as well as representations based on merging of similar stimuli (such as representations E and G in Figure 2) seem more likely than many of the others. These intuitions are ignored using the maximum (and marginal) likelihood method. Lee and Vanpaemel (2008) have recently proposed a theoretically grounded way to formalize a prior distribution over the models that captures these intuitions. The key to their hierarchical Bayesian analysis is the inclusion of a process that generates the different VAM representations. By extending the VAM with a generative process, it is possible to specify a prior distribution over the different models, giving most a priori weight to the pure exemplar and prototype representations and to the representations in which merging is driven by similarity.

Against these limitations, one important strength of the maximum likelihood method is its computational simplicity. It is straightforward to use, requiring only a likelihood and an optimization algorithm. Given the negligible complexity differences among the models in the VAM family and the computational simplicity of the maximal likelihood, it is an attractive method for model evaluation in the VAM.

Abstraction

The second goal in this article was to provide further evidence for abstraction in category learning. Our VAM analyses showed that for most data sets, some abstraction seems to be involved in the representation, thereby generalizing the earlier findings of Vanpaemel and Storms (2008) to a larger set of conditions. This demonstration is important because, on the basis of an MPM-versus-GCM analysis, most of these data sets provide evidence against representational abstraction.

Importantly, for four data sets, using the VAM, no evidence for abstraction was found, which indicates that there are situations in which no abstraction seems to be used. This highlights the intended contribution of the VAM. The VAM does not posit that category learning will *always* involve abstraction. Beyond doubt, conditions exist in which no abstraction will be involved at all. Rather, it hopes to shift the focus from the question of *whether* abstraction is involved to the more interesting and vastly more challenging goal of pinning down these conditions—that is, to understand *when* abstraction is bound to take place.

In this regard, 6 of the 30 data sets deserve special attention. In particular, Data Sets 25, 26, and 27 correspond to test performance from the beginning, middle, and end of one category learning task. Each data set corresponds to a different stage of learning. Also Data Sets 28, 29, and 30 correspond to test performance at different stages of learning. These data sets afford insight into the issue of representational change over time. It has been suggested that at the early stages of learning, representations are abstract and only become sufficiently detailed as learning matures (see Smith & Minda, 1998; but see Nosofsky & Zaki, 2002, for a contrasting view). The present VAM analysis shows that, for both groups, there is some evidence for a gradual loss of abstraction over time. In particular, the VAM analysis shows a shift from a moderately abstract representation (Data Sets 25 and 28) toward a pure (Data Set 27) or a near (Data Set 30) exemplar representation.

Other variables, apart from the time point during learning, that possibly influence the level of abstraction in people's category representation include category size (Homa, Dunbar, & Nohre, 1991; Minda & Smith, 2001), category complexity (Feldman, 2003, 2006), the time point after learning (Homa, Sterling, & Trepel, 1981; Posner & Keele, 1970), time pressure (Lamberts, 1995), and learning instructions (Medin & Smith, 1981; Palmeri & Nosofsky, 1995). We believe that systematically manipulating these and other variables and analyzing the observed data using the VAM will afford detailed insight into what influences the level of abstraction in people's category representations and, as such, will advance progress in our understanding of category learning.

AUTHOR NOTE

The research reported in this article was funded by Grant FWO G.0513.08. The input and help from Eric-Jan Wagenmakers, Dan Navarro, Michael Lee, Wouter Voorspoels, and Marit Daneels are much appreciated. MATLAB code for a VAM analysis is available on the first author's Web site. Correspondence concerning this article should be addressed to W. Vanpaemel, Department of Psychology, University of Leuven, Tiensestraat 102, B-3000 Leuven, Belgium (e-mail: wolf.vanpaemel@psy.kuleuven.be; Web site: http://ppw.kuleuven.be/concat).

REFERENCES

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csáki (Eds.), *Second International Symposium on Information Theory* (pp. 267-281). Budapest: Akadémiai Kiadó.

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

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

Balasubramanian, V. (1997). Statistical inference, Occam's razor, and statistical mechanics on the space of probability distributions. *Neural Computation*, **9**, 349-368.

Barron, A. R., & Cover, T. M. (1991). Minimum complexity density estimation. *IEEE Transactions on Information Theory*, **37**, 1034-1054.

Brooks, L. R. (1978). Nonanalytic concept formation and memory for instances. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 169-211). Hillsdale, NJ: Erlbaum.

Cutting, J. E., Bruno, N., Brady, N. P., & Moore, C. (1992). Selectivity, scope, and simplicity of models: A lesson from fitting judgments of perceived depth. *Journal of Experimental Psychology: General*, **121**, 364-381.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, **70**, 193-242.

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

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

Fried, L. S., & Holyoak, K. J. (1984). Induction of category distributions: A framework for classification learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **10**, 234-257.

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. S. McNamara & J. G. Trafton (Eds.), *Proceedings of the 29th Annual Conference of the Cognitive Science Society* (pp. 323-328). Mahwah, NJ: Erlbaum.

Grünwald, P. (2000). Model selection based on minimum description length. *Journal of Mathematical Psychology*, **44**, 133-152.

Homa, D., Dunbar, S., & Nohre, L. (1991). Instance frequency, categorization, and the modulating effect of experience. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **17**, 444-458.

Homa, D., Sterling, S., & Trepel, L. (1981). Limitations of exemplar-based generalization and the abstraction of categorical information. *Journal of Experimental Psychology: Human Learning & Memory*, **7**, 418-439.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**, 773-795.

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

Lamberts, K. (1995). Categorization under time pressure. *Journal of Experimental Psychology: General*, **124**, 161-180.

Lee, M. D. (2004). A Bayesian analysis of retention functions. *Journal of Mathematical Psychology*, **48**, 310-321.

Lee, M. D., & Vanpaemel, W. (2008). Exemplars, prototypes, similarities and rules in category representation: An example of hierarchical Bayesian analysis. *Cognitive Science*, **32**, 1403-1424.

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

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

Massaro, D. W., Cohen, M. M., Campbell, C. S., & Rodriguez, T. (2001). Bayes factor of model selection validates FLMP. *Psychonomic Bulletin & Review*, **8**, 1-17.

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

Medin, D. L., & Smith, E. E. (1981). Strategies and classification learning. *Journal of Experimental Psychology: Human Learning & Memory*, **7**, 241-253.

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

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, **47**, 90-100.

Myung, I. J., Forster, M. R., & Browne, M. W. (2000). The importance of complexity in model selection [Special issue]. *Journal of Mathematical Psychology*, **44**(1).

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, **4**, 79-95.

Myung, J. I., Pitt, M. A., & Navarro, D. J. (2007). Does response scaling cause the generalized context model to mimic a prototype model? *Psychonomic Bulletin & Review*, **14**, 1043-1050.

Navarro, D. J. (2007). On the interaction between exemplar-based concepts and a response scaling process. *Journal of Mathematical Psychology*, **51**, 85-98.

Navarro, D. J., & Lee, M. D. (2003). Combining dimensions and features in similarity-based representations. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in neural information processing systems* (Vol. 15, pp. 67-74). Cambridge, MA: MIT Press.

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, & Cognition*, **13**, 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 honor of William K. Estes* (Vol. 1, pp. 149-167). Hillsdale, NJ: Erlbaum.

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

Nosofsky, R. M., & Zaki, S. R. (2002). Exemplar and prototype models

revisited: Response strategies, selective attention, and stimulus generalization. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **28**, 924-940.

PALMERI, T. J., & NOSOFSKY, R. M. (1995). Recognition memory for exceptions to the category rule. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **21**, 548-568.

PITT, M. A., KIM, W., & MYUNG, I. J. (2003). Flexibility versus generalizability in model selection. *Psychonomic Bulletin & Review*, **10**, 29-44.

PITT, M. A., & MYUNG, I. J. (2002). When a good fit can be bad. *Trends in Cognitive Sciences*, **6**, 421-425.

PITT, M. A., MYUNG, I. J., & ZHANG, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, **109**, 472-491.

POSNER, M. I., & KEELE, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, **77**, 353-363.

POSNER, M. I., & KEELE, S. W. (1970). Retention of abstract ideas. *Journal of Experimental Psychology*, **83**, 304-308.

REED, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, **3**, 382-407.

ROBERTS, S., & PASHLER, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, **107**, 358-367.

SCHWARZ, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461-464.

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

SHEPARD, R. N. (1991). Integrality versus separability of stimulus dimensions: From an early convergence of evidence to a proposed theoretical basis. In G. R. Lockhead & J. R. Pomerantz (Eds.), *The perception of structure: Essays in honor of Wendell R. Garner* (pp. 53-71). Washington, DC: APA.

SHIFFRIN, R. M., LEE, M. D., KIM, W., & WAGENMAKERS, E.-J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science*, **32**, 1248-1284.

SHIN, H. J., & NOSOFSKY, R. M. (1992). Similarity-scaling studies of dot-pattern classification and recognition. *Journal of Experimental Psychology: General*, **121**, 278-304.

SLOANE, N. J. A. (2003). The on-line encyclopedia of integer sequences. *Notices of the AMS*, **50**, 912-915.

SMITH, J. D., & MINDA, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, & 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, & Cognition*, **26**, 3-27.

SPIEGELHALTER, D. J., THOMAS, A., BEST, N., GILKS, W. R., & LUNN, D. (1994). *BUGS: Bayesian Inference Using Gibbs Sampling*. Retrieved April 9, 2010, from Cambridge University, Institute of Public Health Web site: www.mrc-bsu.cam.ac.uk/bugs/.

STEYVERS, M., LEE, M. D., & WAGENMAKERS, E.-J. (2009). A Bayesian analysis of human decision-making on bandit problems. *Journal of Mathematical Psychology*, **53**, 168-179.

VANPAEMEL, W. (2007). *In search of abstraction*. Unpublished doctoral dissertation, K. U. Leuven.

VANPAEMEL, W. (2009a). BayesGCM: Software for Bayesian inference with the generalized context model. *Behavior Research Methods*, **41**, 1111-1120.

VANPAEMEL, W. (2009b). Measuring model complexity with the prior predictive. In Y. Bengio, D. Shuurmans, J. Lafferty, C. K. I. Williams, & A. Culota (Eds.), *Advances in neural information processing systems* (pp. 1919-1927). Cambridge, MA: MIT Press.

VANPAEMEL, W., & LEE, M. D. (2010). *Using priors to formalize theory: Optimal attention and the generalized context model*. Manuscript submitted for publication.

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: Erlbaum.

WALDORP, L., & WAGENMAKERS, E.-J. (EDS.) (2006). Model selection: Theoretical developments and applications [Special issue]. *Journal of Mathematical Psychology*, **50**(2).

ZAKI, S. R., NOSOFSKY, R. M., STANTON, R. D., & COHEN, A. L. (2003). Prototype and exemplar accounts of category learning and attentional allocation: A reassessment. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **29**, 1160-1173.

NOTES

1. Three of the 30 data sets have been considered by Vanpaemel and Storms (2008), but are included here to keep the present article self-contained.

2. Most values of $-\ln p(\hat{\theta})$ and r^2 are not exactly identical to those reported in the original studies. There are several factors contributing to this discrepancy. One reason is that the data used in the present study might differ slightly from the data used in the analyses reported in the original studies. (The reason for this is explained in Appendix A.) A second reason is that in some of the original studies slightly different models have been used. For example, Nosofsky (1987) used a version of the GCM that made use of stimulus biases, which were estimated from the data of an identification experiment. Both the differences in the data and the models might cause slight variations in $-\ln p(\hat{\theta})$ and r^2.

3. For Data Sets 5, 14, and 15, Table 3 reports the marginal likelihood for a model that is not depicted in Figure 3.

4. Both subprototypes in the category displayed by the squares overlap.

5. In Pitt et al. (2002), instead of the maximal likelihood and the marginal likelihood, AIC and minimum description length (MDL; Grünwald, 2000) were compared. However, when models have the same number of parameters, such as the MPM and the GCM, model evaluation under AIC is the same as that under the maximal likelihood. Furthermore, MDL has been shown to be closely related to the marginal likelihood (see, e.g., Balasubramanian, 1997; Barron & Cover, 1991). As such, contrasting AIC and MDL is, in the present context, essentially equivalent to contrasting the maximal likelihood and the marginal likelihood.

6. One notable exception is the MPM. It has been noted (e.g., Ashby & Maddox, 1993) that the response-scaling parameter γ is not identifiable in the MPM, and therefore cannot be incorporated in the MPM. This discrepancy is expected to introduce large complexity differences between the MPM (without γ) and all other models in the VAM (with γ).

(*Continued on next page*)

APPENDIX A
The 30 Data Sets

This appendix reproduces the 30 previously published data sets. Together with Figure 1, which shows the stimulus assignments, the tables provided in this appendix contain all details needed for testing different category learning models. Table A1 reports the stimulus coordinates. Table A2 reports the observed frequency of Category A decisions for each stimulus, as well as the average number of observations per stimulus. This average sample size equals the number of participants (s) multiplied by the average number of test trials per participant (n) from Table 1. In most tasks, every stimulus was presented the same number of times during the test phase. However, for the 12 data sets collected by Nosofsky (1986, 1989), the number of test trials varied for each stimulus. For these tasks, the number of observations is reported in Table A3, separately for each stimulus. Also in the Saturation (A) condition of Nosofsky (1987), the sample size was not identical for every stimulus. Unfortunately, only an averaged sample size was reported in the original publication and the sample size per stimulus could not be reconstructed.

Most of these data have been reproduced from the tables found in the original publications. The only exceptions are the stimulus coordinates from Nosofsky and Palmeri's (1997) Experiment 2 and Nosofsky and Palmeri's (1998) Experiments 1 and 2. In these studies, the derived MDS representations are only presented graphically, so the values reported in Table A1 have been read off from a figure. Also, apart from the 12 data sets reported in Nosofsky (1986, 1989), the remaining 18 data sets have been reported in the form of response proportions, rather than response frequencies. The frequencies shown in Table A2 have been recalculated from the originally published proportions. For these data sets, the rounding of the proportions in the original publications might have resulted in small deviances between the frequencies presented in Table A2 and the frequencies actually observed in the original studies.

Table A1
Stimulus Coordinates

	Nosofsky (1986) Participant 1		Nosofsky (1986) Participant 2		Nosofsky (1987)		Nosofsky (1989)		Nosofsky, Clark, & Shin (1989) Experiments 1 and 2		Nosofsky, Clark, & Shin (1989) Experiment 3		Nosofsky & Palmeri (1997) Experiment 2		Nosofsky & Palmeri (1998) Experiments 1 and 2	
x_i	x_{i1}	x_{i2}	x_{i1}	x_{i2}	x_{i1}	x_{i2}	x_{i1}	x_{i2}	x_{i1}	x_{i2}	x_{i1}	x_{i2}	x_{i1}	x_{i2}	x_{i1}	x_{i2}
1	−1.855	−1.532	−1.356	−1.430	−2.543	2.641	0.312	−0.241	0	0	0	0	0.6	−1.5	0.2	0.3
2	−0.687	−1.617	−0.492	−1.211	0.943	4.341	0.918	−0.264	3.422	0	7.632	0	1	−0.65	0.8	0.3
3	0.436	−1.633	0.349	−1.235	−1.092	1.848	1.405	−0.187	8.655	0	9.527	0	1.4	1.4	2.25	0.3
4	1.331	−1.647	1.088	−1.121	1.558	2.902	2.062	−0.227	11.941	0	12.655	0	−0.3	−1.2	2.75	0.3
5	−1.615	−0.469	−1.413	−0.528	−2.258	0.430	0.228	0.640	0	2.787	0	2.884	0	−0.2	0.2	0.75
6	−0.531	−0.558	−0.412	−0.425	0.194	0.572	0.844	0.662	3.422	2.787	7.632	2.884	0.5	0.75	0.8	0.75
7	0.500	−0.590	0.518	−0.382	2.806	0.202	1.324	0.687	8.655	2.787	9.527	2.884	−1.8	0	2.25	0.75
8	1.373	−0.535	1.275	−0.215	−1.177	−1.038	1.885	0.623	11.941	2.787	12.655	2.884	−1.2	1.2	2.75	0.75
9	−1.522	0.657	−1.477	0.302	1.543	−1.040	0.374	1.555	0	7.545	0	3.573	–	–	0.2	2.2
10	−0.395	0.518	−0.388	0.294	−2.775	−3.149	0.916	1.501	3.422	7.545	7.632	3.573	–	–	0.8	2.2
11	0.648	0.469	0.607	0.368	0.528	−3.766	1.473	1.544	8.655	7.545	9.527	3.573	–	–	2.25	2.2
12	1.513	0.481	1.317	0.539	1.709	−3.773	2.128	1.520	11.941	7.545	12.655	3.573	–	–	2.75	2.2
13	−1.427	1.770	−1.389	1.189	–	–	0.135	2.352	0	9.518	0	5.422	–	–	0.2	2.75
14	−0.301	1.639	−0.313	1.153	–	–	0.889	2.412	3.422	9.518	7.632	5.422	–	–	0.8	2.75
15	0.767	1.541	0.673	1.187	–	–	1.451	2.493	8.655	9.518	9.527	5.422	–	–	2.25	2.75
16	1.764	1.512	1.414	1.518	–	–	2.061	2.382	11.941	9.518	12.655	5.422	–	–	2.75	2.75

Note—x_i, the ith stimulus; x_{ij}, value of stimulus x_i on dimension j.

(*Continued on next page*)

APPENDIX A (Continued)

Table A2
Average Sample Size and Observed Frequency of Category A Responses for Each Stimulus

Data Set	t	k_1	k_2	k_3	k_4	k_5	k_6	k_7	k_8	k_9	k_{10}	k_{11}	k_{12}	k_{13}	k_{14}	k_{15}	k_{16}
1	219[a]	7	27	183	206	58	73	152	187	193	147	46	66	212	149	35	13
2	225[a]	30	95	164	192	101	88	155	169	176	131	62	75	199	122	22	18
3	250[a]	226	231	165	92	214	206	109	44	208	108	31	12	209	71	13	3
4	225[a]	216	199	139	41	203	193	53	14	219	151	36	6	189	75	6	3
5	225[a]	213	253	192	218	185	193	187	162	24	33	31	40	0	0	0	0
6	200[a]	196	185	214	197	155	150	152	165	59	86	57	58	9	11	7	4
7	256[a]	19	78	76	65	36	179	161	99	60	206	171	101	32	128	116	39
8	144[a]	14	48	54	24	38	95	89	41	61	131	122	33	40	70	34	15
9	490	483	488	467	487	261	420	454	19	47	2	0	4	–	–	–	–
10	240	233	15	217	21	138	136	163	16	196	8	207	228	–	–	–	–
11	180[a]	176	85	163	76	170	102	7	152	10	166	15	10	–	–	–	–
12	400	379	33	344	33	337	85	18	347	97	381	209	160	–	–	–	–
13	167[a]	79	155	48	2	81	190	60	2	262	47	11	4	76	47	24	2
14	150[a]	48	94	45	162	34	34	138	47	53	120	33	24	180	44	63	41
15	175[a]	40	200	242	83	65	49	180	70	8	57	42	199	3	27	58	55
16	150[a]	72	255	72	73	234	66	208	226	23	18	55	58	2	8	3	3
17	610	31	329	104	37	116	506	226	49	110	525	555	238	49	464	531	195
18	150	6	62	49	6	9	130	116	17	12	143	138	25	17	141	138	20
19	140	32	39	36	5	114	117	117	21	111	125	120	44	53	61	76	23
20	150	32	85	107	118	63	47	94	109	59	45	60	82	54	41	51	31
21	150	35	103	119	127	56	58	83	98	46	45	68	90	50	36	42	30
22	150	21	97	140	141	12	65	103	111	8	38	60	81	8	20	26	26
23	248	239	230	246	27	3	17	14	1	–	–	–	–	–	–	–	–
24	248	235	235	243	26	7	29	4	19	–	–	–	–	–	–	–	–
25	492	33	349	97	32	39	385	137	67	90	442	367	262	66	418	413	277
26	492	17	405	57	21	29	433	114	38	64	469	419	259	31	439	435	270
27	492	5	426	28	8	18	464	84	35	46	481	451	265	14	466	451	265
28	120	11	13	18	10	75	76	72	54	79	90	95	91	23	39	60	64
29	120	4	6	4	4	97	101	64	19	83	105	116	102	9	36	56	57
30	360	3	4	5	3	325	347	207	30	270	343	353	327	10	103	175	172

Note—t, number of times each stimulus was presented in the test phase; k_i, number of times stimulus x_i was classified in Category A. [a]Average value; for data sets indicated with a superscript *a* except Set 11, the sample size per stimulus is provided by Table A3.

Table A3
Sample Size for Each Stimulus for Selected Data Sets

Data Set	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	t_{11}	t_{12}	t_{13}	t_{14}	t_{15}	t_{16}
1	210	214	244	215	220	224	206	234	214	211	201	220	216	193	249	229
2	220	256	190	206	240	216	225	231	227	249	214	222	222	223	221	237
3	226	251	234	260	220	273	260	256	228	243	295	257	250	262	224	261
4	219	214	197	236	208	229	231	229	231	218	230	243	213	218	239	244
5	217	254	194	219	242	240	227	198	218	231	221	221	204	235	220	258
6	203	189	223	205	182	185	207	216	179	202	192	228	202	182	202	203
7	257	240	257	284	252	251	226	258	249	268	246	251	270	254	273	262
8	146	129	170	142	144	135	144	140	144	147	167	131	146	141	135	139
13	82	271	306	82	82	287	262	82	287	82	82	263	82	82	251	82
14	216	232	74	211	74	74	240	74	74	223	74	74	232	74	223	236
15	86	258	267	86	293	86	257	86	86	268	86	266	86	269	229	86
16	74	259	74	74	269	74	247	265	74	74	225	237	74	237	74	74

Note—t_i, number of times stimulus x_i was presented in the test phase.

(*Continued on next page*)

APPENDIX B
Calculating the Likelihood

Likelihood Function

Both the maximal likelihood and the marginal likelihood rely on the likelihood function $p(d \mid \theta, M)$, which indicates the fit between data d and model M at parameter value θ. The empirical data d consist of, for each of the m stimuli presented in the test phase, the counts k_i of the number of times stimulus x_i was chosen in Category A out of the t_i trials in which it was presented. It is assumed that the counts k_i follow a binomial distribution:

$$k_i \sim \text{Binomial}(p_i, t_i), \tag{B1}$$

where $p_i \equiv p(\text{A} \mid x_i)$, the response probability as defined in Equation 5.

Parameter Priors

Apart from the likelihood, the marginal likelihood also requires a prior distribution for each of the free parameters involved, which expresses the belief, before the data are collected, in which values of the parameters are likely and unlikely. On the basis of experience with previous related studies, or on theoretical grounds, a researcher can have a priori reasons to assume that some values are more likely than others. This information can be translated in an informative prior. In the absence of such information, an uninformative prior is more apt, which is intended to reflect a state of ignorance by not favoring any parameter values over others. Our analyses assumed uninformative priors for both parameters:

$$w \sim \text{Uniform}(0, 1) \tag{B2}$$

and

$$c^2 \sim \text{Gamma}(\varepsilon, \varepsilon), \tag{B3}$$

with $\varepsilon = .001$ set near zero. The Gamma(ε, ε) distribution is the standard (near) noninformative prior for a precision (see Spiegelhalter, Thomas, Best, Gilks, & Lunn, 1994).

Calculation of the Marginal Likelihood

The marginal likelihood in Equation 10 was evaluated by imposing a $W \times C$ grid on the two-dimensional parameter space, implying the following numerical approximation (see, e.g., Lee, 2004; Pitt et al., 2002):

$$p(d \mid M) \approx \frac{Z}{WC} \sum_{i=1}^{W} \sum_{j=1}^{C} p\left(d \mid w_i, c_j, M\right) p\left(w_i\right) p\left(c_j\right), \tag{B4}$$

where $\{w_1, w_2, \ldots, w_W\}$ and $\{c_1, c_2, \ldots, c_C\}$ are sets of parameter values defining the extent and the resolution of the grid. The constant Z reflects the ranges of w and c. All reported applications made use of a grid that extended from 0 to 1 for w and from 0.01 to 10 for c, increasing in steps of 0.001 for w and 0.01 for c.

APPENDIX C
Recovery Rates

Table C1
Recovery Rates (in %) for the 30 Data Sets Using the Maximal and Marginal Likelihood

Data Set	Maximal Likelihood	Marginal Likelihood	Data Set	Maximal Likelihood	Marginal Likelihood
1	98.00	98.15	16	31.07	31.97
2	97.40	97.36	17	94.44	94.72
3	77.99	78.66	18	86.93	87.32
4	83.37	83.69	19	78.15	79.31
5	17.38	17.63	20	56.74	57.06
6	26.36	26.26	21	66.91	67.21
7	63.31	63.38	22	88.20	88.32
8	48.46	50.07	23	87.76	88.23
9	49.18	49.47	24	87.80	87.05
10	98.82	98.80	25	96.29	96.76
11	48.50	48.64	26	97.27	97.56
12	88.68	88.86	27	97.52	97.72
13	39.61	40.00	28	62.41	62.79
14	87.38	87.55	29	84.79	85.97
15	54.14	54.33	30	93.40	93.71

(Manuscript received July 22, 2009;
revision accepted for publication October 24, 2009.)