Psychon Bull Rev
DOI 10.3758/s13423-012-0300-4

THEORETICAL REVIEW

Using priors to formalize theory: Optimal attention and the generalized context model

Wolf Vanpaemel · Michael D. Lee

© Psychonomic Society, Inc. 2012

Abstract Formal models in psychology are used to make theoretical ideas precise and allow them to be evaluated quantitatively against data. We focus on one important—but underused and incorrectly maligned—method for building theoretical assumptions into formal models, offered by the Bayesian statistical approach. This method involves capturing theoretical assumptions about the psychological variables in models by placing informative prior distributions on the parameters representing those variables. We demonstrate this approach of casting basic theoretical assumptions in an informative prior by considering a case study that involves the generalized context model (GCM) of category learning. We capture existing theorizing about the optimal allocation of attention in an informative prior distribution to yield a model that is higher in psychological content and lower in complexity than the standard implementation. We also highlight that formalizing psychological theory within an informative prior distribution allows standard Bayesian model selection methods to be applied without concerns about the sensitivity of results to the prior. We then use Bayesian model selection to test the theoretical assumptions about optimal allocation formalized in the prior. We argue that the general approach of using psychological theory to guide the specification of informative prior distributions is widely applicable and should be routinely used in psychological modeling.

Keywords Model selection · Bayesian statistics · Bayesian inference · Informative priors

W. Vanpaemel (✉)
Faculty of Psychology and Educational Sciences,
University of Leuven,
3000 Leuven, Belgium
e-mail: wolf.vanpaemel@ppw.kuleuven.be

M. D. Lee
Department of Cognitive Sciences, University of California,
Irvine, USA
e-mail: mdlee@uci.edu

Published online: 07 August 2012

Introduction

The Bayesian statistical framework is becoming increasingly important and popular for implementing and evaluating psychological models, including models of psychophysical functions (Kuss, Jäkel & Wichmann, 2005), stimulus representations (Lee, 2008), category learning (Lee & Vanpaemel, 2008; Vanpaemel & Storms, 2010), signal detection (Rouder & Lu, 2005), response times (Rouder, Lu, Speckman, Sun & Jiang, 2005) and decision making (Wetzels, Grasman & Wagenmakers, 2010). It is widely recognized in statistics (Gelman, Carlin, Stern & Rubin, 2004; Jaynes, 2003) and, increasingly, in psychology (Dienes, 2011; Kruschke, 2010, 2011; Lee & Wagenmakers, 2005) that the Bayesian approach offers a complete and coherent framework for making inferences using models and data. Bayesian parameter estimation allows model parameters to be estimated in a way that naturally represents uncertainty and is applicable even when there are few data. Bayesian model selection allows for models to be compared in a way that automatically implements an Ockham's razor for balancing goodness of fit with complexity, including when models are nonnested (Lee, 2008; Myung & Pitt, 1997; Shiffrin, Lee, Kim & Wagenmakers, 2008; Vanpaemel & Lee, 2012).

Priors as problems, priors as opportunities

Even advocates of the Bayesian approach, however, often view these benefits as coming at a cost, in the form of having to specify prior distributions. In Bayesian parameter estimation, placing priors on parameters is relatively uncontroversial because, as long as the data available are sufficiently informative, the choice of the prior has little impact on inference. In contrast, Bayesian model selection is generally sensitive to the exact choice of the prior. Because priors are often deemed to bring an unwanted level of arbitrariness to the conclusions, placing prior distributions

on parameters is much more controversial. For example, in their seminal paper on Bayesian model selection in psychology, Myung and Pitt (1997, p. 91) say that "the Bayesian method [for model selection], however, has its drawbacks. One is that parameter priors are required to compute the marginal likelihoods [and hence, the Bayes factor]." In practice, these objections have led researchers often to uniform, flat, or otherwise weakly informative priors, in an attempt to limit the information injected into the model selection (e.g., Kemp, Perfors & Tenenbaum, 2004; Lee, 2004; Pitt, Myung & Zhang, 2002).

We do not agree that priors are an unwanted aspect of the Bayesian framework, especially when dealing with psychological models. Unlike generic statistical or psychometric models, such as the generalized linear model, psychological models aim to formalize the mechanisms, processes, and representations underlying observed behavior. The variables that give rise to the behavior of interest are typically unknown and are, therefore, represented by parameters. Generally, knowledge about these parameters is gained by estimating their values from observed data. However, within models that formalize psychological theory, it is almost never the case that these variables are completely unknown. Knowledge about the values parameters are likely to take is often available *before* data are observed, on the basis of previous experience or psychological theorizing.

Expressing knowledge or assumptions about the parameters is difficult in orthodox approaches to inference but is both possible and necessary within the Bayesian approach (Jaynes, 2003; Lindley, 2004). In the Bayesian setting, the prior provides an opportunity to formalize theory in psychological models. The prior is, thus, an integral part of a psychological model, standing on the same footing as the data-generating function formalized by the likelihood. Therefore, we believe that it is wrong to malign priors as a necessary evil. Instead, priors should be welcomed by psychological modelers as yet another advantage of the Bayesian approach.

In this article, we present a concrete example of how an informative prior distribution can be used to capture important existing psychological theory. Our case study considers the generalized context model of category learning (Nosofsky, 1984, 1986). We show how basic existing theoretical assumptions—that people tend to allocate their attention over psychological dimensions so as to optimize their classification performance—can be formalized in an informative prior over parameter values. We highlight that incorporating theoretically meaningful informative priors has at least two important benefits. One is that having an additional mechanism for expressing theoretical assumptions in a model provides a way of constructing richer, stronger, and more complete models that are higher in psychological content and lower in complexity. Another benefit is that, when models have priors to which the theorist is committed, sensitivity to priors is no longer a problematic aspect of Bayesian model selection. This makes all the benefits of Bayesian model selection, such as automatically accounting for model complexity, readily available. Moreover, as we illustrate, Bayesian model selection methods can be used to evaluate empirically the theoretical assumptions represented by the prior.

Theory and standard implementation of the GCM

Nosofsky's (1984, 1986) generalized context model (GCM) is one of the most successful and influential models in cognitive psychology. It provides an account of how people learn and use categories, focusing on how people arrive at a decision concerning the membership of a stimulus in one of several categories. Extensions of the GCM deal with typicality judgments (Nosofsky, 1988) and response times (Nosofsky & Palmeri, 1997).

Like any good model, the GCM is built on a set of clear theoretical assumptions. We first describe those included in the standard implementation and then turn to an assumption—known as the attention-optimization hypothesis—that has failed to be formalized within the model.

The standard implementation of the GCM

Dimensional stimulus representation

The GCM assumes that stimuli can be represented by their values along underlying stimulus dimensions, as points in a multidimensional psychological space. Formally, the *i*th stimulus is represented in a *D*-dimensional psychological space by the point $x_i = (x_{i1}, \ldots, x_{iD})$.

Selective attention

A key contribution of the GCM to modeling category learning is the assumption that the structure of the psychological space is systematically modified by selective attention. Formally, the *k*th dimension has an attention weight, w_k with $0 \leq w_k \leq 1$, and $\sum_k w_k = 1$. These attention weights act to "stretch" attended dimensions and "shrink" unattended ones. The psychological distance between the *i*th and *j*th stimuli is given by the weighted city-block ($r = 1$) or Euclidean ($r = 2$) distance between the points, $d_{ij} = \left[\sum_k w_k \left| x_{ik} - x_{jk} \right|^r\right]^{\frac{1}{r}}$.

Similarity-based classification

The GCM assumes that classification decisions are based on similarity comparisons with the stored exemplars, with

similarity determined by a nonlinearly decreasing function of distance in the psychological space. Formally, the similarity between the ith and jth stimuli is modeled as an exponential ($\alpha = 1$) or Gaussian ($\alpha = 2$) decay function, $s_{ij} = exp\left(-cd_{ij}^{\alpha}\right)$, where c is a generalization parameter.

Exemplar representation

The GCM assumes that categories are represented by individual exemplars. This means that, in determining the overall similarity of a stimulus to a category, every exemplar in that category is considered. Formally, the overall similarity of the ith stimulus to category A is $s_{iA} = \sum_{j \in A} s_{ij}$.

Choice rule

The GCM assumes that classification decisions are based on the Luce choice rule (Luce, 1977), as applied to the overall similarities. Formally, the probability that the ith stimulus will be classified as belonging to category A, rather than to category B, is often modeled as $p_{iA} = \frac{s_{iA}}{s_{iA}+s_{iB}}$. Extended versions of this choice rule, making additional assumptions about response bias, and the determinism of responding, have also been considered (Ashby & Maddox, 1993; Navarro, 2007; Nosofsky, 1989).

Taken together, these five theoretical assumptions lead to the formal implementation of the GCM as it is usually tested against empirical data or used to estimate parameters. The empirical data are generally collected in a category-learning task, which presents participants with stimuli and their accompanying category labels and requires label prediction for novel stimuli.

Typically, the x_i points representing the stimuli are found by multidimensional scaling from separately collected similarity data (Shepard, 1962, 1980). The metric parameter r is set according to knowledge of the separability (r = 1) or integrality (r = 2) of the stimulus domain, and α according to knowledge of the discriminability of the stimuli. With these in place, the generalization parameter c and attention weights w_k remain as free parameters. Since most applications of the GCM rely on non-Bayesian methods for statistical inference, priors for these parameters are never explicitly formulated. Generally, it is implicitly assumed that all parameter values are equally likely. Under a Bayesian interpretation, this corresponds to uniform priors for the attention weights w_k and a flat prior for the generalization parameter c. We will call the implementation of the GCM using these priors GCM_{unif}.

The attention-optimization hypothesis

The idea that people selectively allocate their attention led to the introduction of the free attention parameters w_k. However, the existence of an attention weight is not the limit of theorizing, and some progress has been made in understanding the allocation of attention. In particular, the first GCM article suggested that "it is reasonable to hypothesize that, with learning, subjects will distribute attention among the component dimensions in a way that tends to optimize performance" (Nosofsky, 1984, p. 109), and the following seminal article said, "as a working hypothesis, it is assumed that subjects will distribute attention among the component dimensions so as to optimize performance in a given categorization paradigm. That is, it is assumed that the w_k parameters will tend toward those values that maximize the average percentage of correct categorizations" (Nosofsky, 1986, p. 41). In a later summary paper, Nosofsky (1998a, p. 218) observes that "the theme of optimal performance has always played a central part in theorizing involving the GCM."

Despite being one of the original key assumptions of the theory, the assumption that people learn to allocate their attention over psychological dimensions so as to optimize their classification performance is not incorporated formally in the standard implementation of the GCM. GCM_{unif} assumes all values of the attention weights to be equally likely and, thus, ignores the attention-optimization hypothesis. This is surprising, since the idea that people allocate their attention optimally seems to be a clear theoretical position that should—as with assumptions about representation, selective attention, similarity, selective attention, and choice behavior—be able to be formalized within a model.

We think that the reason the attention-optimization hypothesis was never formally implemented within the GCM is that, unlike the other assumptions, it is an assumption about the distribution of parameters. The exemplar assumption dictates how categories are represented, and the selective attention, similarity, and choice assumptions dictate how information is processed and behavior is produced in a category-learning task. The attention-optimization hypothesis, in contrast, does not directly affect how processing proceeds but, rather, provides information about the values of the attention parameters that partly control the processing. This sort of assumption is usually not easy to express in the likelihood function (i.e., the formalization of the process that is assumed to generate behavioral data). Rather, it is exactly the sort of information that can be expressed in the prior distribution over the possible values of the model parameters. Hence, extending the standard implementation of the GCM to formally incorporate the attention-optimization hypothesis is greatly facilitated by the adoption of the Bayesian framework.

An implementation of the GCM with optimal attention

In this section, we develop an implementation of the GCM's existing theoretical assumptions that extends the standard implementation by formally including the attention-optimization

hypothesis. This implementation of the GCM, which we will refer to as GCM_{opt}, is identical to GCM_{unif}, except for one key aspect. While GCM_{unif} does not include as part of its formal description any mechanism that corresponds to optimal attention, because it assumes all values of the attention parameters to be equally likely, GCM_{opt} explicitly embodies the assumption of optimal attention, because it places an informative prior over the attention parameters.

Our demonstration of completing the GCM to include an informative prior involves four category structures used by Nosofsky (1989). All of these structures are based on a set of 16 semicircles with an embedded radial line drawn from the center of the semicircle to the rim. The stimuli varied in the size of the semicircle and in the angle of orientation of the line, as shown in Nosofsky's (1989) Fig. 2. A two-dimensional representation of the stimulus space was constructed on the basis of similarity data derived from an identification experiment. All four category structures were defined using these 16 stimuli and are shown in Fig. 1. For all of the structures, three or four stimuli were assigned to category A, three or four to category B, and the remaining eight or nine stimuli were left unassigned. In two of the category structures, called "angle" and "size," attending to a single dimension is sufficient to learn the categories. The other structures, "crisscross" and "diagonal," require attention to both dimensions.

Formalizing the attention-optimization hypothesis

The stimuli are two-dimensional, so there is a single free attention weight parameter w, reflecting attention given to the angle dimension. One naive way to include the attention-optimization hypothesis in a formal implementation of the GCM is to fix this attention parameter to the single optimal value. Following Nosofsky (1984, 1986), this value can be found by maximizing the average proportion correct, $\mathrm{APC}(c,w) = \sum_{i\in A} p_{iA}(c,w) + \sum_{i\in B} p_{iB}(c,w)$. One complication with finding the optimal value is that APC simultaneously depends on both c and w, implying that there is no unique attention weight w maximizing APC. Previous authors have pragmatically solved this issue by fixing c to a single value, either estimated from the empirical data or specified independently of the data (Lamberts, 1995; Nosofsky, 1984, 1986; Rehder & Hoffman, 2005). Our remedy starts from the observation that, as c increases, self-similarity starts to dominate the total similarity, and consequently, the GCM starts to behave as a rote memorization model rather than as the exemplar-based generalization model it is intended to be. On the basis of this consideration, we used the following procedure for finding the optimal value of w. Out of a set of 100,000 (c, w) pairs that maximized APC, we considered 100 pairs that implied perfect performance with the smallest possible values of c. The most frequent value of w among these 100 pairs was selected as the optimal value for w. This procedure resulted in 0, 0.47, 0.50 and 1 as the optimal values for w in the angle, crisscross, diagonal, and size conditions, respectively.

Fixing the attention parameter to a single value fails to incorporate important levels of uncertainty and does not correspond well to the theoretical assumption, which is described in terms of tendencies and inclinations. For example, Nosofsky (1998b, p. 330) makes it clear that, for category structures in which both dimensions are relevant, the attention-optimization hypothesis does *not* predict that both dimensions will be allocated exactly equal attention ($w = \frac{1}{2}$). What the hypothesis does predict is that, for such structures, extreme values of the attention weight are very unlikely. Similarly, the attention-optimization hypothesis posits that, for category structures in which only a single dimension is relevant for performing the classification, observers will be inclined to attend exclusively to this dimension, which translates to the attention weights tending toward extreme values.

Thus, for a given category structure, the attention-optimization hypothesis corresponds to a tendency to optimality that does not predict precise values of the attention parameter. Rather, it favors a range of attention values, some of which are considered more likely than others. It is therefore more natural to express the attention-optimization hypothesis in terms of a *distribution* over the possible values of the attention parameter. We view this distribution as uncertainty on the part of the theorist, representing the relative likelihood they assign to different possible attention values a participant in an experiment might be using. Under

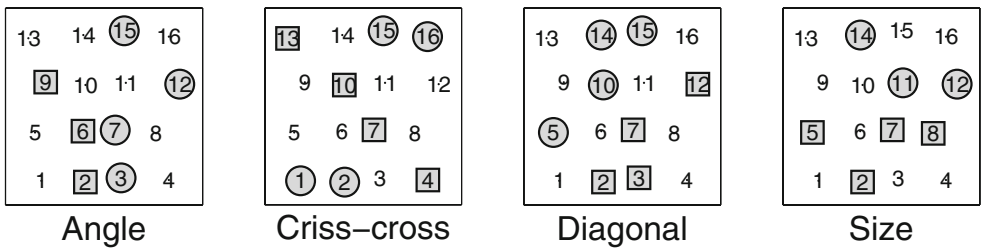


Fig. 1 The four category structures used by Nosofsky (1989). Each numbered location corresponds to a stimulus, with the stimuli assigned to the two categories shown by squares (**c**ategory A) and circles (**c**ategory B) for each structure. The horizontal dimension corresponds to the angle of orientation of the line, whereas the vertical dimension corresponds to the size of the semicircle. Based on Nosofsky (1989, Fig. 1 and Fig. 3)

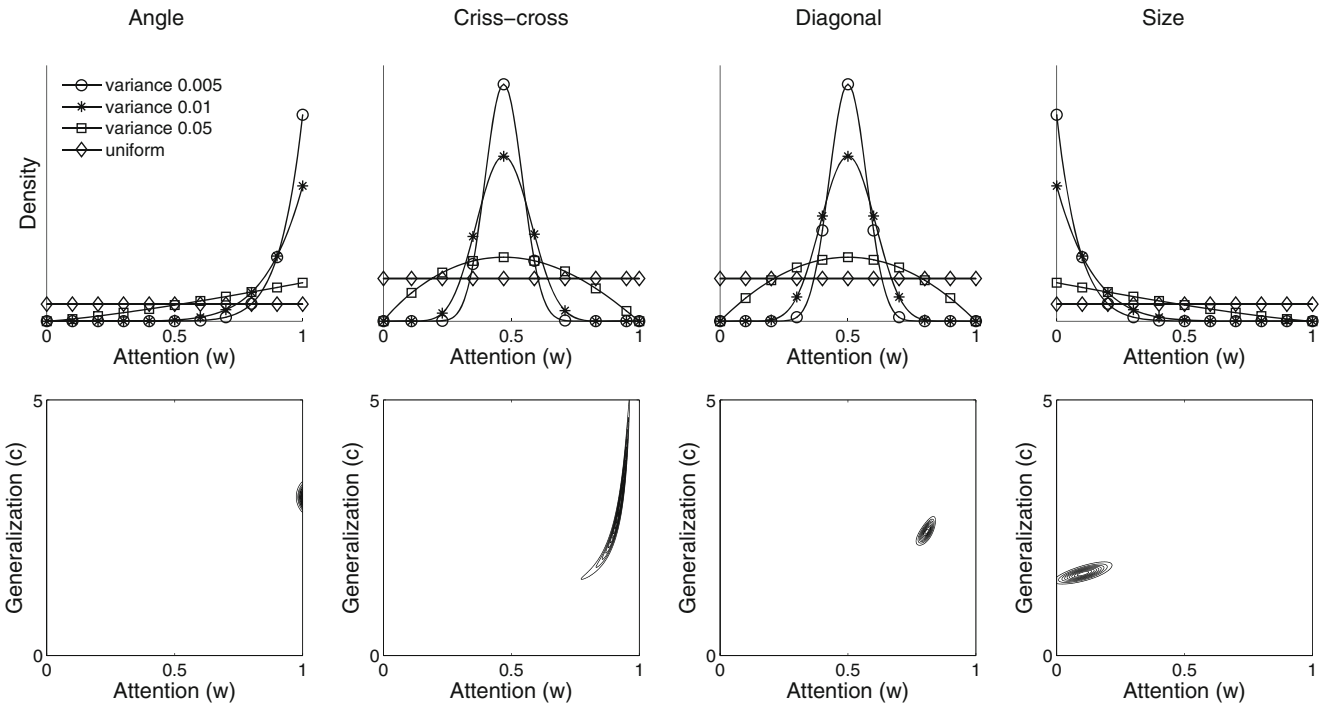


Fig. 2 The first row shows distributions over the *w* parameter, which reflects attention given to the angle dimension. These distributions correspond to expectations about the allocation of attention for Nosofsky's (1989) angle, crisscross, diagonal, and size category structures. Each panel shows the prior assumed in different versions of GCM_{opt}, differing in the variance around the optimal *w* value, and the uniform prior, as assumed in GCM_{unif}. The second row shows the likelihood contour plots for both GCM_{unif} and GCM_{opt}, indicating the level of fit on the Nosofsky (1989) data, across the parameter space ranging from 0 to 1 for the attention parameter *w* and from 0 to 5 for the generalization parameter *c*. Comparing the prior and the likelihood reveals that, in the angle and size conditions, good-fitting values are given a high prior weight by GCM_{opt}, implying a high average fit, whereas in the crisscross and diagonal conditions, GCM_{opt} assigns a low prior weight to good-fitting values, which results in a low average fit

this view, the uncertainty corresponds exactly to the Bayesian prior.

As a verbal expression of theory, the attention-optimization hypothesis is insufficiently informative to specify the exact width of such a distribution. Therefore, formally implementing the hypothesis can be achieved in a number of reasonable ways. The first row of Fig. 2 shows, for each category structure separately, a uniform distribution, as well as three distributions centered on the optimal value of *w*. These distributions differ in their width and, thus, in the level of uncertainty they represent. Formally, beta distributions were derived with variances of 0.005 (most certain), 0.01, and 0.05 (least certain). For the angle and size structures, values close to 1 or 0, respectively, are preferred. For the crisscross and diagonal structures, values around $\frac{1}{2}$ are most favored, but other nonextreme values are considered likely as well. These theory-informed distributions thus seem consistent with the attention-optimization hypothesis and represent the attention values GCM_{opt} believes correspond to people's attention.

Discussion

GCM_{opt}, which expresses existing psychological theory in an informative prior over the attention weight, contrasts favorably with GCM_{unif}, which considers all values of the attention parameter to be equally likely. One important strength of GCM_{opt} is that it is closer to the theoretical foundations of the GCM, as outlined by Nosofsky in the 1980s, than is GCM_{unif}. Additionally, GCM_{opt} is richer in psychological content, makes more assertive claims about learning, and is less complicated, in the formal sense that it makes fewer predictions about possible categorization behavior. The different versions of GCM_{opt} correspondingly differ in psychological content, with a smaller variance corresponding to more precision and less uncertainty and, hence, a stronger, simpler model. In fact, GCM_{unif}, having a variance of $\frac{1}{12} \approx 0.08$, could be viewed as a weak version of GCM_{opt}, as is shown in the first row of Fig. 2.

More generally, theory-informed priors help build models that are more complete, are theoretically richer, and make more stringent predictions than their counterparts with noninformative or weakly informative default priors. Casting theoretical

assumptions in an informative prior increases what Popper (1959) has termed the empirical or informative content of a model. Models with a high empirical content are high in assertive power because they have more to say about our world than do models with a low empirical content. Additionally, using priors to formalize theory tends to reduce the model's complexity. Just as fixing a parameter to a single value reduces the number of free parameters and so decreases the model's complexity, assigning an informative prior to a free parameter makes the parameter less free and so decreases the flexibility of the model (Vanpaemel, 2009).

A further benefit of using theory to define informed prior distributions over parameters involves the merits of model selection methods, such as Bayes factors, that integrate over the parameter space (Kass & Raftery, 1995). While these model selection methods have the well-known and attractive feature of providing an automatic Ockham's razor, a serious reservation about these methods is that the results can be very sensitive to priors, including often near-arbitrary decisions such as how to bound the parameter space (Liu & Aitkin, 2008). Essentially unrealistic or unprincipled bounding can be found in, for example, Lee (2004) or Pitt et al. (2002). These sorts of problems are substantially addressed, however, if the prior distributions are not arbitrary but come directly from theory. Provided the priors capture meaningful theory, as for GCM_{opt}, the prior sensitivity of the Bayes factor is unproblematic, so the Bayes factor becomes a viable and powerful approach for comparing the GCM with other models of category learning, such as the prototype model (Reed, 1972) or models from the general recognition theory framework (Ashby & Townsend, 1986).

Moreover, when priors express theory, the sensitivity of the Bayes factor to the prior is an advantage. Model selection methods that are insensitive to the prior, such as the Akaike information criterion (Akaike, 1974) or the Bayesian information criterion (Schwarz, 1978), fail to evaluate the theory expressed in the prior of GCM_{opt} and, thus, provide only a partial evaluation of the theory formalized in the model (Vanpaemel, 2010). Using the Bayes factor to contrast GCM_{opt} with competing models ensures that the GCM is evaluated in a way that takes into account existing theorizing about attention and, at the same time, acknowledges that this theorizing is incomplete.

Using model selection to test the attention-optimization hypothesis

Besides producing psychologically richer models and freeing theorists from concerns about the sensitivity of model selection results to the prior, the use of theory-informed priors has an important by-product. If priors are used for expressing theories about parameters, these theories themselves can be tested using model selection methods that are sensitive to the prior, such as the Bayes factor. In particular, since the prior over the attention weights is the only difference between GCM_{unif} and GCM_{opt}, the empirical selection between these models amounts to a direct quantitative evaluation of the optimal attention theory itself.

Our demonstration of using the completed implementation of the GCM to test the theory captured in the prior relies on the data Nosofsky (1989) collected using the four category structures shown in Fig. 1. Each category structure was learned by a separate group of participants under a standard training–test procedure. Nosofsky (1989) restricted modeling analyses to those participants who achieved a sufficiently high accuracy in classifying the assigned stimuli during the final 125 trials of the training phase, leaving 41, 37, 43, and 37 participants for the angle, crisscross, diagonal, and size structures, respectively. The data reported in Nosofsky's (1989) Table 4 are the counts of category A responses made during the test phase, aggregated over participants.

Nosofsky (1989) applied the Gaussian similarity and Euclidean distance versions of the GCM and, as in the present article, considered a version of the GCM without response bias and response determinism, among several other versions. Nosofsky (1989) found that the standard implementation of the GCM accounted for 98.7 %, 91.1 %. 98.6 %, and 99.4 % of the variance for the angle, crisscross, diagonal, and size conditions, respectively . Consistent with the attention-optimization hypothesis, Nosofsky (1989) estimated the attention given to the angle dimension to be 1.00 in the angle condition and 0.10 in the size condition. Quite unexpectedly, the attention weight was estimated to be 0.93 and 0.81 in the crisscross and diagonal conditions, respectively (Nosofsky, 1989, Table 5).

Testing the attention-optimization hypothesis

Both GCM_{unif} and GCM_{opt} assume the data-generating process given earlier and assume joint independence in the prior distribution, so that $p(w, c) = p(w)p(c)$. Further, GCM_{unif} and GCM_{opt} use a uniform prior for c, so that $c \sim \text{Uniform}(0, 5)$.[1] The only difference between GCM_{unif} and GCM_{opt} is the prior for w. GCM_{unif} is noncommittal about allocation of attention, so that $w \sim \text{Uniform}(0, 1)$. GCM_{opt}, in contrast, assumes that people tend to optimally allocate attention, as expressed in the informative priors for w. The priors assumed by GCM_{unif} and the three versions of GCM_{opt} are shown in the first row of Fig. 2.

To compare GCM_{unif} and GCM_{opt}, we calculated the Bayes factor. Let $y = (k_1, \ldots, k_{16})$ denote the observed category-learning data, where k_i is the number of times the

[1] This uniform prior reflects the fact that no psychological theory is currently available for the generalization parameter (although Nosofsky & Johansen, 2000, make some initial suggestions in this regard). Ideally, theorizing should be developed that allows the specification of a joint prior over w and c.

*i*th stimulus is classified as being in category A. Formally, the Bayes factor based on these data is given by the ratio of marginal likelihoods $BF_{uo} = \frac{p(y|GCM_{unif})}{p(y|GCM_{opt})}$. Each marginal likelihood integrates over the prior distribution of parameters, so that $p(y|GCM) = \int p(y|w, c, GCM)p(w, c|GCM)d(w, c)$, where the likelihood $p(y|w, c, GCM)$ indicates the level of fit the GCM provides to the data y at parameter values w and c.

A crucial property of the marginal likelihood is that it finds the average fit of each model to the data, across the entire prior distribution of parameters. If a model assigns high prior weights to parameter values likely to have generated the observed data, the model will be rewarded with a high marginal likelihood. In contrast, if the model favors parameter values that are unlikely to have generated the observed data and, thus, lead to bad fits, the model will be punished with a low marginal likelihood. Thus, if people use optimal attention weights consistent with the assumptions of GCM_{opt}, the marginal likelihood will reveal a better average fit. On the other hand, if people make classification decisions consistent with nonoptimal attention allocation, the marginal likelihood will prefer GCM_{unif}. A $BF_{uo} > 1$ then indicates that, on the basis of the evidence provided by the data, GCM_{unif} is preferred, whereas a $BF_{ou} > 1$ corresponds to empirical evidence for GCM_{opt}.

The second row of Fig. 2 shows, for each of the four category structures, contour plots of the likelihood across the relevant parameter space, indicating the level of fit GCM_{unif} and GCM_{opt} provide to the empirical data. Combining these likelihoods with the priors shown in the first row reveals that, for the angle and size structures, GCM_{opt} assigns high prior weights to those values that produce good fits to the data and low prior weights to those values that are unlikely to have produced the data. For the crisscross and diagonal structures, in contrast, the parameter values that are given a high prior weight produce bad fits, and the good-fitting parameter values are given a low prior weight. It thus seems that the Bayes factor supports GCM_{opt} in the angle and size conditions, but not in the crisscross and diagonal conditions.

We evaluated each marginal likelihood using standard numerical grid sampling, using steps of 0.001 for w and 0.005 for c. The calculated marginal likelihoods for GCM_{unif} and the three versions of GCM_{opt}, expressed on the standard logarithmic scale, are presented in Table 1. The choice of 5 as the upper bound of the uniform distribution as the prior for c was largely unprincipled, reflecting a serious incompleteness in theorizing about the generalization parameter. When priors are not completely conceptually grounded, it is crucial that a sensitivity analysis is performed. Reassuringly, the Bayes factors reported below are almost identical using alternative values of 10, 20, 50, and 100 as upper bounds for c.

Table 1 Marginal likelihoods, expressed as minus log likelihoods, for **GCM_{unif}** and for the three versions of **GCM_{opt}** considered, differing in the variance of the prior, for each of the four category-learning tasks considered by Nosofsky (1989)

	GCM_{opt} (variance 0.005)	GCM_{opt} (variance 0.01)	GCM_{opt} (variance 0.05)	GCM_{unif}
Angle	61.48	61.87	63.08	63.88
Crisscross	69.02	67.01	62.87	62.05
Diagonal	63.43	58.84	55.33	55.26
Size	44.75	44.80	45.47	46.14

The difference between the log marginal likelihoods for two models is the log of the Bayes factor that compares them. Thus, for the angle category structure, BF_{ou} shows that the strongest version of GCM_{opt} is $exp(63.88 - 61.48) \approx 11.0$ times more likely than GCM_{unif}, indicating strong evidence for optimal attention. When GCM_{opt} makes less strong assumptions about attention allocation, it becomes more similar to GCM_{unif}. This is reflected by a drop in BF_{ou} to $exp(63.88 - 61.87) \approx 7.5$, which indicates still substantial evidence for GCM_{opt}. The evidence in favor of the weakest version of GCM_{opt} is anecdotal, with a Bayes factor of $exp(63.88 - 63.08) \approx 2.2$. Similarly, for the size category structure, there is substantial evidence for the two strongest versions of GCM_{opt}, with $BF_{ou} = exp(46.14 - 44.75) \approx 4.0$ and $exp(46.14 - 44.80) \approx 3.8$, respectively. The Bayes factor further decreases to $exp(46.14 - 45.47) \approx 2.0$, together with the strength of the theory expressed in the prior of GCM_{opt}. In sum, in the angle and size conditions, the strongest versions of GCM_{opt} are rewarded for assigning a high prior weight to parameter values that provide a good fit and a low prior weight to bad-fitting parameter values.

For the crisscross and diagonal structures, in contrast, BF_{uo} shows that GCM_{unif} is, respectively, $exp(69.02 - 62.05) > 1000$ and $exp(63.43 - 55.26) > 3500$ times more likely than the strongest version of GCM_{opt}. If the precision of the prior of GCM_{opt} declines, GCM_{unif} is still very strongly supported by the data, although to a much lesser degree, with Bayes factors $exp(67.01 - 62.05) \approx 141.6$ and $exp(58.84 - 55.26) \approx 35.7$. Comparing the weakest version of GCM_{opt} with GCM_{unif} still provides evidence for GCM_{unif}, but this evidence is anecdotal, with Bayes factors of $exp(62.87 - 62.05) \approx 2.3$ and $exp(55.33 - 55.26) \approx 1.1$ for the crisscross and diagonal conditions, respectively. In sum, in the crisscross and diagonal conditions, the strongest versions of GCM_{opt} are penalized for assigning a high prior weight to parameter values that provide a bad fit and a low prior weight to good-fitting parameter values.

Discussion

Overall, in the two category structures where a single dimension is relevant for performing a classification, there is some evidence for attention optimization. In the two category structures

where both dimensions are relevant, there is overwhelming evidence for the implementation of the GCM using uniform priors. This implies that, in these conditions, participants do not seem to allocate their attention optimally over both of the stimulus dimensions (Nosofsky, 1998b; Nosofsky & Johansen, 2000). These conclusions are consistent with those of Nosofsky (1989), made on the basis of point parameter estimates found by fitting the GCM without explicit priors on parameters.

Our goal in this article has not been to provide an exhaustive test of the optimal attention hypothesis or to develop additional psychological theorizing about how people allocate attention to stimulus dimensions when learning categories. Our goal was to advocate a more complete Bayesian approach to model building and evaluation that can underpin these sorts of developments. To this end, our model selection approach for testing the attention-optimization hypothesis contrasts favorably with the usual approach of investigating the attention-optimization hypothesis. Standard practice uses a model that explicitly does *not* try to capture the relevant theory and then relies on post hoc assessment of parameter estimates to reach conclusions. The approach taken in this article—testing a model that explicitly formalizes the assumption—follows one of the rationales for building formal models in psychology, which is to make psychological theories precise and testable.

Our case study illustrates the benefits of taking a strong position, since the exact level of the evidence against or in favor of GCM_{opt} depends on the strength of the theory expressed in its prior. A model making strong assumptions has the advantage that when the assumptions are consistent with behavior, the evidential reward is large, as illustrated in the angle and size conditions. If a model making bold claims does not accommodate the data well, however, the evidential penalty is large, as illustrated in the crisscross and diagonal conditions. Stronger tests of theories and higher scientific payoffs come from more complete models.

General discussion

Formal models in psychology are often used to instantiate substantive theory about a particular aspect of cognitive behavior in a quantitative way. Often, however, theory about psychological variables is incomplete, calling for the use of parameters in models. Generally, the parameters are assigned a flat, uniform, or otherwise weakly informative prior distribution. We believe that this state of affairs is unfortunate and presents a challenge to the field. When theory is totally mute on the values of parameters, the use of non- or weakly informative priors logically follows. However, when the models aim to formalize theory, this theory, while usually incomplete, often has at least something to say. Within the coherent Bayesian approach to statistical inference, it is possible to include even incomplete and inexact theorizing about psychological variables within informative prior distributions on the parameters representing these variables. Specifying an informative prior lies between succumbing to the bland specification of a uniform or flat prior distribution, as if no theory at all is available, and overconfidently fixing a parameter to a single value, as if theory is complete and exact.

To illustrate that prior distributions over parameters can capture psychological theory and should, therefore, be considered to be integral parts of a psychological model, we focused on the GCM, which contains parameters corresponding to dimensional attention. Current theorizing about attention allocation is not yet developed to the point at which the attention parameters can be specified precisely, so the attention parameters represent an important source of incompleteness in theorizing about attention during category learning (Nosofsky, 1998b). However, some theory is available, relating to optimal attention allocation. This theory can be formally represented in an informative prior on the attention parameter, which results in an attention parameter that is neither entirely free nor precisely constrained.

More generally, we think our demonstration of the ability of prior distributions to carry theoretical information underscores an underappreciated merit of the Bayesian approach. Non-Bayesian statistical frameworks do not make it easy to include theorizing about more or less likely combinations of parameter values in models. This is unfortunate, since specifying informative prior distributions over the parameters is an important avenue for expressing theoretical ideas and has several attractive advantages. One advantage is that expressing psychological theory in a prior allows theory represented by the prior to be evaluated directly using Bayesian model selection methods. Another benefit is that, once priors are firmly based on theory, concerns about the sensitivity of the Bayes factor to the prior are alleviated. Using the Bayes factor for model selection naturally takes into account existing theory about the variables represented by the parameters, without necessitating their overconfident reduction to precise values. A third benefit is that informative priors simultaneously decrease the flexibility and increase the empirical content of the model.

The approach we have advocated and demonstrated is applicable in any situation where theory has something to say about which combinations of parameter values in a model are more likely than others. In psychological models, this should very often be the case, because parameters usually correspond to meaningful psychological variables and are a primary focus of theory.

There are many ways theorists can formalize theory within the Bayesian approach. One common possibility is to place order constraints on parameters (Hoijtink, Klugkist & Boelen, 2008; Mulder, Klugkist, van de Schoot, Meeus, Selfhout & Hoijtink, 2009) or to constrain the range of a

parameter (Navarro, Pitt & Myung, 2004; Vanpaemel, 2010). More advanced methods involve maximum entropy priors, where complicated constraints on variables can be embedded in distributions (Jaynes, 2003, Ch. 9), and hierarchical Bayesian approaches, where theories about where the basic model parameters themselves come from are implemented formally within the model (Lee & Vanpaemel, 2008; Vanpaemel, 2011).

Of course, it will often be challenging to translate expectations about psychological variables into formal statements about prior probability distributions. This is exactly the same sort of challenge as is faced in building any sort of formal model. The original formulation of the GCM must have presented problematic challenges when casting basic theoretical assumptions as formal model mechanisms. Assumptions about stimulus representation took the form of points in a psychological space, rather than other representational possibilities; assumptions about similarity-based categorization took the form of families of exponential curves relating distance to similarity, rather than other possible generalization gradients; and assumptions about choice took the form of the Luce choice rule, rather than alternative choice functions. In the same way, the particular approach we used to formalize the optimal attention assumption is surely not the only possibility, but it is a formal and testable implementation consistent with the basic theoretical motivation.

Most generally, we think the belief—held, in our experience, by many who construct and test psychological models—that priors on parameters should be as vague and noninformative as possible is misguided. When researchers build models, a primary goal is to capture regularities in data, finding the principles and processes that govern people's behavior. In this sense, model building aims to make informed predictions about data, and the goal is certainly not to be noninformative and render all possible data sets and behaviors equally likely. We think the same argument applies to the psychological variables represented by parameters. A key goal of modeling is to develop theories that make informative statements about these variables, so that placing noninformative priors on parameters is the hallmark of theoretical immaturity. Using a noninformative prior is something that researchers should not celebrate but, rather, strive to overcome.

Author Note We thank Chris Donkin, Eric-Jan Wagenmakers, Jeff Rouder, Rob Nosofsky, Zoltan Dienes, and several anonymous reviewers for comments on earlier drafts. W.V. acknowledges the support of research grants from the KU Leuven Research Council (OT/11/032 and CREA/11/005). M.D.L. acknowledges the support of KULeuven/BOF Senior Fellowship SF/08/015.

References

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

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

Ashby, F. G., & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological Review, 93,* 154–179.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science, 6,* 274–290.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian Data Analysis (Second edition)*. Boca Raton, FL: Chapman & Hall/CRC.

Hoijtink, H., Klugkist, I., & Boelen, P. (2008). *Bayesian Evaluation of Informative Hypotheses*. New York: Springer.

Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. Cambridge, UK: Cambridge University Press.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90,* 377–395.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2004). Learning domain structures. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the 26th Annual Conference of the Cognitive Science Society* (pp. 720–725). Mahwah, NJ: Erlbaum.

Kruschke, J. K. (2010). What to believe: Bayesian methods for data analysis. *Trends in Cognitive Sciences, 14,* 293–300.

Kruschke, J. K. (2011). *Doing Bayesian Data Analysis: A Tutorial with R and BUGS*. Academic Press/Elsevier.

Kuss, M., Jäkel, F., & Wichmann, F. A. (2005). Bayesian inference for psychometric functions. *Journal of Vision, 5,* 478–492.

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

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

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

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.

Lindley, D. V. (2004). That wretched prior. *Significance, 1,* 85–87.

Liu, C. C., & Aitkin, M. (2008). Bayes factors: Prior sensitivity and model generalizability. *Journal of Mathematical Psychology, 52,* 362–375.

Luce, R. (1977). The choice axiom after twenty years. *Journal of Mathematical Psychology, 15,* 215–233.

Mulder, J., Klugkist, I., van de Schoot, R., Meeus, W. H. J., Selfhout, M., & Hoijtink, H. (2009). Bayesian model selection of informative hypotheses for repeated measurements. *Journal of Mathematical Psychology, 53,* 530–546.

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

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., Pitt, M. A., & Myung, I. J. (2004). Assessing the distinguishability of models and the informativeness of data. *Cognitive Psychology, 49,* 47–84.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 104–114.

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

Nosofsky, R. M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 700–708.

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

Nosofsky, R. M. (1998a). Optimal performance and exemplar models of classification. In M. Oaksford & N. Chater (Eds.), *Rational Models of Cognition* (pp. 218–247). London: Oxford University Press.

Nosofsky, R. M. (1998b). Selective attention and the formation of linear decision boundaries: Reply to Maddox and Ashby (1998). *Journal of Experimental Psychology: Human Perception and Performance, 24,* 322–339.

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

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

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

Popper, K. R. (1959). *The Logic of Scientific Discovery*. London: Hutchinson.

Reed, S. (1972). Pattern recognition and categorization. *Cognitive Psychology, 3,* 382–407.

Rehder, B., & Hoffman, A. B. (2005). Thirty-something categorization results explained: Selective attention, eyetracking, and models of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31,* 811–829.

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review, 12,* 573–604.

Rouder, J. N., Lu, J., Speckman, P. L., Sun, D., & Jiang, Y. (2005). A hierarchical model for estimating response time distributions. *Psychonomic Bulletin & Review, 12,* 195–223.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6,* 461–464.

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika, 27,* 125–140.

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science, 214,* 390–398.

Shiffrin, R. M., Lee, M. D., Kim, W.-J., & Wagenmakers, E.-J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science, 32,* 1248–1284.

Vanpaemel, W. (2009). Measuring model complexity with the prior predictive. *Advances in Neural Information Processing Systems, 22,* 1919–1927.

Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology, 54,* 491–498.

Vanpaemel, W. (2011). Constructing informative model priors using hierarchical methods. *Journal of Mathematical Psychology, 55,* 106–117.

Vanpaemel, W., & Lee, M. D. (2012). The Bayesian evaluation of categorization models: Comment on Wills and Pothos (2012). *Psychological Bulletin*. doi:10.1037/a0028551

Vanpaemel, W., & Storms, G. (2010). Abstraction and model evaluation in category learning. *Behavior Research Methods, 42,* 421–437.

Wetzels, R., Grasman, R. P. P. P., & Wagenmakers, E.-J. (2010). An encompassing prior generalization of the Savage-Dickey density ratio test. *Computational Statistics and Data Analysis, 54,* 2094–2102.