Using Priors to Formalize Theory: Optimal Attention and the Generalized Context Model

Wolf Vanpaemel
Department of Psychology
University of Leuven

Michael D. Lee
Department of Cognitive Sciences
University of California, Irvine

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 under-used, method for building theoretical assumptions into formal models, offered by the Bayesian statistical approach. This method involves capturing theoretical assumptions about the psychological variables represented by model parameters, by placing appropriate prior distributions on those parameters. We demonstrate this idea by implementing a version of the seminal Generalized Context Model (GCM) of category learning to incorporate theorizing about the optimal allocation of attention. We show how theoretical assumptions like optimal attention can be formalized within a prior distribution, and then be evaluated against data using standard Bayesian model comparison methods.

Introduction

Bayesian methods are becoming increasingly important and popular tools for implementing and evaluating psychological models, including models of stimulus representation (e.g., Lee, 2008), category learning (e.g., Lee & Vanpaemel, 2008), signal detection (e.g., Rouder & Lu, 2005), response times (e.g., Rouder, Lu, Speckman, Sun, & Jiang, 2005; Vandekerckhove, Tuerlinckx, & Lee, 2008) and decision-making (e.g., Wetzels, Vandekerckhove, Tuerlinckx, & Wagenmakers, in press).

The most obvious, and widely recognized, advantage of the Bayesian approach is that it offers a complete and coherent framework for relating models to data. Bayesian methods allow for principled statistical inferences to be made about models and their parameters, and implement an automatic 'Ockham's razor' for balancing goodness-of-fit with complexity when comparing models (e.g., Lee, 2008; Pitt, Myung, & Zhang, 2002; Shiffrin, Lee, Kim, & Wagenmakers, 2008). In this paper, we discuss a second inherent advantage of the Bayesian approach, involving the (ironically, sometimes maligned) need to specify a prior distribution over model parameters. In particular, we argue that prior distributions can often be a natural vehicle for formalizing theoretical knowledge or assumptions that are otherwise difficult to incorporate in a model.

The goal of building a formal model is to make theoretical ideas precise, and amenable to quantitative evaluation. In traditional non-Bayesian approaches to psychological modeling, all of the relevant theoretical assumptions have to be expressed within a function that describes how psychological variables, represented by parameters, generate behavior on a task. The parameters themselves can then be estimated from data, but there is no modeling mechanism for expressing knowledge about the parameters before data are collected. This seems an incomplete approach, because the values of the psychological variables they represent are rarely, if ever, completely unknown. Usually, something is known about parameters—based on theory, or previous experience—before data are collected.

Recognizing the need to represent information about parameters, the Bayesian framework augments the data-generating function with a prior distribution. This prior expresses knowledge or assumptions about the values parameters are likely to take, and the relationship between them. This means that, in psychological models, theoretical assumptions or other relevant information about psychological variables can be captured by prior distributions over the parameters.

In this paper, we present an example, involving the Generalized Context Model (GCM: Nosofsky, 1984, 1986), where basic theoretical assumptions can be used to derive a richly structured prior distribution over parameter values. We use previous data to compare this Bayesian implementation of the GCM with the traditional implementation, which does not include these theoretical assumptions. In this way, we show how using the priors of Bayesian methods enables a direct evaluation of the theory behind the model.

Theory and Implementation of the GCM

The Standard GCM

Nosofsky's (1984, 1986) GCM model of category learning is one of the most successful and influential models in cognitive psychology. It has been thoroughly empirically evaluated, and widely used. Like any good model, it is built on a set of clear theoretical assumptions, which we describe in this section, to motivate the formal definition of the model.

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, in the model, the ith stimulus is represented in a D-dimensional psychological space by the point $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iD})$.

Selective Attention. A key contribution of the GCM to modeling category learning is the assumption that the structure of psychological space is systematically modified by selective attention. Formally, the dth dimension has an attention weight, w_d with $0 \leq w_d \leq 1$, and $\sum_d w_d = 1$. These attention weights act to 'stretch' attended dimensions, and 'shrink' unattended ones. In the attention-weighted space, the psychological distance between the ith and jth stimuli is given by the attention-weighted Minkowski r-metric distance between the points, $d_{ij} = \left[\sum_d w_d \left|x_{id} - x_{jd}\right|^r\right]^{1/r}$, with $r = 1$ typically chosen for separable stimuli, and $r = 2$ chosen for integral stimuli (Garner, 1974).

Similarity-based Classification. The GCM assumes classification decisions are based on similarity comparisons with the stored exemplars, with similarity determined as a nonlinearly decreasing function of distance in the psychological space. Formally, the similarity between the ith and jth stimuli is usually modeled as an exponential decay function, $s_{ij} = \exp\left(-cd_{ij}\right)$, where c is a generalization parameter. Other similarity functions, usually within the exponential family, such as the Gaussian, have also occasionally been considered.

Exemplar Representation. Another important assumption of the GCM is that categories are represented by individual exemplars. This means that, in determining the overall similarity of a presented stimulus i to Category A, every exemplar in that category is considered. Formally, the overall similarity is $s_{iA} = \sum_{j \in A} s_{ij}$.

Choice Rule. The GCM assumes classification decisions are based on the Luce Choice rule, as applied to the overall similarities. Formally, the probability that the ith stimulus will be classified as belonging to Category A, rather than Category B, is often modeled as $\Pr\left(i \in A\right) = s_{iA} / \left(s_{iA} + s_{iB}\right)$. Extended versions of this choice rule, making additional assumptions about response bias, and the determinism of responding, have also been considered.

The Optimal Attention Assumption

Taken together, these five theoretical assumptions lead to the formal implementation of the GCM as it is usually tested against category learning data. Typically, the points representing the stimuli are found by multidimensional scaling methods from separately collected similarity data (Shepard, 1962, 1980). The Minkowski parameter r is set according to knowledge of the separability or integrality of the stimulus domain. With these in place, the generalization parameter c and attention weights w_d remain as free parameters.

This standard implementation of the GCM theory fails, however, to incorporate one of the original key assumptions of the theory. This assumption relates to optimal attention, and the idea that people learn to distribute their attention over psychological dimensions so as to optimize their classification performance. The first GCM paper suggests 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 seminal following paper 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" (Nosofsky, 1986, p. 41). In a later summary paper, Nosofsky (1998a) argued that "The theme of optimal performance has always played a central role in theorizing involving Nosofsky's (1986) generalized context model."

The idea that, based on experience, people distribute their attention optimally, seems to be a clear theoretical position that should—as with assumptions about representation, similarity, and choice behavior—be able to be formalized within a model. Yet the GCM has never included as part of its formal description any mechanism that corresponds to optimal attention. Usually, conclusions about optimal attention have been drawn from post-hoc hypothesis tests on estimated attention weight parameters (e.g., Nosofsky, 1986). Occasionally, the traditional GCM, in which the attention weight can vary as a free parameter, has been compared—on the basis of a goodness-of-fit measure—to a constrained GCM, in which the attention parameter is fixed at a theoretically interesting value (e.g., Nosofsky, 1989).

It is clear, however, especially from Nosofsky (1998b), that the optimal attention assumption is not intended to remove the attention weights as free parameters by fixing them to single point values. This fixing would be appropriate, for example, if people were assumed always to distribute their attention mathematically optimally in full knowledge of category structures. But the theoretical claim is that optimality makes some patterns of the distribution of attention more likely than others, based on what people can learn from the stimuli they encounter and the feedback they are given. This kind of knowledge is not expressed using a fixed single parameter value, but rather by a distribution over the possible values. Because this distribution is specified before empirical data are collected, it is referred to as a prior distribution. Appropriate priors for the attention weights are therefore not single points corresponding to the optimal values, but distributions based on optimally using the information available during a category learning task.

We think the reason the optimal attention assumption was never formally implemented within the GCM is that, unlike the other assumptions, it is an assumption about parameters. The exemplar assumption dictates how categories are represented, and the similarity and choice assumptions dictate how information is processed and behavior produced in a category learning task. The optimal attention assumption, in contrast, does not affect how processing proceeds, but rather provides information about the values of the parameters that control the processing. This is exactly the sort of information that can be expressed in the prior distribution over model parameters, and so requires the adoption of the Bayesian framework.

An Optimal-Attention GCM

In this section, we extend the traditional GCM to create a new model, GCM_{opt}, that includes a prior distribution over the attention weight parameters corresponding the optimality assumption. Formally, a category learning task involves a person seeing how N stimuli are correctly classified over a total of T trials, with the ith stimulus presented k_i times, and $\sum_{i=1} k_i = T$. With this information, an optimal learner can update a non-informative set of initial beliefs about the appropriate distribution of attention, and update those beliefs using Bayes Rule. These updated beliefs will make some patterns of selective attention better than others, based on the information provided by the experiment. The updated beliefs now correspond to the appropriate prior expectation that the optimal-attention GCM_{opt} model has about people's attention.

This approach makes the prior for the attention parameters not only dependent on the category structure to be learned, but also on the number of trials, the nature of feedback, and other relevant properties of the experiment. The key idea is that the theoretical assumption of optimality proposes people's final category learning behavior is consistent with them having made optimal inferences about attention weights over the course of the experiment. In this sense, the prior over the attention weights in GCM_{opt} corresponds to the optimal distribution of a maximally perfect performer given limited information.

A Demonstration of the Optimal-Attention GCM

In this section, we present a worked example of how GCM_{opt} includes the theoretical assumption of optimal attention weights, and contrast it with the traditional GCM formulation in a formal model comparison exercise. Our demonstration and test relies on four previously published data sets, collected and reported by Nosofsky (1989).

Experimental Data

Stimuli. Nosofsky's (1989) experiment involved 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. A two-dimensional representation of the stimulus space was used, based on similarity data derived from a separate identification experiment.

Category Structures. Four different category structures were defined using the 16 stimuli, as shown in Figure 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 unassigned. In two of the category structures, called 'angle' and 'size', attending to a single dimension is sufficient to learn the categories. The other 'criss-cross' and 'diagonal' structures require attention to both dimensions to be learned.

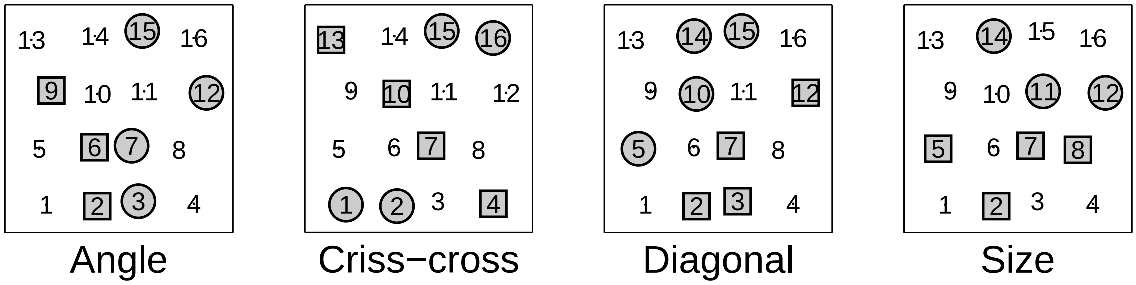

Figure 1. The four category structures used by Nosofsky (1989). Each numbered location corresponds to a stimulus, with the stimuli assigned to the two categories shown by squares (category A) and circles (category B) for each structure. Based on Nosofsky (1989, Figure 1).

Procedure. Each category structure was learned by a separate group of participants under a standard training-test procedure. During the training phase, each stimulus was presented with equal frequency and received deterministic feedback. During the test phase, each unassigned stimulus was presented twice and each assigned stimulus was presented approximately six times. Nosofsky (1989) restricted the modeling analyses to those participants who achieved a sufficiently high accuracy in classifying the assigned stimuli.

Optimal-Attention Prior Distributions

Because the stimuli are two-dimensional, there is a single free attention weight parameter w. To find the optimal attention distributions over this parameter for each category structure, we followed previous authors (e.g., Nosofsky, 1984, p. 110, Rehder & Hoffman, 2005, p. 812), and made the assumption that the generalization parameter c was equal to 3.[1] The optimal distributions correspond to the posterior distribution of the attention parameter, starting from a uniform prior, for a learner who has as data all of the presented stimuli and their correct category assignments. That is, they are the result of the learner applying Bayes theorem to the attention parameter, based on the data provided by the experiment, using the GCM as their likelihood function. In this sense, the optimality assumption can be interpreted as a form of 'rational' analysis, as has recently been employed with great success in many psychological models (see Chater, Tenenbaum, & Yuille, 2006; Griffiths, Kemp, & Tenenbaum, 2008, for overviews).

Figure 2 shows that, for both category structures for which only a single dimen-

[1] Because the optimal distribution of weights depend on the value of c, there is no unique optimal distribution. However, over a large range of psychologically realistic values of the generalization parameter, the optimal distributions are qualitatively very similar to those shown in Figure 2, as are the quantitative model comparison results reported later.

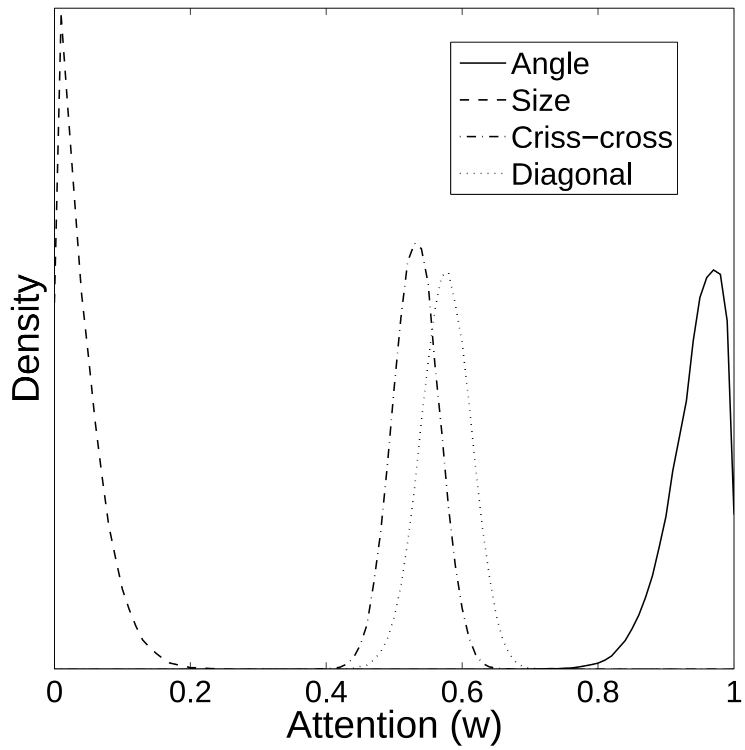

Figure 2. Distributions over the attention parameter w, corresponding to the optimal allocation of attention for each of the four category structures used by Nosofsky (1989).

sion is relevant for performing the classification (i.e., the angle and size structures), the variance of the attention distribution is small, and is centered very close to 0 or 1. In both category structures where both dimensions are relevant (i.e., the criss-cross and diagonal structures), the distribution of attention is broader, and is centered around much less extreme values, indicating that both dimensions are relevant for correct classification.

Testing the Optimal-Attention Assumption

Because GCM_{opt} incorporates the theoretical assumption of optimal attention within the model itself, it is natural to test this assumption by comparing its performance to that of the traditional GCM. We now undertake such a comparison, using a standard Bayesian model selection approach (e.g., Pitt et al., 2002).

Priors for Model Parameters. Bayesian model selection requires that prior distributions be specified for all of the parameters in the models being evaluated. Neither GCM nor GCM_{opt} makes any theoretical claims about which values are likely or unlikely for the c parameter, and so the same uninformative prior is required for both models. Since the c parameter functions as an inverse scale (i.e., $1/c$ scales the distances), implying c^2 functions as a precision, the standard near non-informative prior $c^2 \sim \text{Gamma}\left(\varepsilon, \varepsilon\right)$ with $\varepsilon = .001$ was used (e.g., Jaynes, 2003, Ch. 12; Lunn, Thomas, Best, & Spiegelhalter, 2000; but see also Gelman, 2006).

The traditional GCM model also makes essentially no formal committment to the attention weights parameter w, and so requires a standard uninformative prior. We used the obvious prior $w \mid \text{GCM} \sim \text{Uniform}\,(0, 1)$. For the w parameter in the GCM_{opt} model, we used the theory-informed prior distributions shown for each category structure shown in Figure 2. This, of course, is the only difference between the models, and the key assumption being tested by the formal model comparison.

Model Selection Measures. We based model selection on the marginal likelihoods of the two models. Each marginal likelihood integrates over the prior distribution of parameters, so that, for example, $p\,(\boldsymbol{y} \mid \text{GCM}) = \int p\,(\boldsymbol{y} \mid c, w, \text{GCM})\, p\,(c, w \mid \text{GCM})\ \ \text{d}\,(c, w)$, where $\boldsymbol{y} = (k_1, \ldots, k_N)$ represents the observed data. We assumed for both models independence in the prior distribution, so that $p\,(c, w) = p\,(w)\, p\,(c)$.

Marginal likelihoods have the well-known and important property of automatically accounting for model complexity by finding the *average* fit of each model to the data across the entire prior distribution of parameters. Crucially, for our application, marginal likelihoods—unlike simple approximations like the Akaike Information Criterion or Bayesian Information Criterion that rely on maximum-likelihood fits (Pitt et al., 2002)—are also sensitive to the prior distributions on model parameters. Thus, if people use optimal attention weights consistent with the prior assumptions of the GCM_{opt}, the marginal likelihood will reveal a better average fit. On the other hand, if people make classification decisions consistent with non-optimal attention allocation, marginal likelihood will prefer the traditional GCM.

In terms of interpretation, it is also convenient that marginal likelihoods together define the well-known Bayes Factor (Kass & Raftery, 1995). Formally, the Bayes Factor for the category learning data $\boldsymbol{y}$ is given by the ratio of marginal likelihoods $\text{BF} = p\,(\boldsymbol{y} \mid \text{GCM})\,/p\,(\boldsymbol{y} \mid \text{GCM}_{\text{opt}})$.

Model Selection Results. We evaluated each marginal likelihood using a standard numerical method based on grid sampling, based on 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 or 0.01 for c. This resolution proved high enough that further refinement did not change the results significantly.

The calculated marginal likelihoods, expressed on the standard logarithmic scale, are presented in Table 1. The difference between these log marginal likelihoods for the two models is the log of the Bayes Factor that compares them. Thus, for the angle and size category structures, the Bayes Factor shows that GCM_{opt} is, respectively, $\exp\,(76.4 - 74.2) \approx 9.0$ and $\exp\,(51.9 - 50.6) \approx 3.7$ times more likely than the traditional GCM, based on the evidence provided by the data. For the criss-cross and diagonal structures, in contrast, the Bayes Factor shows that the traditional

Table 1: Marginal likelihoods, expressed as minus log likelihoods, for the traditional GCM and optimal-attention GCM_{opt}, over each of the four category learning tasks considered by Nosofsky (1989).

	GCM	GCM_{opt}
Angle	76.4	74.2
Criss-cross	68.3	75.2
Diagonal	65.0	77.8
Size	51.9	50.6

GCM is, respectively, $\exp(75.2 - 68.3) \approx 990$ and $\exp(77.8 - 65.0) > 10,000$ times more likely than GCM_{opt}.

Overall, therefore, in the two category structures where a single dimension is relevant, there is more evidence for GCM_{opt} than for the traditional GCM. This can be interpreted as support for the attention optimization hypothesis. In the two remaining category structures, there is overwhelming evidence for the traditional GCM. It seems that in these cases, participants did not distribute their attention optimally over both of the stimulus dimensions. These conclusions are completely consistent with those of Nosofsky (1989), made on the basis of post-hoc likelihood ratio tests and chi-squared model comparisons of the point parameter estimates found by fitting only the traditional GCM.

Discussion

Our demonstration of how an optimal-attention GCM can be developed as a model, applied to Nosofsky's (1989) data, and compared to the traditional GCM, yielded the same basic findings as the original investigations. It is reasonable to ask, therefore, what has been gained. The basic answer is that one of main rationales for building formal models in psychology is to test theoretical assumptions. The ability of the GCM to describe, interpret and predict category learning data speaks to the reasonableness of theoretical assumptions like those about exemplar representation, similarity-based categorization, and the Luce choice decision rule.

It seems strange, therefore, to use a different approach to test an assumption like the optimal distribution of attention. Rather than testing a model that formalizes the assumption, most previous analyses rely on a model that explicitly did not try to capture the relevant theory, and then required post-hoc assessment of parameter estimates to reach conclusions. What our demonstration shows is that it is straightforward—within a coherent Bayesian statistical framework for implementing and evaluating models—to include theoretical assumptions about parameters within prior distributions. The fact that the same conclusions are reached in our demonstration is a comforting confirmation of the original tests of the optimality assumption.

But we believe our comparison of models embodying competing theoretical ideas provides a more direct and principled method for evaluating whether the data support the theory.

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. 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 method we used to formalize the distribution corresponding to optimal attention is surely not the only possibility, but it is a formal and testable implementation consistent with the basic theoretical motivation.

More generally, we think our demonstration of the ability of prior distributions in the Bayesian setting to carry theoretical information underscores an under-appreciated merit of the Bayesian approach. Standard statistical frameworks do not make it easy to theorize about more or less likely combinations of parameter values in models, but this seems an important avenue for expressing theoretical ideas. There are many ways theorists could formalize their ideas within the Bayesian approach. One simple idea is to place order constraints on parameters to capture expectations such as the relative difficulty of different tasks (e.g., Hoijtink, Klugkist, & Boelen, 2008), or to constrain the range of behavior a model can produce to those consistent with an experimental setting (e.g., Liu & Aitkin, 2008). More advanced methods involve maximum entropy priors, where complicated constraints on variables can be embedded in prior distributions (e.g. 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). We think these sorts of ideas should be pursued much more often than they are now, and that, as they mature, theorists will be able to develop and evaluate formal models that more closely capture and test our understanding of psychological processes.

Acknowledgments

MDL acknowledges the support of a Visiting Fellowship from the University of Leuven.

References

Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences*, *10*(7), 287–291.

Garner, W. R. (1974). *The Processing of Information and Structure*. Potomac, MD: Erlbaum.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, *1*(3), 515–534.

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

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.

Lee, M. D. (2008). Three case studies in the Bayesian analysis of cognitive models. *Psychonomic Bulletin & Review*, *15*(1), 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*(8), 1403–1424.

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

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS – a Bayesian modelling framework: Concepts, structure and extensibility. *Statistics and Computing*, *10*, 325–337.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(1), 104–114.

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

Nosofsky, R. M. (1989). Further tests of an exemplar-similarity approach to relating identificationand 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.* 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*(1), 322–339.

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

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 and Review*, *12*, 195–223.

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknowndistance function. I. *Psychometrika*, *27*(2), 125–140.

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science*, *214*(24), 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*(8), 1248–1284.

Vandekerckhove, J., Tuerlinckx, F., & Lee, M. D. (2008). A Bayesian approach to diffusion process models of decision-making. In V. Sloutsky, B. Love, & K. McRae (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 1429–1434). Austin, TX: Cognitive Science Society.

Wetzels, R., Vandekerckhove, J., Tuerlinckx, F., & Wagenmakers, E. (in press). Bayesian parameter estimation in the Expectancy Valence model of the Iowa gambling task. *Journal of Mathematical Psychology.*