Cognitive Science 32 (2008) 1248–1284
Copyright © 2008 Cognitive Science Society, Inc. All rights reserved.
ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1080/03640210802414826

A Survey of Model Evaluation Approaches With a Tutorial on Hierarchical Bayesian Methods

Richard M. Shiffrin[a], Michael D. Lee[b], Woojae Kim[a], Eric-Jan Wagenmakers[c]

[a]*Departments of Psychology & Cognitive Science, Indiana University*
[b]*Department of Cognitive Sciences, University of California, Irvine*
[c]*Department of Psychology, University of Amsterdam*

Received 25 July 2007; received in revised form 14 July 2008; accepted 18 August 2008

Abstract

This article reviews current methods for evaluating models in the cognitive sciences, including theoretically based approaches, such as Bayes factors and minimum description length measures; simulation approaches, including model mimicry evaluations; and practical approaches, such as validation and generalization measures. This article argues that, although often useful in specific settings, most of these approaches are limited in their ability to give a general assessment of models. This article argues that hierarchical methods, generally, and hierarchical Bayesian methods, specifically, can provide a more thorough evaluation of models in the cognitive sciences. This article presents two worked examples of hierarchical Bayesian analyses to demonstrate how the approach addresses key questions of descriptive adequacy, parameter interference, prediction, and generalization in principled and coherent ways.

Keywords: Model selection; Model evaluation; Bayesian model selection; Minimum description length; Prequential analysis; Model mimicry; Hierarchical Bayesian modeling

1. Introduction

Models play a central role in cognitive science. They provide the formal bridge between theories and empirical evidence. They make explicit ideas about how cognitive processes operate and how psychological variables guide those processes. They allow theoretical assumptions to be tested in the laboratory, and make predictions about how cognition will behave in new or different circumstances. The central role models play makes their evaluation an important issue. It is necessary to be able to choose between models and decide whether a model is

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

"good." There is, however, no simple or unitary answer to the question of what makes a model good. Our view is that good models should help in achieving at least the following five related, but different, goals:

1. *Achieve a basic level of descriptive adequacy:* A model should agree with observed data well enough that something sensible can be said about how a cognitive process behaves. For example, a model of memory retention that shows a negatively accelerating decrease in retention over time describes some basic aspects of the data. The model serves to give a formal expression of important empirical regularities.
2. *Provide insight and understanding:* A model should help us understand things not directly evident from looking at the data, thereby leading to further studies and tests. It should allow us to deepen, refine, and elaborate our understanding of the cognitive processes at work. For example, a category learning model may account for data only when using a particular value of a selective attention parameter. The value of this parameter has psychological meaning and provides part of an explanation of how category learning was achieved.
3. *Facilitate prediction and generalization:* A good model should help make predictions about what will be observed in the future or generalizations about what would be observed under altered circumstances.
4. *Direct new empirical explorations:* A good model should lead us to develop new empirical studies that have the greatest chance of increasing our understanding and adding to our knowledge.
5. *Foster theoretical progress:* There is a sense in which the goal of modeling is not to find answers but to sharpen questions. Modeling forces theoretical ideas to take precise forms and to encounter empirical evidence head-on. Models help make clear the predictions of theories and suggest critical tests. To the extent that a model clarifies where theory is working and where it is failing, it makes a valuable contribution.

In this article, we begin by reviewing current methods for evaluating models in the cognitive sciences, many of which have been covered in recent special issues and review articles (e.g., I. J. Myung, Forster, & Browne, 2000; Pitt, Myung, & Zhang, 2002; Wagenmakers & Waldorp, 2006). These include theoretically based approaches, such as Bayes factors and minimum description length (MDL) measures; simulation approaches, including model mimicry evaluations; and practical approaches, such as validation and generalization measures. We point out that, although often useful in specific settings, most of these approaches are limited in their ability to give a general assessment of a model. Many of the measures focus on only one narrow aspect of what makes a model good, and often provide too concise a summary to guide further model development. We suggest that hierarchical methods—which are rapidly becoming the approach of choice in scientific and statistical modeling—can provide a more thorough evaluation of models in the cognitive sciences. Such methods are easily implemented in a Bayesian framework, and in light of the general advantages of Bayesian inference, we restrict our attention to hierarchical Bayesian methods. We present two worked examples to demonstrate how the hierarchical Bayesian approach addresses questions of descriptive

adequacy, parameter interference, and prediction and generalization in principled and coherent ways.

2. Current approaches to model evaluation

Throughout the cognitive science literature, even with a restricted focus on quantitative and statistical goals, a large number of approaches to model selection and evaluation have been used. It is useful to classify these roughly into three classes: Theoretical approaches develop formal measures designed to assess models on the basis of some well-justified criterion, simulation approaches use computational methods to explore the relationship between models and data, and applied approaches evaluate the ability of models to predict new or different data.

Each of these approaches potentially offers different and useful information in evaluating and comparing models. No two approaches, whether within or between these classes, always agree with each other, but many pairs agree in substantial numbers of cases. Theoretical approaches have the potential to offer deep and general insights, but are not guaranteed to give better answers than the other approaches, especially in light of the many differing goals of model evaluation. In addition, it is not always feasible to apply the best theoretical approaches, especially for models so complex that it is hard to understand their full range of behavior. Simulation approaches offer some insight into how and why models perform differently, and usually scale better to the more complex models. Applied approaches are almost always possible, but will not necessarily provide any insight into what underlies the observed success or failure of a model to predict unseen data. Yet applied approaches map well onto the goal of predicting new data that many researchers have in mind for selecting models.

2.1. Theoretical approaches

One way of understanding the range of theoretical measures available to cognitive science is to think in terms of three underlying philosophies (see Grünwald, 2005, p. 13). The Bayesian philosophy champions the model or model class that is most likely, in the sense of providing the most robust fit to the data (i.e., fits well at the most parameterizations), given a prior state of knowledge. The MDL philosophy champions the model that best compresses the data, in the sense that the length of the description of the model and the data as encoded by the model is minimal. The prequential philosophy champions the model that best predicts unseen data.

These philosophies are not necessarily in direct competition, and there are many consistencies, both in terms of conceptual motivation and technical results, in their relationships to each other. Often, a model that compresses the data the best is the one that fits the data most robustly, and also predicts unseen data better. Nevertheless, the three philosophies give theoretical primacy to different aspects of what makes a model good. It should be expected that the three philosophies will give rise to measures that reach different conclusions for specific

model evaluation problems. Such differences should be interpreted not in terms of flaws of the measures, but in terms of the differing goals: robustness, compression, and prediction.

2.1.1. Bayesian methods

Bayesian inference has a clear and compelling foundation in probability theory (Jaynes, 2003; Lee & Wagenmakers, 2005). What makes a statistical inference Bayesian is the way uncertainty is handled. The Bayesian assumption is that uncertainty is always represented by probability distributions. This allows probability theory, in the form of Bayes Rule, to provide an automatic method of inferential updating when useful information, such as empirical observation, becomes available.

The use of probability theory also allows for "marginalization" in which what is known about one variable can be conditioned on what is known about every other relevant variable. Marginalization is the key component of Bayesian Model Selection because it allows the robustness of the fit between a model and data to be measured, and embodies an automatic form of Ockham's razor.

As a simple example, consider the problem of testing whether a coin is fair or biased. The fair model, M_f, asserts that the rate of observing heads is $\theta = 1/2$. The biased model, M_b, asserts that the rate could be $0 < \theta < 1$. If we assume that both heads and tails are observable outcomes, there is a justification for making the prior assumption that each possible rate θ is equally likely, therefore using a uniform prior (see Lee & Wagenmakers, 2005). For observed data D giving k heads out of n tosses, the likelihood functions are as follows:

$$\begin{aligned} p(D \mid M_f) &= \binom{n}{k}\left(\frac{1}{2}\right)^k\left(\frac{1}{2}\right)^{n-k} \\ &= \binom{n}{k}\left(\frac{1}{2}\right)^n, \end{aligned} \tag{1}$$

for the fair model and

$$p\left(D \mid M_b, \theta\right) = \binom{n}{k}\theta^k\left(1-\theta\right)^{n-k}, \tag{2}$$

for the biased model.

The fair model has no parameters, and so its likelihood provides a complete description of how the model relates to the data. For the biased model, the Bayesian approach averages the likelihood over all possible values of the parameter θ, as weighted by the prior. This average is the marginal likelihood $p\left(D \mid M_b\right)$, and provides a measure of how robustly the model—in *all* of its possible forms—fits the data. Formally,

$$\begin{aligned} p(D \mid M_b) &= \int p(D \mid M_b, \theta)p(\theta)\,\mathrm{d}\theta \\ &= \int_0^1 \binom{n}{k}\theta^k(1-\theta)^{n-k}\,\mathrm{d}\theta \\ &= \frac{1}{n+1}. \end{aligned} \tag{3}$$

The ratio of marginal likelihoods for the two models is known as the Bayes factor (Kass & Raftery, 1995), and can also be thought of as the ratio of posterior to prior odds. In this sense, the Bayes factor measures the evidence data provided for one model over another. The Bayes factor is widely used for model selection, and is optimal under the assumption of 0–1 loss (i.e., that one model is the true data-generating model, and the other model is false). The Bayesian framework has the potential to define alternative model selection measures, corresponding to different loss assumptions, that deserve wider exposure and exploration for assessing models of cognition (see Gelfand, 1996, p. 148).

For our coin example, the Bayes factor is:

$$\frac{p\left(D \mid M_f\right)}{p\left(D \mid M_b\right)} = \frac{\binom{n}{k}\left(\frac{1}{2}\right)^n}{\frac{1}{n+1}}, \tag{4}$$

and therefore, for data giving $k = 12$ heads out of $n = 20$ tosses, gives approximately the value 2.52. This means that it is about $2^1/_2$ times more likely the coin is fair, rather than biased, based on the evidence provided by the data. As this example makes clear, a feature of Bayesian approaches to model selection is that its measures have naturally meaningful interpretations because they are probability statements and do not require a separate calibration. It is arguable that none of the other model selection approaches we discuss have this highly desirable property.

For most cognitive models, however, it is difficult or impossible to produce an analytic expression for the marginal probabilities, so some sort of simulation or approximation must be undertaken. Popular Bayesian measures such as the Bayesian (or Schwarz) Information Criterion (BIC; Schwarz, 1978), the Laplace Approximation (Kass & Raftery, 1995), and the Geometric Complexity Criterion (I. J. Myung, Balasubramanian, & Pitt, 2000) are asymptotic analytic approximations to the marginal probabilities. The differences between them derive from how close an approximation they are to the exact marginal probability and, in particular, whether they are good enough approximations to be sensitive to the part of model complexity that arises from the functional form of the interactions between parameters (I. J. Myung & Pitt, 1997).

Computational methods have recently emerged in cognitive science and in computational Bayesian statistics (Courville, Daw, & Touretzky, 2006; Lee, 2008) as an alternative approach to approximating the integrated probability across the parameterized class. There exist various computational algorithms, most often developed within the Markov chain Monte Carlo (MCMC) framework, that are based on drawing samples from the joint posterior distribution of the parameters of a model. Approximating the desired probability of the model class from posterior samples is often difficult, but useful techniques are being developed along a number of different lines (e.g., Carlin & Chib, 1995; Chib, 1995; Raftery, Newton, Satagopan, & Krivitsky, 2007). A particularly useful approach may be that termed Reversible-Jump MCMC (Green, 1995). We provide a concrete example of obtaining the Bayes factor using MCMC methods later, when we discuss hierarchical Bayesian approaches.

2.1.2. MDL methods

The MDL approach (see Grünwald, 2007, for a thorough and recent treatment) has its foundation in information and coding theory, and particularly in the theory of Kolmogorov or algorithmic complexity. The basic idea is to view models as codes that express expected regularities, and prefer those codes or models that best compress the observed data. The key measure for model evaluation under the MDL approach is the code length, or stochastic complexity, of the data under the model. The minimal code length used is the combined code length for the model and data described by the model.

Various MDL measures have been developed as approximations to this code length. Initial two-part code approximations (Rissanen, 1978) were later extended to the Stochastic Complexity Criterion (SCC; Rissanen, 1987, 1996), also sometimes described as the Fisher Information Approximation. The most recent MDL measure is the normalized maximum likelihood (NML; Rissanen, 2001), which is based in a reconceptualization of the original stochastic complexity measure. NML measures how well a model fits observed data, relative to how well that model could fit any possible data, and has found application to cognitive models (I. J. Myung, Navarro, & Pitt, 2006). Formally, the NML is given by the following:

$$\mathrm{NML} = \frac{p\left(D \mid \theta^*(D)\right)}{\sum_{D'} p\left(D' \mid \theta^*(D')\right)}, \tag{5}$$

where D denotes the observed data, D' denotes any possible data, and $\theta^*(\cdot)$ denotes the maximum likelihood parameter values for a given set of data.

Loosely, it is reasonable to think of two-part code MDL measures as being like the BIC approximation, which equates the complexity of a model with a count of parameters; the SCC with the Bayesian Laplacian-style approximations, which are also sensitive to functional form complexity; and the NML with the Bayesian marginal probability, which are exact.

To make the NML idea concrete, we apply it to our earlier coin example. The possible observations involve $k' = 0, \ldots, n$ heads out of the 20 tosses. Thus, for the fair model,

$$\mathrm{NML}_{M_f} = \frac{\binom{n}{k}\left(\frac{1}{2}\right)^n}{\sum_{k'=0}^{n}\binom{n}{k'}\left(\frac{1}{2}\right)^n}. \tag{6}$$

For the biased model, the best fitting (i.e., maximum likelihood) value of the parameter is $\theta^*(D) = k/n$ for observed data with k heads out of n tosses; therefore,

$$\mathrm{NML}_{M_b} = \frac{\binom{n}{k}\left(\frac{k}{n}\right)^k\left(1-\frac{k}{n}\right)^{n-k}}{\sum_{k'=0}^{n}\binom{n}{k'}\left(\frac{k'}{n}\right)^{k'}\left(1-\frac{k'}{n}\right)^{n-k'}}. \tag{7}$$

Given the observed data $k = 12$ heads out of $n = 20$ tosses we considered earlier, the NML values are approximately 0.12 for the fair model and 0.04 for the biased model; therefore, the fair model is preferred, as in the Bayesian analysis. It is interesting to note that although the MDL and Bayesian approaches can give similar answers, their theoretical bases are quite different. For example, the Bayesian approach relies on the likelihood principle: the idea that only what has been observed matters, not what could have been observed. The Bayesian approach, therefore, calculates the probability of the observed data for every parameterized model in each model class in contention, and then normalizes so the probabilities add to one.

In contrast, the normalized maximum likelihood approximation to MDL considers only the model in each parameterized class that has the highest probability of the observed data, and for each model class normalizes by the sum of these maximum probabilities for every data set that could have been observed. These approaches appear almost polar opposites, and deep and non-obvious analysis is needed to understand why they often produce similar results.

2.1.3. Prequential methods

The prequential method (from "sequential prediction"; Dawid, 1984, 1991, 1992; Dawid & Vovk, 1999; Skouras & Dawid, 1998) is based on the philosophy that the relevant criterion for model selection is the minimization of prediction error for unseen data. An ideal model—one that captures only replicable structure and ignores all idiosyncratic noise—has the smallest prediction errors for future data coming from the same source. The prequential method estimates the predictive power of a model by using one part of the data to estimate the parameters the model, and using another part of the data to assess the predictions of the model. The model with the best predictive performance is preferred. The distinguishing features of the prequential method are that the size of the data set used for estimation continually grows, and that the method concerns itself only with one-step-ahead predictions.

For instance, when you want to know which of two weather forecasting systems is more accurate, the prequential method prescribes that you consider only the forecasts for the next day. As the days go by and more information becomes available, the forecasting systems are free to continually update their predictions; their predictive performance, however, is assessed only for the next-day weather that has not yet been observed. At each point in time, the relative merit of the weather forecasting systems is given by the difference in the sum of their prediction errors that have been observed so far (i.e., the Accumulated one–step–ahead Prediction Errors or APE).

More specifically, assume that we have a data set $x^n = (x_1, x_2, \ldots, x_n)$, and a model M_j for which one wants a prequential performance estimate. The calculation then proceeds as follows (Wagenmakers, Grünwald, & Steyvers, 2006):

1. Based on the first $i - 1$ observations, calculate a prediction $\hat{p}_i$ for the next observation i.
2. Calculate the prediction error for observation i (e.g., $(x_i - \hat{p}_i)^2$).
3. Increase i by 1 and repeat Steps 1 and 2 until $i = n$.
4. Sum all of the one-step-ahead prediction errors calculated in Step 2. This yields the accumulative prediction error (APE). Thus, for model M_j, the accumulative prediction error is given by

$$\mathrm{APE}(M_j) = \sum_{i}^{n} d[x_i, (\hat{p}_i \mid x^{i-1})], \tag{8}$$

where d indicates the specific function that quantifies the discrepancy between what is observed and what is predicted.

In our coin example, suppose that the successive predictions for n coin tosses $x^n = (x_1, \ldots, x_n)$ are based on the logarithmic loss function $-\ln \hat{p}_i(x_i)$, so that the larger the

probability that $\hat{p}_i$ (determined based on the previous observations x^{i-1}) assigns to the observed outcome x_i, the smaller the loss. As in the Bayesian analysis, we assume a uniform prior distribution. This distribution is used to provide a prediction for the first datum. Under these conditions, the prequential method will always prefer the same model as the Bayes factor. To see why this is the case, note that from the definition of conditional probability, $p(x_i \mid x^{i-1}) = p(x^i)/p(x^{i-1})$, it follows that

$$p(x_1, \ldots, x_n) = p(x_n \mid x^{n-1})p(x_{n-1} \mid x^{n-2}) \ldots p(x_2 \mid x_1)p(x_1). \tag{9}$$

This equation shows that the probability of the data may be decomposed as a series of sequential probabilistic predictions $p(x_i \mid x^{i-1})$. The APE with logarithmic loss and the Bayesian predictions satisfy

$$-\ln p(x^n \mid M_j) = \sum_{i=1}^{n} -\ln p(x_i \mid x^{i-1}, M_j). \tag{10}$$

The Bayes factor prefers the model M_j that minimizes the left-hand side, whereas the prequential method prefers the model that minimizes the right-hand side; hence, the two procedures are equivalent (for details, see Wagenmakers et al., 2006, pp. 152–153).

This procedure may appear reasonable, but so do many others. For instance, why not use "two-step-ahead prediction error," or why not weight the most recent prediction errors more than the older prediction errors? The reason is that the prequential method as previously formulated has strong theoretical ties to both Bayesian methods and the predictive version of the MDL principle (Rissanen, 1986b). In certain situations (i.e., logarithmic loss and Bayesian one-step-ahead predictions), the prequential method and Bayesian model selection are exactly equivalent, and in other situations (e.g., squared error loss and maximum likelihood predictions) the methods converge as many data are collected (for details, see Wagenmakers et al., 2006). This theoretical connection provides both a theoretical foundation for the prequential method and a predictive interpretation of the Bayesian and MDL methods.

Although the prequential method has seen little application in psychological research, its advantages are readily apparent. The prequential method is a data-driven procedure that at least partly approximates Bayesian model selection, and yet it does not require the specification of priors. It lends itself readily to simulation methods, and model complexity is taken into account easily and automatically through the focus on predictive performance. The only requirement for the prequential procedure to work is that the models under consideration are able to generate predictions for the next observation. This means that the general method is very flexible and can, for instance, also be applied to complex models or architectures of cognition that may not even have a likelihood function. Finally, prequential model selection is *consistent* in the sense that when the set of candidate models contains the true data-generating model, the prequential model will start to prefer it over the other models as the number of observations increases. In our opinion, the prequential method and its variants are under-appreciated and deserve more study.

The prequential method, promising as it is, also has its limitations (Wagenmakers et al., 2006, pp. 154–155). One problem is that it is not immediately clear how the method should be applied in case the data do not have a natural ordering. In weather forecasting, data

arrive sequentially, and there can be no discussion about what observation to predict next. In most psychological experiments, however, the data may arrive sequentially, but this is often considered accidental. When the sequential predictions are not made by a Bayesian system, the ordering of the data can lead to different results, at least for small data sets. One solution to this problem calculates the final APE as an average of APEs for many random orderings of the same data set (Kontkanen, Myllymäki, & Tirri, 2001; Rissanen, 1986a).

2.2. Simulation approaches

In many psychological applications, only a handful of candidate models carry substantive interest. Often, the investigation centers on which one of two competing models is to be preferred. In such cases, an important question concerns the extent to which the models are able to mimic each other's behavior (Navarro, Pitt, & Myung, 2004; Wagenmakers, Ratcliff, Gomez, & Iverson, 2004).

Suppose, for instance, that a researcher sets out to determine what model of response time is best: the diffusion model (e.g., Ratcliff, 1978) or the mutual inhibition model (e.g., Usher & McClelland, 2001). Both models may differ in the number of parameters, and the parameters of the two models certainly differ in functional form. It may transpire that for a particular data set, the diffusion model provides a better fit to the data. To what extent is this diagnostic? Perhaps the diffusion model is a chameleon-like model that is able to fit well not just its own data, but also data generated by the mutual inhibition model. The mimicry method involves generating data sets with each of the two model classes and fitting each generated data set with both models. At the end, one has two distributions of fit differences—one when Model A generated the data sets and one when Model B generated the data sets. The potential ability to select the actual generating model, with choice of a suitable choice criterion, is determined by the separation of the two distributions. The mimicry characteristics are determined by the placement of each distribution relative to the zero difference of fit.

To be more specific, the following steps are used to produce the desired distributions (Navarro et al., 2004; Wagenmakers et al., 2004):

1. Generate data from Model A.
2. Fit data from Step 1 with Model A and Model B, and calculate the difference in goodness-of-fit ΔGOF_A.
3. Generate data from model B.
4. Fit data from Step 3 with Model A and Model B, and calculate the difference in goodness-of-fit ΔGOF_B.
5. Repeat Steps 1 through 4 n times to yield $\Delta GOF_A[1, 2, \ldots, n]$ and $\Delta GOF_B[1, 2, \ldots, n]$.

An important question involves the best way to generate data from the model classes because one would want to generate data sets with different parameter choices and perhaps with differences across subjects. One local or *data-informed* method generates data from the posterior distribution for the parameters. This method is computationally straightforward, but from a model selection perspective, it often assigns too much weight to the complex model (cf. Wagenmakers et al., 2004, Fig. 8). The global or *data-uninformed* method generates data from the prior distribution for the parameters (Navarro et al., 2004). The choice of parameters for

generating data sets is critical to the method, and involves choices of the range and distributions of parameters within that range for each model class. How to choose these in a way that is fair to the model classes under comparison is an open question, analogous to questions about the way in Bayesian methods to choose priors that are fair to the model classes under comparison. For an example of the use of these methods in practice, applied to model selection for models in different classes for different tasks, see Cohen, Sanborn, and Shiffrin (2008). Whatever method is used to generate data sets, one can find the criterion (on the difference-of-fit axis) that maximizes the choice of the actual generating model. When the data sets are generated in a way that does match the actual environment, then classifying them with this optimal criterion will optimize the total number of correct classifications.

In the coin example, model M_f assumes a fair coin, $\theta = \frac{1}{2}$, whereas model M_b assumes a biased coin, $0 < \theta < 1$. Model M_b will therefore always be able to mimic model M_f perfectly, whereas model M_f will only be able to mimic M_b for values of θ that are close to $\frac{1}{2}$. To illustrate the model mimicry method, we applied the data-uninformed parametric bootstrap cross-fitting method and generated 1,000,000 simulated data sets from both M_f and M_b. Under model M_b, each data set was generated by first sampling a particular θ from a uniform distribution. Note that, just as the observed data, each simulated data set contained exactly 20 observations.

The two models were fitted to the 2,000,000 simulated data sets, and the difference in goodness of fit was quantified by the differences in the log probability that M_f and M_b assign to the data; that is,

$$\log\left(\frac{\Pr(k \mid n = 20, M_b)}{\Pr(k \mid n = 20, M_f)}\right).$$

For instance, for the observed data (i.e., $k = 12$) we calculate

$$\Pr(k = 12 \mid n = 20, M_f) = \binom{n}{k}\frac{1}{2}^n \approx 0.12$$

and

$$\Pr(k = 12 \mid n = 20, M_b) = \binom{n}{k}\hat{\theta}^k(1-\hat{\theta})^{n-k} \approx 0.18,$$

where $\hat{\theta} = k/n$ is the maximum likelihood estimate. The difference between 0.12 and 0.18 is $\log(\frac{0.18}{0.12}) \approx 0.403$.

An examination of the discrete distributions of log differences revealed that the observed difference of approximately 0.403 is more likely to occur under M_f than it is under M_b. Specifically, the difference of approximately 0.403 is observed with a probability of 0.24 under M_f and 0.10 under M_b. It is tempting to conclude that the data are 2.4 times more likely under M_f than they are under M_b. Note the quantitative correspondence to the Bayes factor conclusion that the data are about $2^1/_2$ times more likely under the fair coin model than under the biased coin model.

It is not always the case, however, that the Bayes factor analysis and the model mimicry analysis produce almost exactly the same answer. For instance, J. I. Myung, Pitt, and Navarro (2007) showed that adding parameters may substantially increase a model's complexity (as indicated by MDL or Bayes factor methods) but does not necessarily increase a model's

ability to mimic a competing model. The advantages of the model mimicry method are clear: It is easy to apply to complex models, and it yields an intuitive measure for the capacity of models to mimic each other. It also can be used to optimize the goal of the selection of the actual generating model, although it could be argued that this goal does not penalize complexity sufficiently. Also, the method can be used to determine how many experimental trials are necessary to distinguish two models. The model mimicry method is still relatively unexplored. Future work will have to study more carefully the extent to which the mimicry method is related to other model selection procedures.

2.3. Practical validation approaches

The philosophy that underlies validation methods is the same as the one that underlies the prequential method: The preferred model class is the one whose (weighted) parameterized models best predict unseen data from the same source. In usual approaches, the best parameterized model in the class is used for prediction, but it is also possible to predict by weighting the predictions of all parameterized models in the class. In these approaches, the models are fitted to one part of the data—the "calibration" or training set—and their predictive performance is assessed for the remaining part of the data—the "validation" or test set.

Although validation methods divide the observed data in a training set and a test set, there are many ways in which this can be done. This is illustrated by a summary of the most popular methods:[1]

1. *Split-sample or hold-out method:* This method is often used to assess predictive performance of neural networks. In the split-sample method, only one part of the data is ever used for fitting (i.e., the training set and the test set do not change roles), and this leads to results with a relatively high variance.
2. *Split-half cross-validation:* In split-half cross-validation, the first half of the data forms the training set, and the second half of the data forms the test set. After the prediction error for the test set has been assessed, the same procedure is repeated, but now the second half of the data forms the training, set, and the first half of the data forms the test set (i.e., training and test set "cross"). The overall prediction error is the average of the prediction error on the two test sets. Note that each time the model is fitted to only 50% of the data—a procedure that yields relatively large prediction errors.
3. *Leave-one-out cross-validation:* In leave-one-out cross-validation, a data set of n observations is repeatedly split into a training set of size $n - 1$ and a test set of size 1. The overall prediction error is given by the average prediction error for the n test sets (Browne, 2000; Stone, 1974). The computational advantage of this procedure is that it only requires a sequence of n model predictions.
4. *K-fold cross-validation:* In K-fold cross-validation, the data are split in K blocks, and one of those blocks is successively selected to be the test set (i.e., the training set is always $K - 1$ blocks large). The overall prediction error is the average of the prediction error on the K test sets. The problem with this method is that different choices of K may lead to different results.

5. *Delete-d cross-validation:* This method is the same as K-fold cross-validation, except that the test blocks consist of every subset of d observations from the data. As with K-fold cross-validation, different choices of d may lead to different results.
6. *Bootstrap model selection:* The bootstrap method (e.g., Efron & Tibshirani, 1993) is usually applied to obtain standard errors for parameter estimates. The bootstrap procedure works by resampling the observed data (with replacement) in order to use the variability in the observed data as a plug-in estimate for the variability in the population. The bootstrap method can, however, also be applied to smooth the results obtained by cross-validation. In particular, the so-called .632+ bootstrap procedure has been shown to improve on cross-validation in a number of problems (Efron & Tibshirani, 1997). Because the bootstrap resamples are supported by approximately 63.2% of the original sample points, results from the .632+ bootstrap method generally correspond closely to those from split-half cross-validation.

Model selection by validation has a number of clear advantages. Validation methods are data-driven, and replace complicated mathematical analysis by raw computing power (Efron & Gong, 1983). Validation methods are relatively easy to implement, and they can be applied to complicated models without much thought. Despite the intuitive and practical appeal of validation, the many variants of the method show there are open questions about the best implementation. In particular, the balance point between training data and test data remains an open question. Relatively large training sets lead to overfitting, but relatively small training sets lead to underfitting. Further, as the number of observations increases, most cross-validation methods will start to prefer models that are overly complex (i.e., the methods are not consistent; for a discussion, see Shao, 1993; Stone, 1977). These problems of choice and consistency go back to the fact that cross-validation does not make explicit its underlying assumptions. These considerations negate some of the appeal of the validation methods.

We illustrate both leave-one-out cross-validation and split-half validation by revisiting the example of a coin that comes up heads 12 out of 20 tosses. For the leave-one-out method, the biased coin model M_b is fit to training sets of 19 observations, and the maximum likelihood estimate $\hat{\theta}$ is then used to determine the fit to the remaining data point (i.e., the test set). The fair coin model M_f does not learn from the training data, as it always predicts that heads will come up with probability $\theta = \frac{1}{2}$. The difference in goodness of fit between the models for the data from the test set, D_v, is calculated by the ratio of the probabilities that the models assign to the data: $R = \Pr(D_v \mid \hat{\theta}) / Pr(D_v \mid \theta = \frac{1}{2})$. For the leave-one-out method, this procedure only needs to be repeated 20 times. Somewhat surprisingly, the leave-one-out method prefers the biased coin model in 12 out of the 20 cases. The average value of R is 1, indicating no preference for one or the other model.

For the split-half validation method, the data are divided in 10 observations that form the training set and 10 observations that form the test set. This procedure was repeated 10,000 times for random permutations of the data. The assessment of the ratio of probabilities R proceeds in the same way as it did for the leave-one-out model. In 34% of cases, the split-half method preferred the biased coin model H_b, in 42% of cases it preferred the fair coin model H_f, and in 24% of cases their was an exact tie (a tie occurs when the biased coin

model has $\hat{\theta} = \frac{1}{2}$). The average value of R is about 0.80, indicating a preference for the fair coin model. In sum, the results show that leave-one-out cross-validation has a slight preference for the biased coin model, whereas the split-half procedure prefers the fair coin model.

When carrying out simulations, it is possible to explore predictive validation as a model selection criterion in a way that eliminates many of the problems that arise in practice because one knows the "true" model and its generating parameter values. Thus, for a given data set one can estimate parameters for the models in contention, and then determine how well those estimated parameters predict the (infinite) data distributions produced by the true model with its true generating parameters. The model that on the average does the best job of predicting the true distribution of future data is to be preferred. In one sense, this simulation method can be used to compare and contrast different model selection methods. For an example, see Cohen, Sanborn, and Shiffrin (2008).

2.3.1. The generalization criterion method

The goal of the generalization criterion method (Busemeyer & Wang, 2000) is to quantify model adequacy by assessing predictive performance. As in cross-validation, the observed data are divided in two sets: a calibration or training set to estimate the model parameters, and a validation or test set to assess predictive performance. The crucial difference with cross-validation is that in the generalization criterion method, the training set and the test set do not overlap in terms of experimental design. For instance, Ahn, Busemeyer, Wagenmakers, and Stout (2008) compared several models of reinforcement learning by fitting them to one experiment (e.g., the Iowa gambling task; Bechara, Damasio, Damasio, & Anderson, 1994) and evaluating them on a different experiment (e.g., the Soochow gambling task; Chiu et al., 2005).

Thus, in the generalization criterion method, parameters are fit to different conditions than those that are used to evaluate predictive performance. This way, the model comparisons "are based on *a priori* predictions concerning *new* experimental conditions. Essentially, this tests the models ability to accurately interpolate and extrapolate, which is one of the major goals of a general scientific theory" (Busemeyer & Wang, 2000, p. 179). In contrast to cross-validation, the generalization criterion method does not necessarily favor complex models over simple models when the sample size grows large.

3. Worked hierarchical Bayesian examples

All of the approaches to model evaluation we have reviewed have important limitations in their ability to address the basic goals of modeling—achieving descriptive adequacy, enhancing explanation through inference about parameters, making predictions and generalizations, and furthering theoretical development—that we identified at the outset. A theoretical Bayes factor, MDL or predictive measure, or a validation or generalization test result provide useful information about which of a number of competing models has better performance. These measures will usually give some indication of likely parameter values, and give a basis for

inferring which model will predict future data better. However, they do not give a full account of how and why the models succeed and fail to various degrees, and provide little direct information to drive subsequent theorizing.

3.1. Hierarchical methods

We believe hierarchical methods, in general, and hierarchical Bayesian methods, in particular, represent an approach to model development and evaluation in the cognitive sciences that address many of these concerns. Hierarchical Bayesian methods are standard and powerful ways of analyzing models and drawing inferences about parameters from data, and are widely used in statistics, machine learning, and throughout the empirical sciences. The hierarchical Bayesian approach employs the basic machinery of Bayesian statistical inference, with all the advantages it entails (e.g., Jaynes, 2003; Sivia, 1996), but is designed to work with richly structured hierarchical models. Introductions to hierarchical Bayesian methods can be gained from textbook accounts in statistics and machine learning (e.g., Gelman, Carlin, Stern, & Rubin, 2004; MacKay, 2003) or from recent expositions aimed at psychologists (e.g., Griffiths, Kemp, & Tenenbaum, 2008; Lee, 2008). We do not, of course, claim this approach is a final solution or the only sensible approach. However, it has a number of important and useful features, including the ability to check descriptive adequacy, allow inferences about parameters, make predictions and generalizations, compare models, and suggest modeling extensions and refinements that we hope to make clear.

We emphasize that hierarchical models should not be confused with models having a tree structure, such as a neural net with a hidden layer. For present purposes, we may define hierarchical models as models in which some parameters are partly determined (e.g., chosen from distributions defined by) other parameters. The determining parameters are typically termed *hyperparameters*. As a concrete example, consider modeling data from several subjects. Each subject is assumed to produce data according to the same class of model, but with different parameter values. In a hierarchical model, one might assume that the parameters for each subject are chosen from a normal distribution with mean and variance parameters, and the mean and variance would be the hyperparameters. As usual, one would determine the likelihood of the observed data for all subjects for each combination of the two hyperparameters and each choice of individual parameters. In this example, we see the usual tension between fitting each subject as well as possible (optimal choice of individual parameters) and fitting the group as a whole (by choosing a small variance for the Gaussian hyperparameter). This tension results in a movement of the individual parameters toward the group mean, a desirable characteristic given that we do not desire to overfit the data, and fit the noise in each individual's data.

3.2. Our two examples

Our examples present *generative* models for a cognitive task. Such models describe how the theory, with its probabilities and parameters, produce the observed data. For a given model class, and a given set of parameters, the observed data is produced with a specifiable and derivable probability. Of course, some parameters give rise to higher probabilities than others.

If our model is "heads occur with probability θ" and we observe 20 heads in 25 coin flips, a θ of 0.8 (say) gives rise to a high probability of the observed outcome, and a θ of 0.2 (say) gives rise to a low probability of the outcome. The general Bayesian approach is to convert these differing probabilities into degrees of plausibility or belief for the θ values, based on both the probabilities assigned to the observed data by the different θ values, and also their prior probabilities.

It is becoming common to represent probabilistic generative models as graphical models (for introductions, see Griffiths et al., 2008; Jordan, 2004; Lee, 2008). We believe these conventions are quite useful, and deserve to be seen and understood by members of our field. Hence, we present our models using these descriptive formalisms, and try to aid understanding by showing how the example models are represented in this format.

4. Example 1: Memory retention

Finding a lawful relationship between memory retention and time is about the oldest cognitive modeling question, going back to Ebbinghaus in the 1880s. The usual experiment involves giving people (or animals) many items of information on a list, and then testing their ability to remember items from the list after different periods of time have elapsed. Various mathematical functions, usually with psychological interpretations, have been proposed as describing the relation between time and the level of retention. These include models like exponential decay, power, and hyperbolic functions (Rubin, Hinton, & Wenzel, 1999; Rubin & Wenzel, 1996).

Our case study relies on a simplified version of the exponential decay model. The model assumes that the probability an item will be remembered after a period of time t has elapsed is $\theta_t = \exp(-\alpha t) + \beta$, with the restriction $0 < \theta_t < 1$. The α parameter corresponds to the rate of decay of information. The β parameter corresponds to a baseline level of remembering that is assumed to remain even after very long time periods. This model may or may not be regarded as a serious theoretical contender in the memory retention modeling literature, but is useful for simulation and illustrative purposes. Our analyses are based on fictitious data from a potential memory retention study.

Our fictitious data are given in Table 1, and relate to 4 participants tested on 18 items at 10 time intervals: 1, 2, 4, 7, 12, 21, 35, 59, 99, and 200 sec. The number of items tested and the first 9 time intervals are those used by Rubin et al. (1999) in an attempt to consider data that realistically could be measured in a psychological experiment. Each datum in Table 1 simply counts the number of correct memory recalls for each participant at each time interval. Included in Table 1 are missing data, shown by dashes, so that we can test the prediction and generalization properties of models. All of the participants have missing data for the final time period of 200 sec, so we can test the ability of the model to generalize to new measurements. For Participant 4, there are no data at all, so we can test the ability of models to generalize to new participants.

Table 1
Fictitious memory retention data, giving the number out of 18 items correctly recalled for three participants over nine time intervals and including an extra retention interval of 200 sec and an extra participant as missing data

	Time Interval In Seconds									
Participant	1	2	4	7	12	21	35	59	99	200
1	18	18	16	13	9	6	4	4	4	—
2	17	13	9	6	4	4	4	4	4	—
3	14	10	6	4	4	4	4	4	4	—
4	—	—	—	—	—	—	—	—	—	—

4.1. No individual differences

4.1.1. Graphical model

The graphical model for our first attempt to account for the data is shown in Fig. 1. In the graphical model, nodes represent variables of interest, and the graph structure is used to indicate dependencies between the variables, with children depending on their parents. We use the conventions of representing continuous variables with circular nodes and discrete variables with square nodes, and unobserved variables without shading and observed variables with shading. For unobserved variables, we distinguish between stochastic variables with single borders and deterministic variables with double borders. We also use plate notation, enclosing with square boundaries subsets of the graph that have independent replications in the model.

The model in Fig. 1 assumes that every participant has the same retention curve, and so there is one true value for the α and β parameters. The outer plate with $j = 1, \ldots, T$ corresponds to the $T = 10$ different time periods, whose values are given by the observed t_j variable. Together with the α and β parameters, these time periods define the probability and item to

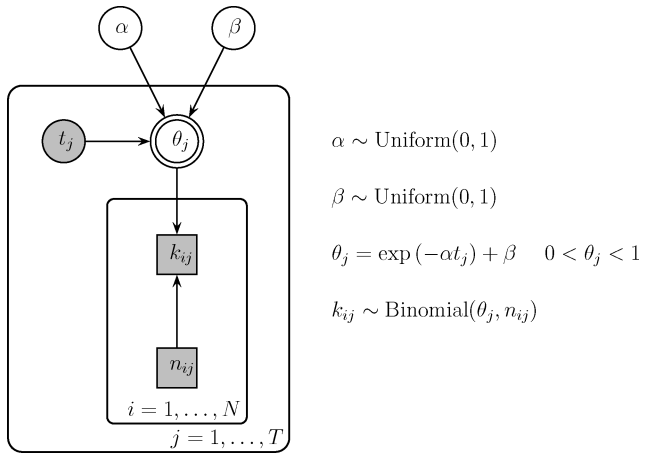

Fig. 1. Graphical model for the exponential decay model of memory retention, assuming no individual differences.

be remembered. The probability of remembering for the jth time period is the deterministic θ_j node.

The inner plate with $i = 1, \ldots, N$ corresponds to the $N = 4$ participants. Each has the same probability of recall at any given time period, but their experimental data, given by the success counts k_{ij} and (potentially) the number of trials n_{ij}, vary, and so are inside the plate. For the data in Table 1, the k_{ij} data are the counts of remembered items and $n_{ij} = 18$ because 18 items were presented for every participant at every time interval. The success counts are binomially distributed according to the success rate and number of trials.

4.1.2. Inference via posterior sampling

The graphical model in Fig. 1 defines a complete probabilistic relation between the model parameters and the observed data. The graphical model is a generative one, specifying how an α rate of decay rate and a β level of permanent retention combine to produce observed retention performance. Once the data in Table 1 are observed, each set of parameter values assigns a probability to that data set, and Bayesian inference allows us to reverse the generative process and assign probabilities to the various parameter sets. The posterior probability distribution represents this information, specifying the relative probability of each possible combination of α and β being the ones that generated the data.

Modern Bayesian inference approximates the posterior distribution by drawing samples using computational methods. Throughout this case study, we implement the graphical models using WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000), which uses a range of MCMC computational methods including adaptive rejection sampling, slice sampling, and Metropolis–Hastings (e.g., see Chen, Shao, & Ibrahim, 2000; Gilks, Richardson, & Spiegelhalter, 1996; MacKay, 2003) to perform posterior sampling. The basic idea is that, over a large number of samples, the relative frequency of a particular combination of parameter values appearing corresponds to the relative probability of those values in the posterior distribution. This correspondence allows the information that is conceptually in the exact joint posterior distribution to be accessed approximately by simple computations across the posterior samples. For example, a histogram of the sampled values of a variable approximates its marginal posterior distribution, and the arithmetic average over these values approximates its expected posterior value.

4.1.3. Results

We evaluated the retention model in Fig. 1 using the data in Table 1, by drawing 10^5 posterior samples after a "burn-in" period (i.e., a set of samples that are not recorded, so that the sampling algorithms can adapt) of 10^3 samples. The joint posterior distribution over α and β is shown in the main panel of Fig. 2, as a two-dimensional scatterplot. Each of the 50 points in the scatterplot corresponds to a posterior sample selected at random from the 10^5 available. The marginal distributions of both α and β are shown below and to the right, and are based on all 10^5 samples. The marginals show the distribution of each parameter, conditioned on the data, considered independently from (i.e., averaged across) the other parameter.

It is clear from Fig. 2 that the joint posterior carries more information than the two marginal distributions. If the joint posterior were independent, it would be just the product of the two marginals and would carry no extra information. However, the joint posterior shows a mild

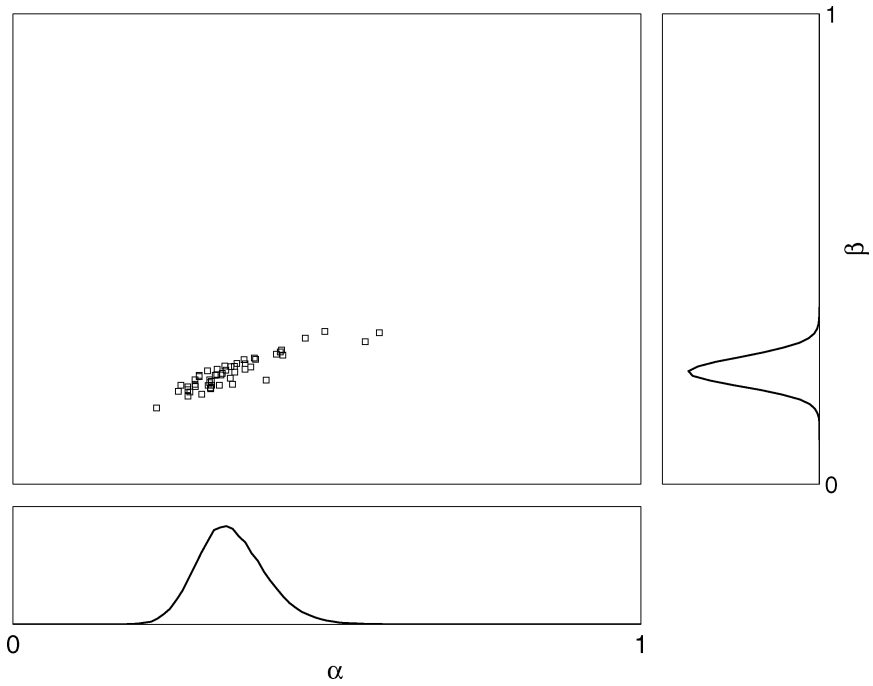

Fig. 2. The joint posterior over the decay and permanent retention parameters α and β for the model that assumes no individual differences.

relationship, with larger values of α generally corresponding to larger values of β. This can be interpreted psychologically as meaning the relatively higher baselines are needed to model the data if relatively greater rates of decay are used.

Fig. 3 shows the posterior predictive distribution over the number of successful retentions at each time interval. The posterior predictive is the prediction about observed data for each possible combination of parameter values under the model, weighted according to the posterior probability of each combination (as represented in Fig. 2). For each participant, at each interval, the squares show the posterior mass given to each possible number of items recalled. These correspond to the models predictions about observed behavior in the retention experiment, based on what the model has learned from the data. Also shown, by the black squares and connecting lines, are the actual observed data for each participant, where available.

It is important to understand that the predictions shown are not generated for each time lag independently. Rather, for each sampled posterior parameter value we generate predictions for all time points, and this procedure is repeated to produce the observed predictions. This is the same generative procedure used to determine the likelihood of the observed data in the process of determining the posterior for the model parameters.

The obvious feature of Fig. 3 is that the current model does not meet a basic requirement of descriptive adequacy. For both Participants 1 and 3, the model gives little posterior probability to the observed data at many time periods. It predicts a steeper rate of decay than shown by the data of Participant 1, and a shallower rate of decay than shown by the data of

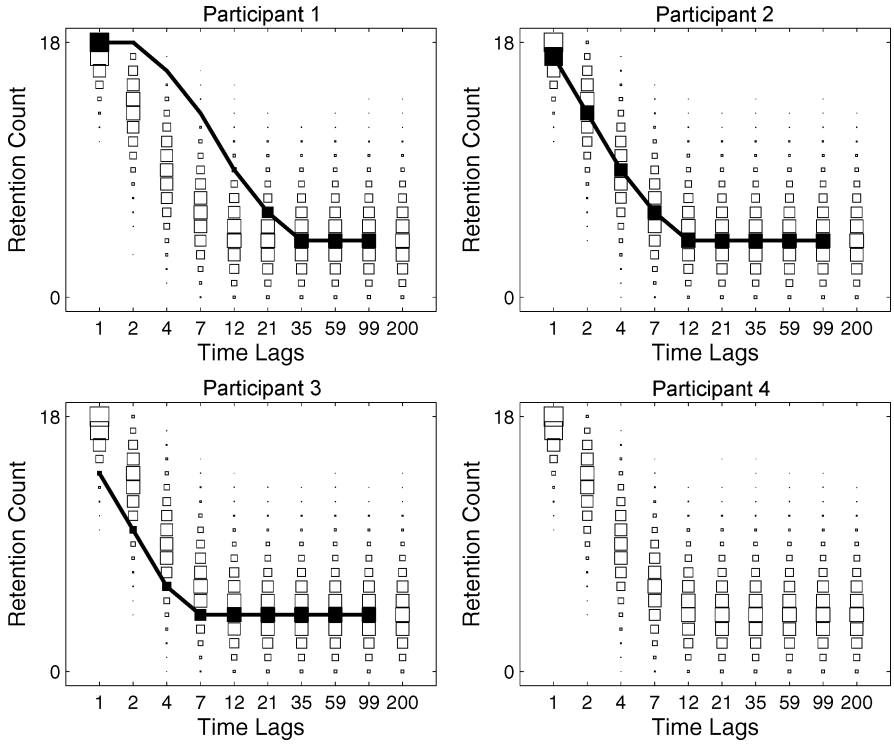

Fig. 3. The posterior predictive for the model that assumes no individual differences against data from the four participants.

Participant 3. Evaluating the model using the posterior predictive analysis, we conclude the assumption that there are no individual differences is inappropriate, and needs to be relaxed in subsequent model development. It is important to understand that this conclusion negates the usefulness of the posterior distribution over parameters, as shown in Fig. 2. This posterior distribution is conditioned on the assumption that the model is appropriate, and is not relevant when our conclusion is that the model is fundamentally deficient.

4.2. Full individual differences

A revised graphical model that does accommodate individual differences is shown in Fig. 4. The change from the previous model is that every participant now has their own α_i and β_i parameters, and that the probability of retention for an item θ_{ij} now changes for both participants and retention intervals.

Once again, we evaluated the model by drawing 10^5 posterior samples after a burn-in period of 10^3 samples. The joint posterior distributions for each participant are shown in the main panel of Fig. 5. Each point on the scatterplot corresponds to a posterior sample, with different markers representing different participants. The first, second, third, and fourth participants use

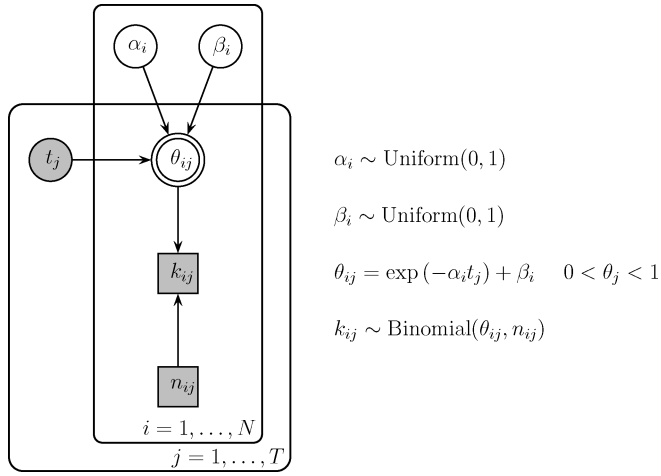

Fig. 4. Graphical model for the exponential decay model of memory retention, assuming full individual differences.

"+," "⋆," "x," and "o" markers, respectively. The marginal distributions are shown below and to the right and use different line styles to represent the participants.

Fig. 6 shows the same analysis of the posterior predictive distribution over the number of successful retentions at each time interval for each participant. It is clear that allowing

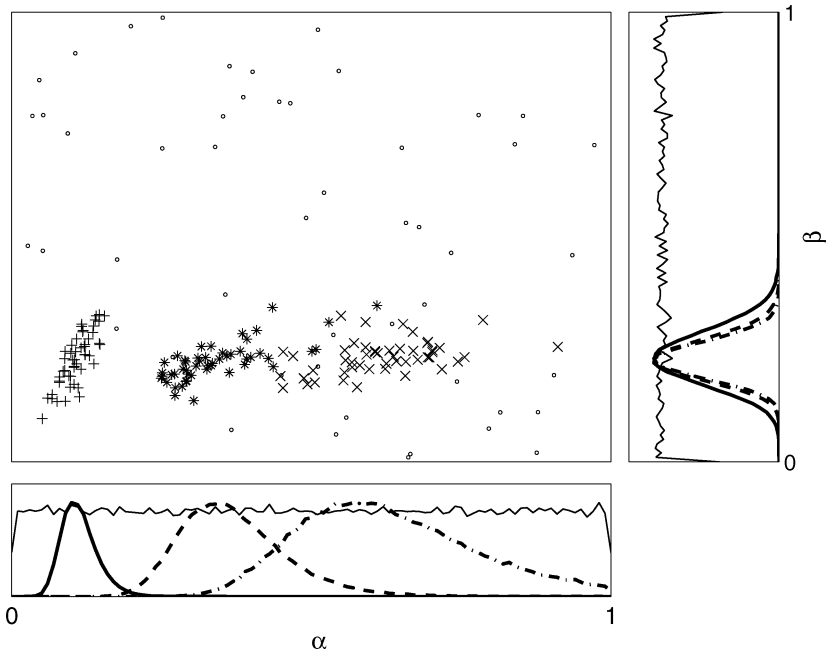

Fig. 5. The joint posterior of all four participants over the decay and permanent retention parameters α and β, for the model that assumes full individual differences.

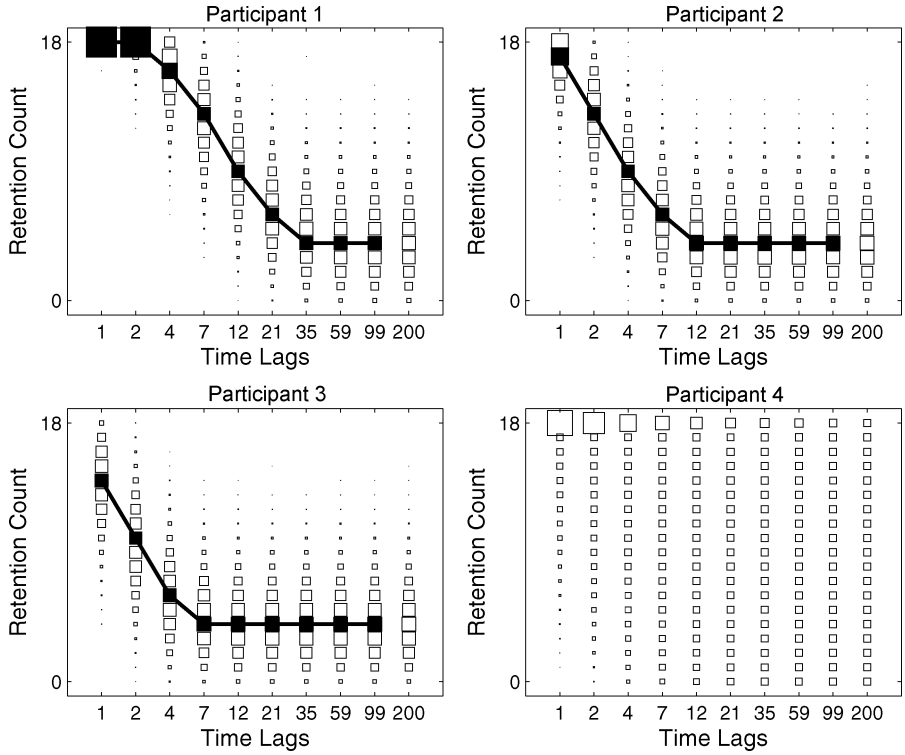

Fig. 6. The posterior predictive for the model that assumes full individual differences, against data from the four participants.

for individual differences lets the model achieve a basic level of descriptive adequacy for Participants 1 and 3. The posteriors in Fig. 5 show that different values for the α decay parameter are used for Participants 1, 2, and 3, corresponding to our intuitions from the earlier analysis.

The weakness in this model is evident in its predictions for Participant 4. Because each participant is assumed to have decay and permanent retention parameters that are different, the only information the model has about the new participant are the priors for the α and β parameters. The relations between parameters for participants that are visually evident in Fig. 5 are not formally captured by the model. This means, as shown in Fig. 5, the posteriors for Participant 4 are just the priors, and so the posterior predictive does not have any useful structure. In this way, this model fails a basic test of generalizability because it does not make sensible predictions for the behavior of future participants.

Intuitively, one might want to predict that Participant 4, will be likely to have model parameters represented by some sort of average of Participants 1 to 3. Carrying this intuition a bit further, one might also want Participants 1 to 3 to have their highest likelihood parameters closer to their group mean than is the case when choosing individual parameters independently. These intuitions are captured formally in the hierarchical model we turn to next.

4.3. Structured individual differences

The relation between the parameters of structures is naturally addressed in a hierarchical model, which is able to represent knowledge at different levels of abstraction in a cognitive model. Just as the data have been assumed to be generated by the latent decay and permanent retention parameters for individual participants, we now assume that those parameters themselves are generated by more abstract latent parameters that describe group distributions across participants.

The specific graphical model we used to implement this idea is in Fig. 7. The key change is that now we are modeling the variation in the different α_i and β_i parameters for each participant by assuming they have a Gaussian distribution across participants. This means that the α_i and β_i parameters are now sampled from over-arching Gaussian distributions, themselves with unknown parameters in the form of means μ_α and μ_β and precisions λ_α and λ_β.

Because they are now sampled, the α_i memory decay and β_i permanent retention parameters no longer have priors explicitly specified, but inherit them from the priors on the means and precisions of the Gaussian distributions. It is important to understand this means inferences made for one participant influence predictions made for another. Because the means and precisions of the group-level distributions are common to all participants, what is learned about them from one participant affects what is known about another. It is in this way the hierarchical model formally represents the relations between participants.

Once again, we evaluated the model by drawing 10^5 posterior samples after a burn-in period of 10^3 samples. The joint and marginal posterior distributions for this model are

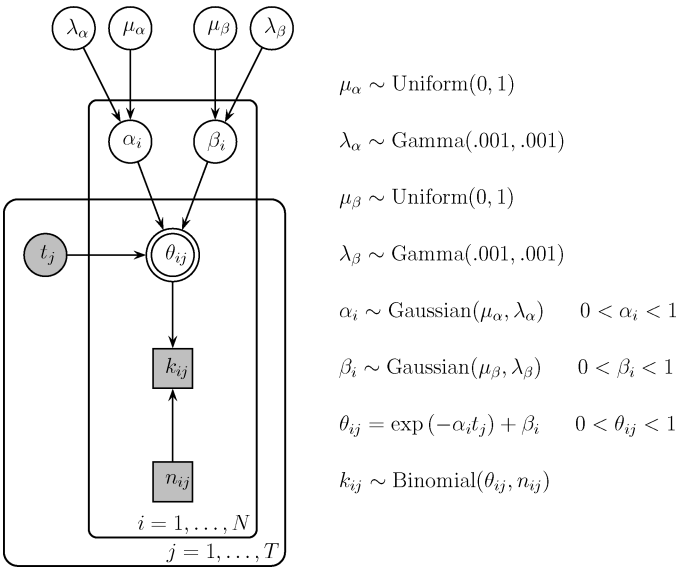

Fig. 7. Graphical model for the exponential decay model of memory retention, assuming structured individual differences.

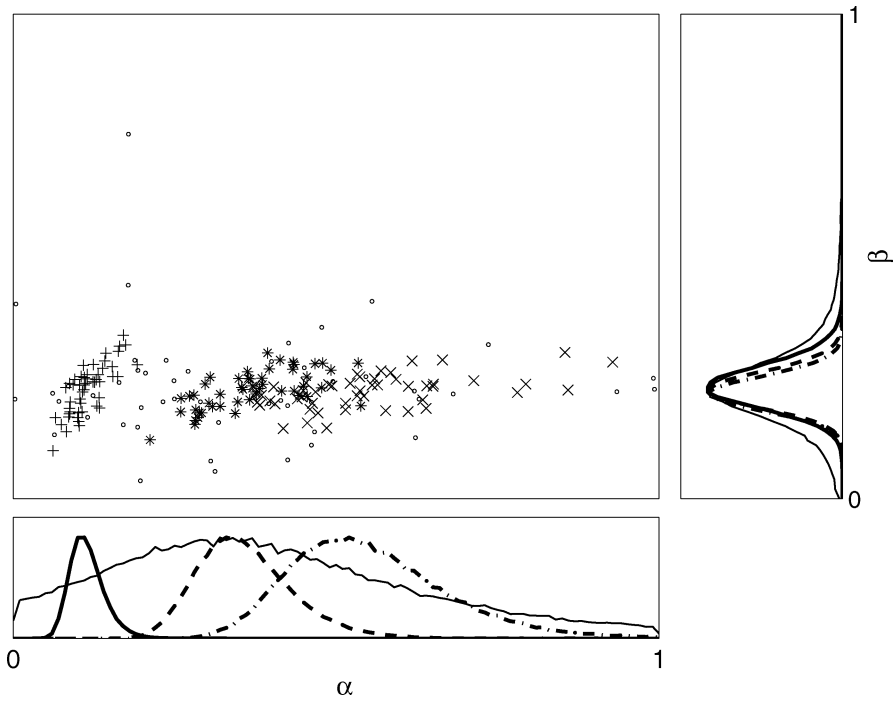

Fig. 8. The joint posterior of all four participants over the decay and permanent retention parameters α and β, for the model that assumes full individual differences.

shown in Fig. 8 using the same markers and lines as before. For Participants 1, 2, and 3, these distributions are extremely similar to those found using the full individual differences model. The important difference is for Participant 4, who now has sensible posterior distributions for both parameters. For the decay parameter α, there is still considerable uncertainty, consistent with the range of values seen for the first three participants; but for the permanent retention parameter β, Participant 4 now has a much more constrained posterior.

The posterior predictive distributions for each subject under the hierarchical model are shown in Fig. 9. The predictions remain useful for the first three participants, and are now also appropriate for Participant 4. This effective prediction for a participant from whom no data have yet been collected arises directly from the nature of the hierarchical model. Based on the data from Participants 1, 2, and 3, inferences are made about the means and precisions of the group distributions for the two parameters of the retention model. The new Participant 4 has values sampled from the Gaussians with these parameters, producing the sensible distributions in Fig. 8 that lead to the sensible predictions in Fig. 9.

4.4. Comparing models

At this point, we have developed a model that seems to describe adequately the observed retention data, makes sensible predictions about a future time interval for which no data

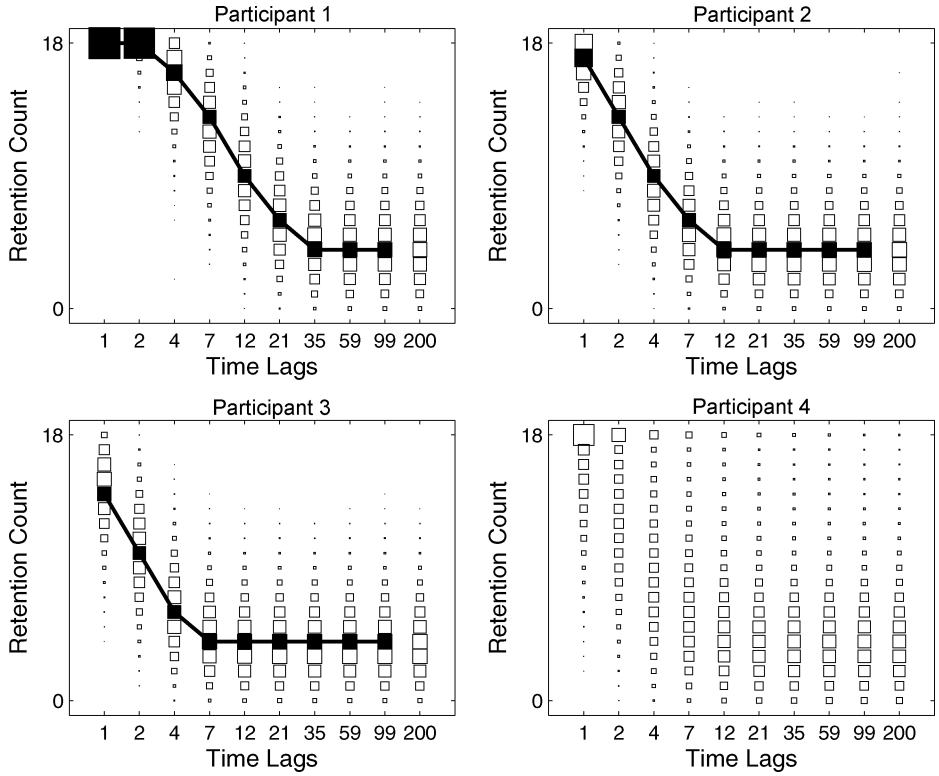

Fig. 9. The posterior predictive for the model that assumes structured individual differences, against data from the four participants.

have been collected, and generalizes reasonably to a new participant for whom no data have been collected. A useful question to ask is whether the same properties could be achieved with a simpler version of the model. Visual examination of the marginal distribution of the permanent retention parameter β for each participant in Figs. 5 and 8 suggests that there might not be individual differences for this aspect of retention. This observation could also be supported by examination of the marginal posterior for the precision λ_β, which we have not shown.

The obvious possibility for a simpler model, then, is one that assumes a single β parameter for all participants, but retains the full hierarchical account for the α decay parameter. It is straightforward to formulate the corresponding graphical model, and its parameter estimation and prediction properties are indeed extremely similar to the hierarchical model in Fig. 7. Our analysis here seeks to evaluate formally the simpler model against the more complicated one from which it was developed.

To do the evaluation, we use the graphical model in Fig. 10 as a means of calculating the Bayes factor. The graphical model represents the full hierarchical model on the left as Model A; and the simplified version with the single β parameter on the right as Model B. These two models independently generate their predicted retention rates θ_{ij}^A and θ_{ij}^B for each participant

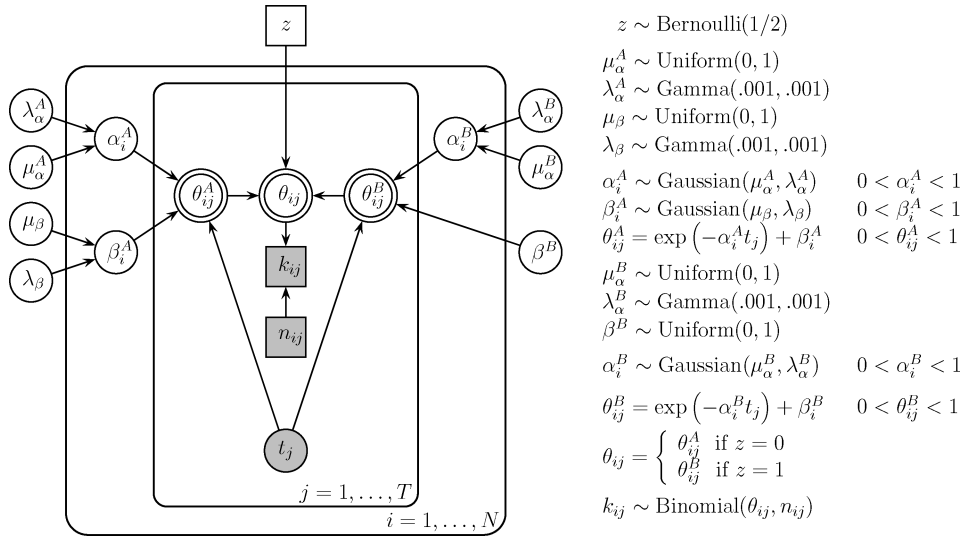

Fig. 10. Graphical model for comparing a full hierarchical model of retention (Model A on the left) to a simpler version that assumes no individual differences in the β parameter (Model B on the right) using a latent model indicator variable z to move between the models.

at each time interval. Which of these is used to model the observed data is determined by the latent binary variable z. When $z = 0$, the retention rate of the simpler model is used; but when $z = 1$, the rate of the full model is used.

The posterior sampling of z, counting the proportion of times it is 0 and 1, then amounts to an evaluation of the relative usefulness of each model. By setting a Bernoulli $(1/2)$ prior on z, its posterior mean $\bar{z}$ estimates the Bayes factor as $\bar{z}/(1 - \bar{z})$.[2]

We evaluated the Bayes factor using the graphical model in Fig. 10, and drawing 10^5 posterior samples after a burn-in period of 10^3 samples for four independent chains (i.e., separate runs of the sampling algorithm, with two initialized with $z = 0$ and the other two initialized with $z = 1$). We observed that the four chains converged to give the mean $\bar{z} = 0.998$, corresponding to a Bayes factor of about 900 in favor of the simpler model.

As mentioned earlier, because probabilities and odds lie on a meaningful scale calibrated by betting, this Bayes factor can be interpreted in the context of the research question it addresses. Our conclusion would be that there is strong evidence that the permanent retention level does not differ across participants, and the simpler model is the better one.

5. Example 2: The SIMPLE model

In this example, we move beyond toy models and fabricated data, and consider a recently proposed model of memory and seminal data. Brown, Neath, and Chater (2007) proposed a temporal ratio model of memory called SIMPLE. The model assumes memories are encoding with a temporal component, but that the representations are logarithmically compressed, so

that more distant memories are more similar. The model also assumes distinctiveness plays a central role in performance on memory tasks, and that interference rather than decay is responsible for forgetting. Perhaps most importantly, SIMPLE assumes the same memory processes operate at all time scales, unlike theories and models that assume different short- and long-term memory mechanisms.

Brown et al. (2007) evaluated SIMPLE on a wide range of memory tasks, fitting the model to many classic data sets from the memory literature. All of the parameter fitting is based on minimizing a sum-squared error criterion, producing point parameter estimates, and goodness of fit is primarily assessed using R^2 variance explained measures. Although this approach provides a useful first look at how the model relates to data, it allows only limited exploration of what the model tells us about human memory.

Brown et al. (2007) seemed aware of these limitations, saying, "We report R^2 values as a measure of fit despite the problems with the measure; direct log-likelihood calculations and model comparison are infeasible in most cases" (p. 545). It is probably worth pointing out the claim about infeasibility of direct log-likelihood calculations is technically inaccurate. The sum-squared error criterion used corresponds exactly to a log-likelihood if the data are assumed to be drawn from Gaussian distributions with common variance (see I. J. Myung, 2003, for a tutorial). In this sense, the analyses reported by Brown et al. already incorporate direct log-likelihood calculations, although with an unusual choice of likelihood function. The current (implied) Gaussian choice assumes, among other things, that a 0.99 probability of recall is as variable as a 0.50 probability, and allows for the possibility of recall probabilities less than 0 and greater than 1. A more natural choice of likelihood function, which we adopt, is a binomial that relates the k_i times the ith item was recalled in the n total trials across all participants to a θ_i probability of recall.

More importantly, it is not difficult to implement fully Bayesian analyses of the SIMPLE model. Our goal in this example is to show how the straightforward application of hierarchical Bayesian analysis permits stronger evaluation and deeper exploration of the model.

5.1. Bayesian analysis of SIMPLE

We focus our demonstration on the first application considered by Brown et al. (2007), which involves seminal immediate free recall data reported by Murdock (1962). The data give the proportion of words correctly recalled averaged across participants, for lists of 10, 15, and 20 words presented at a rate of 1 sec per word; and lists of 20, 30, and 40 words presented at a rate of 2 sec per word.

Brown et al. (2007) made some reasonable assumptions about undocumented aspects of the task (e.g., the mean time of recall from the end of list presentation), to set the time, T_i, between the learning and retrieval of the ith item. With these times established, the application of the SIMPLE free-recall data involves five stages, as conveniently described in the Appendix in Brown et al.

First, the ith presented item, associated with time T_i, is represented in memory using logarithmic compression, given by $M_i = \log T_i$. Second, the similarity between each pair of items is calculated as $\eta_{ij} = \exp(-c|M_i - M_j|)$, where c is a parameter measuring the "distinctiveness" of memory. Third, the discriminability of each pair of items is calculated

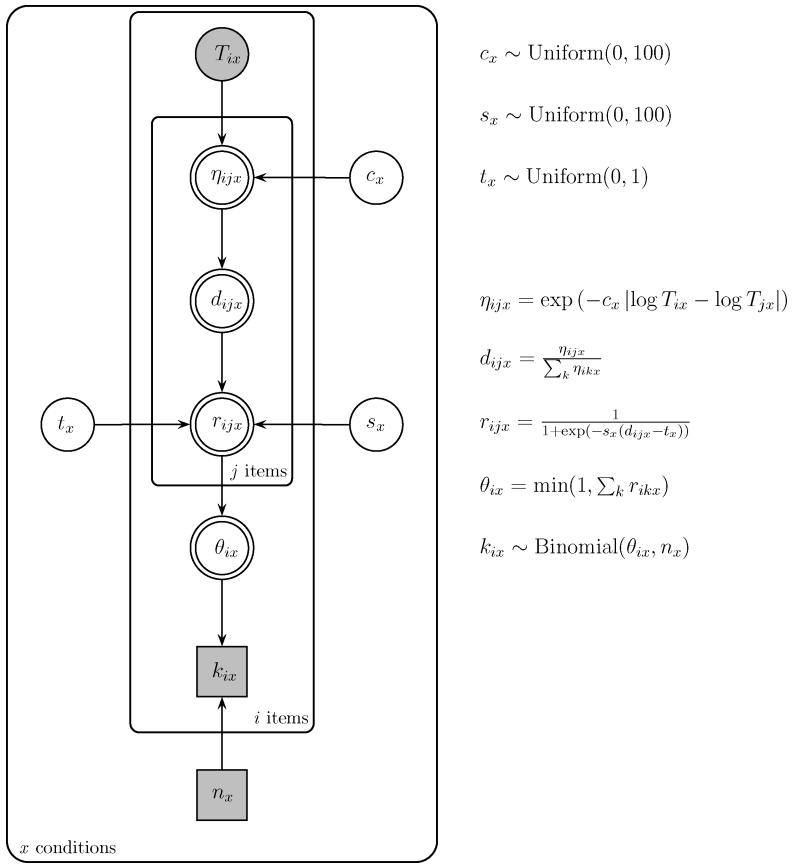

Fig. 11. Graphical model implementing the SIMPLE model of memory.

as $d_{ij} = \eta_{ij} / \sum_k \eta_{ik}$. Fourth, the retrieval probability of each pair of items is calculated as $r_{ij} = 1/(1 + \exp(-s(d_{ij} - t)))$, where t is a threshold parameter and s is a threshold noise parameter. Finally, the probability the ith item in the presented sequence will be recalled is calculated as $\theta_i = \min(1, \sum_k r_{ik})$.

5.1.1. Graphical model

These stages are implemented by the graphical model shown in Fig. 11, which makes it possible to subject SIMPLE to a fully Bayesian analysis. The graphical model has nodes corresponding to the observed times between learning and retrieval, T_i, and the observed number of correct responses k_i for the ith item and total trials n. The similarity (η_{ij}), discriminability (d_{ij}), retrieval (r_{ij}), and free-recall probability (θ_i) nodes are deterministic and simply link the time properties of the items to their accuracy of recall according to the SIMPLE model and its three parameters.

In Fig. 11 the times, responses, and free-recall probabilities apply per item, and so are enclosed in a plate replicating over items. The similarity, discriminability, and retrieval measures

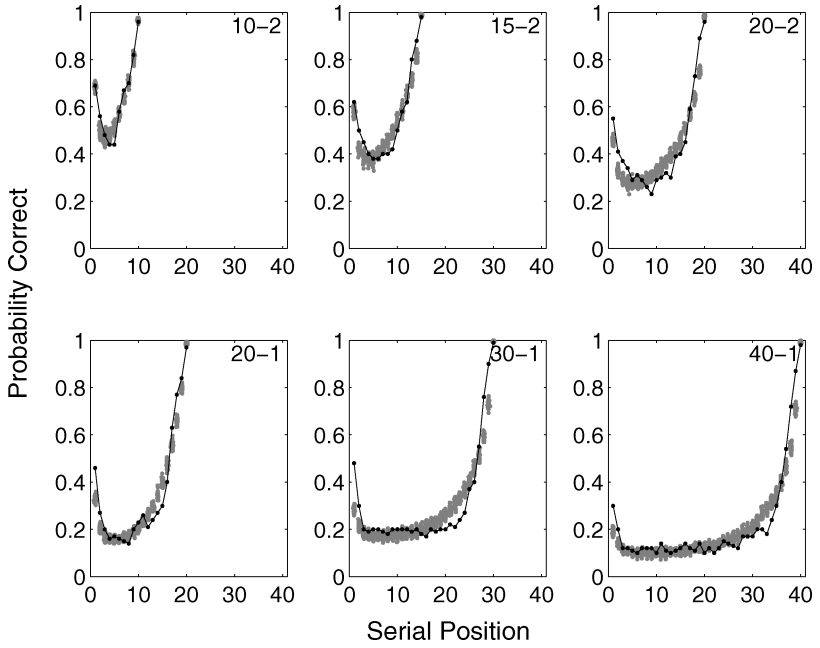

Fig. 12. Posterior prediction of SIMPLE model for the six conditions of the Murdock (1962) immediate free-recall data. The solid lines show the data, and the gray areas show 50 posterior predictive samples for the item at each serial position. The conditions are labeled according to the number of items and the rate of presentation so that, for example, the "10-2" condition had 10 items presented at 1 sec per item.

apply to pairs of variables, and so involve an additional plate also replicating over items. We follow Brown et al. (2007) by fitting the c, t, and s parameters independently for each condition. This means the entire graphical model is also enclosed in a plate replicating over the $x = 1, \ldots, 6$ conditions in the Murdock (1962) data.

5.1.2. Results

Our results are based on 10^5 posterior samples, collected after a burn-in of 10^5 samples, and using multiple chains to assess convergence. Fig. 12 shows the posterior prediction of the SIMPLE model for the six Murdock (1962) data sets. The solid lines show the probability the item in each serial position was correctly recalled. A total of 50 samples from the posterior predictive are shown for each serial position as gray points, making a gray area that spans the range in which the model expects the data to lie. It is clear that, consistent with the excellent R^2 model fits reported by Brown et al. (2007), the SIMPLE model accounts well for all of the serial position curves.

Where the Bayesian approach significantly extends the original model fitting is in understanding the inference made about the parameters. Fig. 13 shows the joint posterior parameter distribution as a three-dimensional plot, with 20 posterior samples for each condition shown by different markers. Also shown, projected onto the planes are the pairwise joint distributions

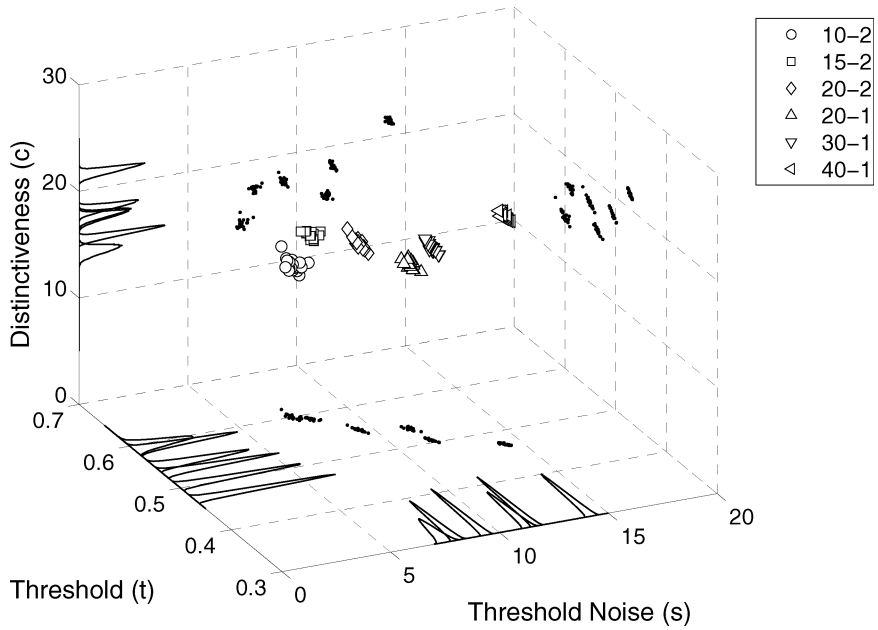

Fig. 13. Joint posterior parameter space for the SIMPLE model for the six conditions of the Murdock (1962) immediate free-recall data.

of each possible combination of parameters (marginalized over the other parameter in each case). Finally, the marginal distributions for each parameter are shown along the three axes.

Fig. 13 attempts to convey the detailed information about the distinctiveness, threshold, and threshold noise parameters provided by the computational Bayesian approach. The point estimates of the original analysis are now extended to include information about variability and co-variation. This additional information is important to understanding how parameters should be interpreted and for suggesting model development. For example, the lack of overlap of the three-dimensional points for the six conditions suggests that there are important differences in model parameters for different item list lengths and presentation rates. In particular, it seems unlikely that an alternative approach to fitting the six conditions using a single discriminability level and threshold function will be adequate.

Another intuition, this time coming from the two-dimensional joint posteriors, is that there is a trade-off between the threshold and threshold noise parameters because their joint distributions (shown by the points in the bottom plane) show a high level of correlation for all of the conditions. This means that the data in each condition are consistent with relatively high thresholds and relatively low levels of threshold noise, or with relatively low thresholds and relatively high levels of threshold noise. This is probably not an ideal state of affairs: Generally, parameters are more easily interpreted and theoretically compelling if they operate independently of each other. In this way, the information in the joint parameter posterior suggests an area in which the model might need further development or refinement.

As a final example of the information in the joint posterior, we note that the marginal distributions for the threshold parameter shown in Fig. 13 seem to show a systematic relation

with item list length. In particular, the threshold decreases as the item list length increases from 10 to 40, with overlap between the two conditions with the most similar lengths (i.e., the "10-2" and "15-2" conditions, and the "20-2" and "20-1" conditions). This type of systematic relation suggests that, rather than treating the threshold as a free parameter, it can be modeled in terms of the known item list length. We now consider how this idea can be implemented in a hierarchical extension to the SIMPLE model.

5.2. A hierarchical Bayesian extension of SIMPLE

5.2.1. Graphical model

Our hierarchical Bayesian extension of SIMPLE is represented by the graphical model shown in Fig. 14. There are two important changes from the model that replicated the assumptions of Brown et al. (2007). First, the distinctiveness (c) and threshold noise (s) parameters are now assumed to have the same value for all experimental conditions. In Fig. 14, their nodes are outside the plate replicated over conditions, and they are no longer indexed by x.

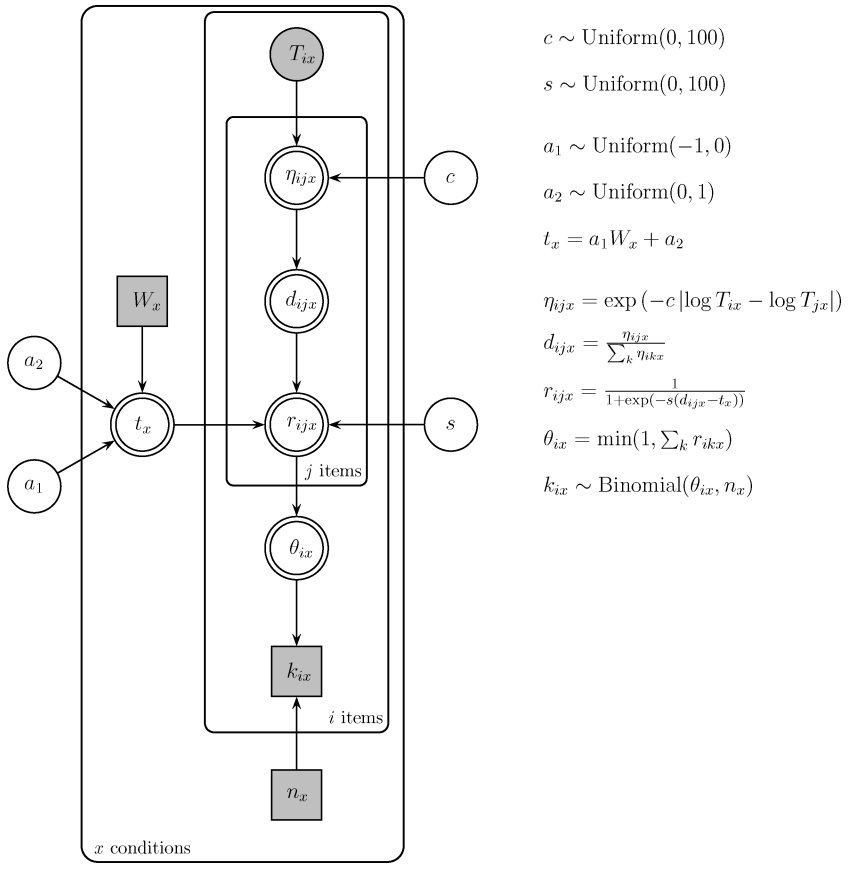

Fig. 14. Graphical model implementing a hierarchical extension to the SIMPLE model of memory.

We do not believe this is a theoretically realistic assumption (indeed, as we pointed out, the joint posterior in Fig. 13 argues against it), but it allows us to construct a tutorial example to demonstrate our main point.

It is the second change that captures this main point and corresponds to the way the thresholds t_x are determined. Rather than being assumed to be independent, these thresholds now depend on the item list length, denoted W_x for the xth condition, via a linear regression function $t_x = a_1 W_x + a_2$ parameterized by the coefficients a_1 and a_2. Consistent with the intuitions gained from Fig. 13, we make the assumption the linear relationship expresses a decrease in threshold as item list length increases, by using the prior $a_1 \sim \text{Uniform}(-1, 0)$.

The goal of our hierarchical extensions is to move away from thinking of parameters as psychological variables that vary independently for every possible immediate serial recall task. Rather, we now conceive of the parameters as psychological variables that themselves now need explanation, and attempt to model how they change in terms of more general parameters.

This approach not only forces theorizing and modeling to tackle new basic questions about how serial recall processes work, but also facilitates evaluation of the prediction and generalization capabilities of the basic model. By making the threshold parameter depend on characteristics of the task (i.e., the number of words in the list) in systematic ways, and by treating the other parameters as invariant, our hierarchical extension automatically allows SIMPLE to make predictions about other tasks.

5.2.2. Results

To demonstrate these capabilities, we applied the hierarchical model in Fig. 14 to the Murdock (1962) conditions, and also to three other possible conditions for which data are not available. These generalization conditions all involve presentations rates of 1 sec per item, but with 10, 25, and 50 items corresponding to both interpolations and extrapolations relative to the collected data.

Our results are again based on 10^5 posterior samples from the graphical model, collected after a burn-in of 10^5 samples, and using multiple chains to assess convergence. The posterior predictive performance is shown in Fig. 15. The top two rows show the Murdock (1962) conditions, whereas the bottom row shows the predictions the model makes about the generalization conditions.

Fig. 16 shows the modeling inferences about the distinctiveness, threshold noise, and threshold parameters. For the first two of these, the inferences take the form of single posterior distributions. For the threshold parameter, however, the posterior inference is now about its functional relationship to item list length. The posterior distribution for this function is represented in Fig. 16 by showing 50 posterior samples at each possible length $W = 1, \ldots, 50$. These posterior samples are found by taking joint posterior samples (a_1, a_2) and finding $t = a_1 W + a_2$ for all values of W.

We emphasize that the particular model being evaluated is not being proposed as a useful or realistic one, but simply to demonstrate what hierarchical Bayesian methods can provide for evaluating a cognitive process model like SIMPLE. In this context, we emphasize that Fig. 15 shows the descriptive adequacies and inadequacies of the hierarchical model in relation to the available data, and details its prediction to new experimental situations for which data are not available. We also emphasize that Fig. 16 shows not only the posterior distribution

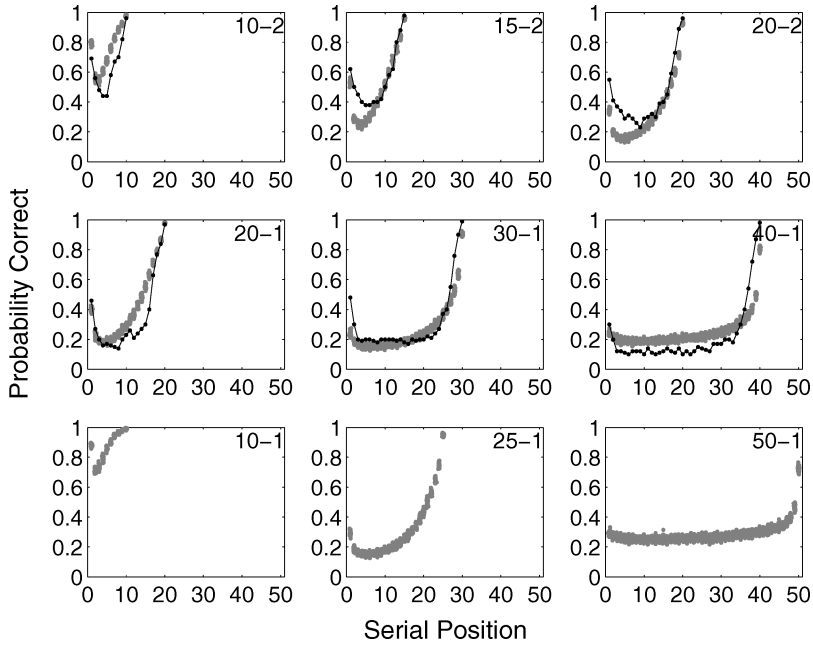

Fig. 15. Posterior prediction of the hierarchical extension of the SIMPLE model for the six conditions of the Murdock (1962) immediate free-recall data and in generalizing to three new conditions. The solid lines show the data, and the gray areas show 50 posterior predictive samples for the item at each serial position.

of individual parameters, but the posterior distribution over the functional relation between parameters and characteristics of the experimental task.

6. Discussion

In this article, we have motivated the use of hierarchical Bayesian methods by arguing they provide the sorts of inferential power and flexibility to evaluate and refine cognitive models. We have tried to demonstrate this power and flexibility using two worked examples, relying

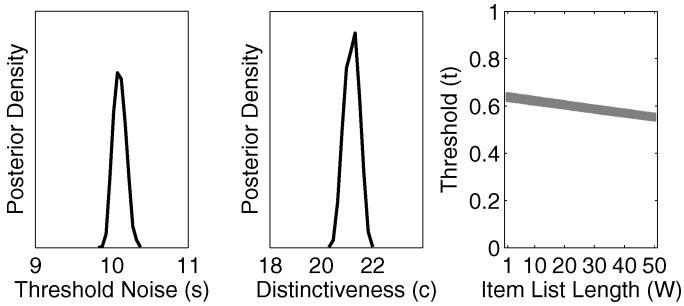

Fig. 16. Posterior parameter inference for the SIMPLE model parameters in its hierarchically extended form.

on graphical modeling and posterior sampling as a means of specifying and doing inference with psychological models and data.

One possible practical objection is that it may require considerable intellectual investment to master Bayesian methods, leaving a temptation to continue applying traditional frequentist techniques. It is true that traditional frequentist statistical approaches are commonly applied to hierarchical models. However, we think it is extremely difficult to make sure these traditional methods are adequately implemented, and almost impossible to insure they reach the same standards as the Bayesian approach. Traditional approaches do not naturally represent uncertainty, automatically control for model complexity, easily marginalize nuisance parameters, represent relevant prior information, work in the same way for all sample sizes, or do a range of other things needed to guarantee good inference. For some hierarchical models and data sets and analyses, it may be that these deficiencies do not affect the conclusions, but there must always be the concern that one or more of them is causing a problem. Remedying the inherent weakness of traditional statistical methods requires *ad-hoc* workarounds, which may have to be tailored for each specific problem, and usually demand considerable statistical sophistication. All of these difficulties in retaining traditional statistical methods stand in stark contrast to the conceptual and practical simplicity of implementing the Bayesian approach. For this reason, we believe the investment in learning Bayesian methods will be worthwhile for any researcher interested in making complete, coherent, and principled inferences about their models and data.

In this context, our examples show advantages both of Bayesian analysis generally and hierarchical Bayesian analysis in particular, for the purposes of model evaluation and comparison. The information in posterior distributions over parameters, and posterior predictive distributions over data, both provide very direct information about how a model accounts for data and allow strengths in a model to be identified and weaknesses to be remedied. Allowing hierarchical development also means that modeling can take place at different levels of psychological abstraction, so that both the parameters controlling memory retention for individuals, and the parameters controlling individual differences can be considered simultaneously. We showed that it is conceptually straightforward to test alternative models using MCMC methods that can provide measures like Bayes Factors.[3]

Although our examples considered relatively straightforward models, the hierarchical Bayesian framework can, in principle, be applied to any model amenable to probabilistic characterization. Lee (2008) presented a number of additional working examples including the multidimensional scaling model of stimulus representation, the Generalized Context Model of category learning, and a signal detection theory account of decision making. Other excellent applications of hierarchical Bayesian methods to cognitive science models are provided by Rouder and Lu (2005); Rouder, Lu, Morey, Sun, and Speckman (2008); and Rouder, Lu, Speckman, Sun, and Jiang (2005). Finally, Lee and Vanpaemel (this issue) presented an involved category learning example showing, among other things, how hierarchical Bayesian methods can be used to specify theoretically based priors over competing models.

We see the hierarchical Bayesian framework as a powerful and general one for developing, evaluating, and choosing between computational models of cognition. Of course, no method

is perfect, and hierarchical Bayesian methods share with all Bayesian methods the problem of finding suitably formal ways to convert all relevant forms of prior knowledge into prior probability distributions for parameters. Also, the computational demands in fully Bayesian analyses may mean they do not scale to the most simulation-intensive cognitive models. However, we hope our examples make it clear that hierarchical Bayesian methods can make a significant, positive contribution to the enterprise of model development and testing for many areas in cognitive science.

It is noteworthy that advances in the field have brought hierarchical modeling within the reach of many researchers who produce models for data. For example, in a recent article, Cohen, Sanborn, and Shiffrin (2008) examined model selection methods when there are few data for individual participants. They focused solely on two approaches: analyzing participants separately and then combining the results, or analyzing grouped data formed by combining the data from all participants into a single pseudo-participant. They pointed out the obvious advantages of using hierarchical approaches, but did not pursue these partly on the basis that such approaches would be out of reach not only for most non-modelers, but also most modelers. As illustrated in this article, the field is advancing rapidly and, with useful and sophisticated software like WinBUGS increasingly available, we should see hierarchical Bayesian modeling increasingly considered and used by modelers. We would be surprised if this approach does not become the method of choice for years to come.

To sum up, the ability to evaluate whether a model is useful, and to choose between numbers of competing models, is a basic requirement for progress in cognitive science. In this review, we have tried to emphasize the multidimensional nature of model evaluation and selection, arguing that good models should describe data well, allow for inferences about psychologically meaningful variables, be able to predict new data, and facilitate future empirical and theoretical progress. We have reviewed a number of theoretical, simulation-based, and practical validation methods for evaluation and selecting models, but highlighted their limitations in addressing the general question of model evaluation. We think that a practical and useful alternative involves hierarchical Bayesian methods. We have tried to demonstrate in worked examples that these methods offer very general and powerful capabilities for developing, evaluating, and choosing between models of cognition.

Notes

1. This summary is based in part on http://www.faqs.org/faqs/ai-faq/neural-nets/part3/section-12.html, a neural net FAQ maintained by Warren S. Sarle.
2. Technically, our approach could be regarded as a variant of the Carlin and Chib (1995) product space method, which requires specifying "pseudo-priors" through which the parameters of the model not being indexed by the indicator variable are updated. For conceptual simplicity, these pseudo-priors were not included in the graphical model (i.e., we view them as part of the sampling method rather than as part of the substantive probabilistic model).

3. Although our Bayes factor example was between nested models, there is no theoretical difficulty in using the same approach to compare any classes of models.

References

Ahn, W. Y., Busemeyer, J. R., Wagenmakers, E.-J., & Stout, J. C. (2008). Comparison of decision learning models using the generalization criterion method. *Cognitive Science, 32*, 1376–1402.

Bechara, A., Damasio, A. R., Damasio, H., & Anderson, S. (1994). Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition, 50*, 7–15.

Brown, G. D. A., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. *Psychological Review, 114*(1), 539–576.

Browne, M. (2000). Cross-validation methods. *Journal of Mathematical Psychology, 44*, 108–132.

Busemeyer, J. R., & Wang, Y.-M. (2000). Model comparisons and model selections based on generalization criterion methodology. *Journal of Mathematical Psychology, 44*, 171–189.

Carlin, B. P., & Chib, S. (1995). Bayesian model choice via Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B, 57*, 473–484.

Chen, M. H., Shao, Q. M., & Ibrahim, J. G. (2000). *Monte Carlo methods in Bayesian computation*. New York: Spinger-Verlag.

Chib, S. (1995). Marginal likelihood from Gibbs output. *Journal of the American Statistical Association*, *90*, 1313–1321.

Chiu, Y. C., Lin, C. H., Huang, J. T., Lin, S., Lee, P. L., & Hsieh, J. C. (2005, September). *Immediate gain is long-term loss: Are there foresighted decision makers in Iowa gambling task?* Paper presented at the 3rd annual meeting of the Society for Neuroeconomics, Kiawah Island, South Carolina.

Cohen, A., Sanborn, A. N., & Shiffrin, R. M. (2008). Model evaluation using grouped or individual data. *Psychonomic Bulletin & Review, 15*, 692–712.

Courville, A. C., Daw, N. D., & Touretzky, D. S. (2006). Bayesian theories of conditioning in a changing world. *Trends in Cognitive Sciences, 10*(7), 294–300.

Dawid, A. P. (1984). Statistical theory: The prequential approach. *Journal of the Royal Statistical Society A, 147*, 278–292.

Dawid, A. P. (1991). Fisherian inference in likelihood and prequential frames of reference. *Journal of the Royal Statistical Society B, 53*, 79–109.

Dawid, A. P. (1992). Prequential analysis, stochastic complexity and Bayesian inference. In J. M. Bernardo, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian statistics 4* (pp. 109–121). Oxford, England: Oxford University Press.

Dawid, A. P., & Vovk, V. G. (1999). Prequential probability: Principles and properties. *Bernoulli, 5*, 125–162.

Efron, B., & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife and cross-validation. *American Statistician, 37*, 36–48.

Efron, B., & Tibshirani, R. (1997). Improvements on cross-validation: The .632+ bootstrap method. *Journal of the American Statistical Association, 92*, 548–560.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Gelfand, A. E. (1996). Model determination using sampling-based methods. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice* (pp. 148–162). Boca Raton, FL: Chapman & Hall/CRC.

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

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (Eds.). (1996). *Markov chain Monte Carlo in practice*. Boca Raton, FL: Chapman & Hall/CRC.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika, 82*, 711–732.

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.

Grünwald, P. D. (2005). Introducing the MDL principle. In P. D. Grünwald, I. J. Myung, & M. A. Pitt (Eds.), *Advances in minimum description length: Theory and application* (pp. 5–22). Cambridge, MA: MIT Press.

Grünwald, P. D. (2007). *The minimum description length principle*. Cambridge, MA: MIT Press.

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

Jordan, M. I. (2004). Graphical models. *Statistical Science, 19*, 140–155.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795.

Kontkanen, P., Myllymäki, P., & Tirri, H. (2001). Comparing prequential model selection criteria in supervised learning of mixture models. In T. Jaakkola & T. Richardson (Eds.), *Proceedings of the 8th international workshop on artificial intelligence and statistics* (pp. 233–238). San Maleo, CA: Kaufmann Publishers.

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. (this issue). 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.

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.

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

Murdock, B. B., Jr. (1962). The serial position effect in free recall. *Journal of Experimental Psychology, 64*, 482–488.

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

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

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

Myung, I. J., Navarro, D. J., & Pitt, M. A. (2006). Model selection by normalized maximum likelihood. *Journal of Mathematical Psychology, 50*, 167–179.

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

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

Raftery, A. E., Newton, M. A., Satagopan, J. M., & Krivitsky, P. (2007). Estimating the integrated likelihood via posterior simulation using theharmonic mean identity (with discussion). In J. M. Bernado (Ed.), *Bayesian statistics 8* (pp. 1–45). New York: Oxford University Press.

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

Rissanen, J. (1978). Modeling by shortest data description. *Automatica, 14*, 445–471.

Rissanen, J. (1986a). A predictive least-squares principle. *IMA Journal of Mathematical Control and Information*, *3*, 211–222.

Rissanen, J. (1986b). Stochastic complexity and modeling. *The Annals of Statistics, 14*, 1080–1100.

Rissanen, J. (1987). Stochastic complexity. *Journal of the Royal Statistical Society, Series B*, *49*, 223–239, 252–265.

Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory, 42*(1), 40–47.

Rissanen, J. (2001). Strong optimality of the normalized ML models as universal codes and information in data. *IEEE Transactions on Information Theory, 47*, 1712–1717.

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., Morey, R. D., Sun, D., & Speckman, P. L. (2008). A hierarchical process dissociation model. *Journal of Experimental Psychology: General, 137*, 370–389.

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.

Rubin, D. C., Hinton, S., & Wenzel, A. (1999). The precise time course of retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*, 1161–1176.

Rubin, D. C., & Wenzel, A. E. (1996). One hundred years of forgetting: A quantitative description of retention. *Psychological Review, 103*, 734–760.

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

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association*, *88*(422), 286–292.

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

Skouras, K., & Dawid, A. P. (1998). On efficient point prediction systems. *Journal of the Royal Statistical Society B, 60*, 765–780.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society B, 36*, 111–147.

Stone, M. (1977). Asymptotics for and against cross-validation. *Biometrika, 64*, 29–35.

Usher, M., & McClelland, J. L. (2001). On the time course of perceptual choice: The leaky competing accumulator model. *Psychological Review, 108*, 550–592.

Wagenmakers, E.-J., Grünwald, P., & Steyvers, M. (2006). Accumulative prediction error and the selection of time series models. *Journal of Mathematical Psychology, 50*, 149–166.

Wagenmakers, E.-J., Ratcliff, R., Gomez, P., & Iverson, G. J. (2004). Assessing model mimicry using the parametric bootstrap. *Journal of Mathematical Psychology, 48*, 28–50.

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