Psychological Review
2013, Vol. 120, No. 3, 667–678

0033-295X/13/$12.00 DOI: 10.1037/a0032458

THEORETICAL NOTE

Likelihood-Free Bayesian Analysis of Memory Models

Brandon M. Turner
Stanford University

Simon Dennis and Trisha Van Zandt
The Ohio State University

Many influential memory models are computational in the sense that their predictions are derived through simulation. This means that it is difficult or impossible to write down a probability distribution or likelihood that characterizes the random behavior of the data as a function of the model's parameters. In turn, the lack of a likelihood means that these models cannot be directly fitted to data using traditional techniques. In particular, standard Bayesian analyses of such models are impossible. In this article, we examine how a new procedure called approximate Bayesian computation (ABC), a method for Bayesian analysis that circumvents the evaluation of the likelihood, can be used to fit computational models to memory data. In particular, we investigate the bind cue decide model of episodic memory (Dennis & Humphreys, 2001) and the retrieving effectively from memory model (Shiffrin & Steyvers, 1997). We fit hierarchical versions of each model to the data of Dennis, Lee, and Kinnell (2008) and Kinnell and Dennis (2012). The ABC analysis permits us to explore the relationships between the parameters in each model as well as evaluate their relative fits to data—analyses that were not previously possible.

Keywords: likelihood-free inference, ABCDE, BCDMEM, REM, list-length effect

Supplemental materials: http://dx.doi.org/10.1037/a0032458.supp

A great deal of effort in psychological research, especially in cognitive psychology, is focused on the development of mathematical or statistical models that explain how people behave. In cognitive psychology, these models are mainly concerned with how people's performance of a task changes under different task demands. Performance is usually assessed by measuring response accuracy or response time, and task demands are changed when people respond under time pressure, with increased cognitive loads, varying levels of stimulus discriminability, and so forth.

A mathematical or statistical model is a set of equations that describes the relationship between a set of independent variables and the behavior of the subject. These equations, derived from a particular psychological theory, represent the influence of the independent variables on the system as changes in a set of parameters. For instance, consider a high-threshold model of performance in a recognition memory task (Batchelder & Riefer, 1990, 1999). In such tasks, subjects are asked to discriminate items that have been previously encountered in a study episode (targets) from those that have not (distractors). One possible explanation for how people recognize targets is that, with probability R, a memory trace is formed during study. At test, if a trace exists, the subject will respond that the target is "old." If there is no trace, the subject will guess "old" with probability g. If the item is a distractor, no trace will exist, and so, the subject will have to guess. There are, therefore, two kinds of "old" responses: "old" responses to targets, which are called hits, and "old" responses to distractors, which are called false alarms. The probability of a hit is then $R + g(1 - R)$, and the probability of a false alarm is g.

The theory that drives the construction of the model states how the parameters of the model must change as task demands change. For example, as memory researchers, we could use the high-threshold model to test hypotheses about the structure and processes in recognition. Suppose we ask subjects to study targets in two conditions. In the first condition, targets are only presented for 500 ms, whereas in the second condition, targets are presented for 5 s. We might expect the probability of forming a memory trace to be higher in the 5-s study condition than in the 500-ms study condition, which the model would accommodate as a change in the

This article was published Online First April 15, 2013.

Brandon M. Turner, Department of Psychology, Stanford University; Simon Dennis and Trisha Van Zandt, Department of Psychology, The Ohio State University.

Portions of this work were submitted by Brandon M. Turner in partial fulfillment of the requirements for the doctoral degree in psychology at The Ohio State University and were presented at the 7th Annual Context and Episodic Memory Symposium, Philadelphia, Pennsylvania, and the 44th Annual Meeting of the Society for Mathematical Psychology, Portland, Oregon. This work was funded by National Science Foundation Grant SES-1024709 and National Institutes of Health Award F32GM103288. We thank Jay Myung and Per Sederberg for helpful discussions on the research plan for this project and Andrew Heathcote, Erik Reichle, Rich Shiffrin, and Eric-Jan Wagenmakers for helpful comments that improved an earlier version of the manuscript.

Correspondence concerning this article should be addressed to Brandon M. Turner, Department of Psychology, Stanford University, 342 Jordan Hall, Building 420, Stanford, CA 94305. E-mail: turner.826@gmail.com

parameter R. The parameter g, however, should not necessarily change.[1]

Because we expect to see particular changes in a model's parameters over changes in experimental conditions, we must worry about how to estimate those parameters. Sometimes, as in the high-threshold model, the estimation problem is trivial. Let the observed proportions of hits $P(H) = O_T/N_T$, where O_T is the number of "old" responses among the N_T target test trials, and let the observed proportion of false alarms $P(FA) = O_D/N_D$, where O_D is the number of "old" responses among the N_D distractor test trials. It is easy to show that one estimate for g is $\hat{g} = P(FA)$, while an estimate for R is $\hat{R} = [P(H) - P(FA)]/[1 - P(FA)]$. For more complex models, however, the estimation problem is not as straightforward.

Parameter estimation for more complex models typically is performed by minimization of least squares (which is the basis for regression and analysis of variance) or maximization of likelihood. For these methods, we select a discrepancy function (e.g., sums of squares or the likelihood) that gives some indication of the distance between the model's predictions and the observed data. We proceed iteratively, proposing potential parameter values and evaluating the discrepancy function until we arrive at parameter values that bring us model predictions that are as close to the data as possible.

However, to compute the model predictions for a set of parameter values, we usually need a mathematical statement relating the parameters to the dependent variables (for least squares minimization) or the distribution function that describes the random behavior of the dependent variables (for maximum likelihood). For example, for the high-threshold model, the sum-of-squares discrepancy function for least squares minimization could be

$$SSE = \left[P(H) - \left(\hat{R} + \hat{g}\left(1 - \hat{R}\right)\right)\right]^2 + \left[P(FA) - \hat{g}\right]^2.$$

The estimates $\hat{R}$ and $\hat{g}$ we proposed above are, in fact, least squares estimates that make SSE equal to zero.

Maximum likelihood is a slightly different approach in which we attempt to maximize the probability of the observed data rather than minimize the distance between predicted and observed dependent variables. For the high-threshold model, we start by using the parameters R and g to compute the probabilities of hits and false alarms predicted by the model. These probabilities can then be used to compute the probability of the observed data given the parameters R and g. If we observe O_D false alarms in N_D distractor test trials and O_T hits in N_T target test trials, then the probability of a subject's sequence of $N_T + N_D$ responses is

$$Cg^{O_D}(1 - g)^{N_D - O_D}\left[R + g(1 - R)\right]^{O_T}\left[1 - \left(R + g(1 - R)\right)\right]^{N_T - O_T}, \quad (1)$$

where C is a constant reflecting the proportions of target and distractor trials. The function in Equation 1 is called the likelihood function. The estimates for R and g we proposed above maximize this likelihood, and so, we say that they are maximum likelihood estimators.

We now have, especially in memory research, a number of models that have been constructed to mimic (arguably) plausible brain mechanisms (e.g., O'Reilly, 2001, 2006; O'Reilly & Munakata, 2000; Shadlen & Newsome, 2001; Usher & McClelland, 2001). The structure of the models is constrained by our understanding of relevant neuroscience, and hence, they are more bottom-up and frequently rely on simulation to generate predictions. This means that they are sufficiently complex that we cannot always write down the mathematical statements relating dependent variables to parameters, nor can we always write down the probability distributions that the dependent variables follow.[2] All we can do is simulate the models and analyze their data in exactly the same way that we analyze the data from human subjects.

The lack of mathematical expressions for a simulation-based model's predictions makes estimating that model's parameters very tricky. Presently, the most popular estimation algorithms use what is called approximate least squares estimation (e.g., Criss & McClelland, 2006; Malmberg, Zeelenberg, & Shiffrin, 2004). The discrepancy function that guides the search for parameter values is a function of the difference between the data simulated from the model and the observed data. For example, if we were to take this approach with the high-threshold model, we would generate a simulated data set using some candidate values for the parameters g and R and obtain a predicted hit rate $\hat{P}(H)$ and false-alarm rate $\hat{P}(FA)$. This simulation would proceed in the following way: First, using the candidate value for R, we would simulate a target test trial by determining whether or not a memory trace exists. This determination would occur by sampling a Bernoulli random variable (0 or 1) with probability R. If this variable is equal to 1, indicating that a trace exists, we obtain an "old" response from the model. If the variable is equal to 0, indicating that no trace exists, we would then sample a second Bernoulli random variable with probability g, and if this variable is equal to 1, we obtain an "old" guess from the model. Otherwise, we obtain a "new" response. We proceed similarly for distractor test trials, only without having to determine if a trace exists.

After simulating responses for each trial of an experiment, we would then have the observed hit and false-alarm rates $P(H)$ and $P(FA)$ from a subject and the predicted hit and false-alarm rates $\hat{P}(H)$ and $\hat{P}(FA)$ from the simulation using the proposed values of g and R. Our approximate least squares discrepancy function might be

$$\widehat{SSE} = \left(P(H) - \hat{P}(H)\right)^2 + \left(P(FA) - \hat{P}(FA)\right)^2.$$

We could embed this discrepancy function in an optimization routine to find the values of g and R that minimize it.

The approximate least squares approach is unsatisfactory for a number of reasons. First, least squares discrepancy functions that compare samples of random variables (the observed statistics $P(H)$ and $P(FA)$) to samples of other random variables (the simulated data $\hat{P}(H)$ and $\hat{P}(FA)$) can produce less accurate parameter estimates than other estimation methods like maximum likelihood (e.g., Myung, 2003; Rouder, Sun, Speckman, Lu, & Zhou, 2003; Van Zandt, 2000). Second, the evaluation of the discrepancy

[1] In this illustrative and overly simplistic treatment of the high-threshold model, we are not considering the empirical evidence demonstrating that not only does the hit rate increase with increased strength but the false-alarm rate decreases as well. This mirror effect is not explained by the high-threshold model we present here.

[2] For more detailed, statistical explanations of the difficulties in deriving the likelihood function, see Myung, Montenegro, and Pitt (2007) and Turner and Van Zandt (2012).

function is very time consuming and inefficient. Because there is sampling variability in each simulated data set, to obtain accurate values of the discrepancy function for each set of parameters, we cannot just simulate data ($\hat{P}(H)$ and $\hat{P}(FA)$) once. Instead, we must simulate the data many times for each set of proposed parameter values. Third and finally, the kinds of inferences we can make about the estimated parameters are very limited. We do not know how the estimates are distributed, which makes null hypothesis testing difficult, and (most importantly for our purposes) we obtain, after the estimation is over, only point estimates of the parameters; we cannot perform Bayesian analyses.

Bayesian Model Fitting

The approximate least squares approach can be contrasted with the Bayesian approach, which is steadily growing in popularity (e.g., Craigmile, Peruggia, & Van Zandt, 2010; Dennis, Lee, & Kinnell, 2008; Klugkist, Laudy, & Hoijtink, 2010; Kruschke, 2011; Lee, 2004, 2008, 2011; Lee, Fuss, & Navarro, 2006; Lee & Vanpaemel, 2008; Lee & Wagenmakers, 2010; Oravecz, Tuerlinckx, & Vandekerckhove, 2009; Rouder & Lu, 2005; Rouder, Lu, Speckman, Sun, & Jiang, 2005; Rouder et al., 2003; Shiffrin, Lee, Kim, & Wagenmakers, 2008; Steyvers, Lee, & Wagenmakers, 2009; Vandekerckhove, Tuerlinckx, & Lee, 2011; Wagenmakers, 2007). There are at least three reasons for the increased popularity of the Bayesian approach. The first is that Bayesian modeling provides the distribution of the parameters given the data that were observed at each point in the parameter space. This distribution is known as the *posterior* distribution, and it can be used to answer questions about the model that ordinary frequentist approaches cannot.

For example, the posterior distribution can reveal delicate tradeoffs between subsets of parameters that would not be visible under standard model-fitting techniques. Having a clear understanding of how parameters interact with one another and how these parameters change under experimental conditions is essential to fully understanding the model of interest.

The second reason is that the Bayesian approach also provides a convenient framework for performing hierarchical analyses (e.g., Lee, 2011; Lee & Wagenmakers, 2010; Shiffrin et al., 2008). Hierarchical analyses allow for individual differences in the parameters of each subject and circumvent the incorrect inferences that are sometimes drawn when data are averaged across subjects (Estes & Maddox, 2005). Because data are generally collected from many subjects in several different experimental conditions, accounting for individual differences is essential to assessing which model provides the best fit to experimental data.

Finally and most importantly for this article, Bayesian analysis provides a method of selecting the best model from several competing models, taking into account both fits to the data and the complexity of the models with respect to the data. In general, a more complex model will fit better simply by virtue of its increased flexibility. Thus, to select the best model, one must balance the fit of the data against the complexity of the model. However, model complexity is somewhat of a misnomer. As we demonstrate below, the complexity of a model can depend on the experimental design to which it is applied. The Bayesian approach provides a principled way to balance the various facets that must be considered when performing model selection.

In this article, we show how a new approach to evaluating simulation-based models solves these problems. Most importantly, it permits us to treat the model parameters in a Bayesian fashion: We can explore the models' parameter spaces in more detail and investigate meaningful alternative hypotheses concerning those parameters without abandoning the theoretical structure that makes the models interesting in the first place.

We begin by first providing an overview of our likelihood-free Bayesian modeling approach. We then present the results of a simulation study comparing the predictions of the two models when list length is manipulated. By first examining where the two models make different predictions, we can better identify experimental manipulations that will allow for greater model discriminability. We then fit the two models to data that involve such a manipulation. We show how our Bayesian modeling approach can be used to compare the two models on the basis of the interpretations they provide, as well as their relative fits to the data.

In this article, we advocate for likelihood-free Bayesian techniques at a conceptual, rather than technical, level. Readers are encouraged to consult the online supplemental materials. These materials provide (a) more details on the memory models we used here, (b) an overview of the fundamentals of Bayesian inference, (c) technical details of a basic approximate Bayesian computation algorithm, (d) a simulation study comparing likelihood-free and likelihood-informed methods for parameter inference, and (e) the mathematical and methodological details of the hierarchical models used in Studies 1 and 2.

Likelihood-Free Model Evaluation

To demonstrate the utility of our likelihood-free approach, we chose two prominent cognitive models that make different assumptions about how interference occurs in memory: the bind cue decide model of episodic memory (BCDMEM; Dennis & Humphreys, 2001) and the retrieving effectively from memory (REM) model (Shiffrin & Steyvers, 1997). These two models make different assumptions about the source of noise in recognition decisions. BCDMEM posits that interference from different contexts in which a probe item appeared makes recognition difficult, whereas REM posits that interference from the other items in the study context makes recognition difficult.[3] These two different ideas about interference are built into the structure of the models, leading them to make different predictions about performance under a range of experimental conditions. Both models are very powerful and can fit a range of different experimental effects. However, their relative goodness-of-fit to data has not been examined in a rigorous way because their analytic forms are very difficult (and, until very recently, unknown; see Myung et al., 2007). For this reason, the models' predictions and fits to data have been determined by simulation.

Fitting the Models to Data

Both models were originally proposed as simulation-based models, meaning that their likelihood functions were not explicitly

[3] Although later instantiations of REM incorporate both item and context noise, in this article we only consider the pure item–noise version for demonstration.

known at the time of their inception. For BCDMEM, however, Myung et al. (2007) derived both exact and asymptotic equations for the likelihood function. Despite this advance, the exact equations are computationally difficult to evaluate, resulting in long computation times, and the asymptotic equations can lead to inaccurate estimates of the posterior distribution.

By contrast, a closed-form, general expression for the REM model's likelihood has not yet been published.[4] Estimates of REM's parameters have been obtained by "handheld" fits in which parameter values have been adjusted manually over a restricted range (Shiffrin & Steyvers, 1997) or by simulating the model and using approximate least squares procedures that rely on the match between simulated and observed data (e.g., Criss & McClelland, 2006; Malmberg et al., 2004). These procedures severely limit the extent to which inference can be made about the parameters—in particular, how these parameters vary with changes in experimental conditions.

Standard methods for applying Bayesian techniques to model fitting require models to be sufficiently simple that a closed-form expression for the likelihood of the data given the model parameters can be derived. For models like BCDMEM and REM, depending as they do on simulations to obtain point estimates of parameters, Bayesian analyses have been impossible to perform. This means that we have not, to this point, been able to objectively evaluate the relative merits of the context versus item noise theories.

To fit the models to the data, we use approximate Bayesian computation (ABC), an approach that allows full Bayesian inference despite the lack of an expression for the REM likelihood (Turner & Van Zandt, 2012). The ABC approach we use combines two recently developed algorithms designed to make estimation of the posterior distribution as efficient as possible. First, we use a method for proposal generation called differential evolution (DE; see ter Braak, 2006; Turner, Sederberg, Brown, & Steyvers, in press). The DE proposal mechanism has proven useful in estimating high-dimensional, multimodal posterior distributions of highly correlated parameters (Hu & Tsui, 2005; Storn & Price, 1997; ter Braak, 2006; Vrugt et al., 2009). Turner and Sederberg (2012) used the DE approach to develop the approximate Bayesian computation with differential evolution (ABCDE) algorithm and showed that the ABCDE algorithm could outperform other proposal methods. At its core, ABCDE uses a kernel-based weighing scheme (see Wilkinson, 2011) to evaluate how likely a given proposal is to have generated the observed data. To do so, the kernel provides a weight whose magnitude decreases as a function of a distance metric that compares the simulated data to the observed data. Thus, proposal parameter values producing data that are close to the observed data are given a higher weight and are more likely to be retained as an estimate of the posterior distribution (see Turner & Van Zandt, 2012, for a tutorial).

The second algorithm allows us to efficiently fit hierarchical models. Turner and Van Zandt (2011) noted that the conditional distribution of the group-level parameters does not depend on the unknown likelihood function because the group-level parameters depend on the data only through the subject-level parameters. Thus, if the parameter space is partitioned appropriately, we can obtain samples directly from the posterior distribution of the group-level parameters using a technique called Gibbs sampling (see Gelman, Carlin, Stern, & Rubin, 2004) and without any need for ABC. Turner and Van Zandt called their approach, which alternates between Gibbs sampling for the group-level parameters and ABC sampling for the subject-level parameters, Gibbs ABC.

The discussion of our algorithm, which embeds the ABCDE algorithm within the Gibbs ABC framework, is necessarily brief. Interested readers should consult Turner and Sederberg (2012) and Turner and Van Zandt (2011) for more technical details about how the algorithms are implemented. We use our algorithm to fit REM and BCDMEM to data, giving the first fully Bayesian analyses of the models. With estimates of the posterior distributions, we can contrast the different models with each other, particularly with regard to their flexibility, interpretability, and quality of fit.

While our approach for fitting the models to data will provide accurate estimates of the posterior, estimates of the posterior alone will not allow us to directly contrast BCDMEM and REM. Instead, we rely on conventional measures of model fit, such as the deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & van der Linde, 2002) and Bayesian predictive information criterion (BPIC; Ando, 2007), to assess which model matches the data most closely. While measuring model fit may sound straightforward, there are many factors that must be considered, such as model complexity, number of parameters, goodness of model fit to the data, and experimental design.

Model Constraints From Experimental Design

A great deal of attention has been paid to the role that model complexity plays in selecting between different models (e.g., Liu & Aitkin, 2008; Myung & Pitt, 1997; Pitt, Kim, Navarro, & Myung, 2006; Pitt, Myung, & Zhang, 2002). Models that are more complex can fit a wider range of data and therefore often exhibit larger goodness-of-fit values than less complex models. However, more complex models may tend to overfit data, explaining variations in the data that are due to error rather than systematic changes in behavior with experimental conditions. A model's complexity is a function of its likelihood (e.g., Ando, 2007), and so, any independent variable that enters into that likelihood will influence complexity. This fact is not well appreciated: It means that the range of predictions of a model may be in part determined by experimental design.

For BCDMEM and REM, we can exploit the constraint provided by experimental design to provide a stronger test of the models. In particular, we can fit the models to data that include a list-length manipulation. Because neither model has parameters that by themselves can explicitly capture the list-length effect,[5] the degree of model fit depends on the model's architecture. It follows that if the data exhibit the list-length effect, then a model that is designed to produce a list-length effect should be preferred over a model that does not make such a prediction.

In this article, experimental design plays a vital role in testing the two competing list-length theories. Thus, it is essential that we

[4] An unpublished manuscript by Montenegro, Myung, and Pitt (2011) derived a numerical solution for the likelihood for a simple, special-case version of REM. However, this solution requires the numerical integration of difficult expressions, which would need to be modified to fit the models we use in this article. Thus, we do not discuss it here.

[5] We arrived at this conclusion after a number of simulation studies, which are not reported here (also see Myung et al., 2007; Montenegro et al., 2011).

fully understand how the models REM and BCDMEM interact with experimental design prior to fitting the models to data. In the next section, we present the results of a simulation study meant to provide insight into the models' behavior as a function of list length.

Model predictions for list length. Unlike BCDMEM, REM predicts that as the number of items in the study list increases, recognition performance will decrease. This prediction conflicts with findings from Dennis and Humphreys (2001) and Dennis et al. (2008), who showed that with controls for retention interval, attention, displaced rehearsal and variance in contextual reinstatement, there is no list-length effect in recognition memory for words, although there may be a small effect for other stimuli such as fractals and faces (Kinnell & Dennis, 2012).

Because REM has no simple likelihood, there have been few efforts devoted to determining how each of the model's parameters contributes to its ability to produce the list-length effect. However, knowing how changes in each parameter influence the model's predictions is essential to understanding the complexity of the model and where the list-length effect comes from. One way to explore the parameter space is to look at the distribution of all possible data generated by parameters sampled from uninformative priors. The resulting distribution is known as the prior predictive distribution (Gelman et al., 2004; Vanpaemel, 2010).

We simulated the model under different list lengths with parameters that were selected from uninformative priors.[6] The distribution of the resulting simulated data, the prior predictive distribution, gives an image of the model's predictions under the full possible range of parameter values.

We considered four different list lengths: 10, 20, 80, and 2,000 items. The 2,000-item list simulations represented the limiting behavior of the two models. The test list was composed of 10 targets selected at random from the studied items and 10 distractors. To simulate the data, we sampled a set of parameters from the prior, used those parameters to generate simulated data for 20 study-test trials, and recorded the mean hit and false-alarm rates over these trials. We repeated this procedure 10,000 times under each Model $\times$ List Length combination. Given our choice of priors, the only differences we might have observed between the two models' predictions would be due entirely to differences in the models' structures and how the parameters interact with list length.

Figure 1 shows the prior predictive distributions of hit and false-alarm rates in the receiver operating characteristic (ROC) space over the four list-length conditions. The top panels show the data from REM, whereas the bottom panels show the data from BCDMEM. In each panel, the line drawn from the top left corner to the bottom right corner represents unbiased responding, whereas the line drawn from the bottom left corner to the top right corner represents chance performance. BCDMEM shows no systematic changes in its ROC predictions over different list lengths, which we expected because BCDMEM does not predict a list-length effect.

Figure 2 shows the prior predictive distribution separated by hit (left panels) and false-alarm (right panels) rates for REM (top panels) and BCDMEM (bottom panels). In this figure, the rates are plotted as a function of list length for a single parameter value (gray lines). The black lines show the mean of the prior predictive density. Figure 2 shows that the predictions of REM are a function of list length and not the individual parameter values that were sampled.

The prior predictive distributions show that memory performance as measured by the hit and false-alarm rates drops with increasing list lengths in REM but is unaffected in BCDMEM. This means that, even with no changes in the model parameters over conditions, there are areas in the ROC space that will favor one model for short lists and the other model for long lists.

An important fact demonstrated by Figures 1 and 2 is that, as list length increases, the range of effects predicted by REM systematically changes—REM's complexity decreases. This occurs because REM's intractable likelihood is a function of list length. Choosing which model, REM or BCDMEM, fits a set of data best must therefore take into account a possibly complicated interaction between the complexity of the model (Myung & Pitt, 1997) and the experimental design. This is an important consideration that will influence how well REM can compete with other models when explaining data from these designs. In the next two sections, we exploit this implicit constraint by fitting both models to data from experiments with list-length manipulations. We begin with the data presented in Dennis et al. (2008) and then proceed to data from Kinnell and Dennis (2012).

Study 1: Data From Dennis, Lee, and Kinnell (2008)

Dennis et al. (2008) presented subjects with high- and low-frequency words in short and long study lists but only tested the first items in the study list. In this way, they could equate the retention intervals across conditions by delaying the test phase in the short-list conditions by an amount of time equal to the difference between the long and short study phases.

In another condition, they added an additional filler task after both long- and short-list retention intervals. Without such a filler task, the time between study and test for long lists is negligible, whereas, for short lists, it is equal to the duration of the filler task that equates the retention intervals for short and long lists. The additional filler task provides a strong cue that the study context needs to be reinstated to begin the test phase. Without the filler task, long lists do not provide such a reinstatement cue, and so, subjects may not reinstate the study context from the beginning of the study list.

Dennis et al. (2008) found that with no filler task (with no cue for long lists to reinstate the study context from the beginning of the study list), recognition performance was poorer for long than for short lists—a list-length effect. However, with an additional filler task (where both long and short lists were cued to reinstate the study context), there was no list-length effect.

Models

A full description of the hierarchical models appears in the online supplemental materials. We made standard choices about how to extend the two models to hierarchical data. We defined parameters at two levels—some parameters captured behavioral patterns at the group level, whereas other parameters captured subject-specific effects. We pay most attention to the group-level

[6] Each of the models' parameters are bound by zero and one, so we specified a uniform prior with these bounds.

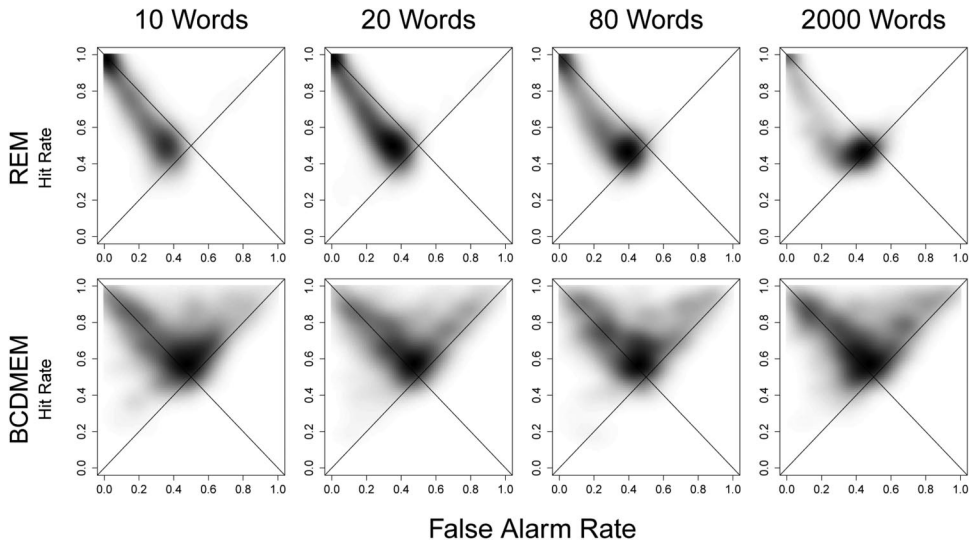

Figure 1. The prior predictive density under four different list lengths for REM (top panels) and BCDMEM (bottom panels): 10 (left column), 20 (middle left column), 80 (middle right column), and 2,000 (right column) items. Darker regions indicate higher density. BCDMEM = bind cue decide model of episodic memory; REM = retrieving effectively from memory model.

parameters, which quantify the effects of the experimental manipulations across all subjects.

Results

We now present the group-level results of the model fitting in three sections. We first discuss the interpretation provided by each model separately, and then, we compare and contrast the two models.

BCDMEM. The two parameters of greatest interest in BCDMEM describe the probability of context reinstatement and the extent of context noise. The probability that context nodes become inactive after longer lists with no filler activity should be greater than for all other experimental conditions because context will be less likely to be reinstated after long lists with no additional filler activity. We denote the probability that nodes become inactive as δ for the long list, no filler activity condition, and as d for the other conditions.

The extent of context noise is dictated by the probability with which nodes become active in the retrieved context. This probability should be lower for low-frequency words than it is for high-frequency words because high-frequency words have been encountered in more contexts than low-frequency words. We will denote the probability of nodes becoming active as τ for the high-frequency words and p for the low-frequency words.

The context reinstatement and context noise parameters for each subject in each experimental condition were drawn from group-level distributions with means ω_δ, ω_d, ω_p, and ω_τ. Figure 3 shows the estimated posterior distributions for the effect of filler activity $\omega_\delta - \omega_d$ (left panel) and the effect of word frequency $\omega_\tau - \omega_p$ (right panel). In the frequentist setting, we would want to reject the null hypotheses that $\omega_\delta - \omega_d \leq 0$ and $\omega_\tau - \omega_p \leq 0$. In the Bayesian setting, we can evaluate explicitly the probability that the alternative hypotheses are true, that is, the probability that $\omega_\delta - \omega_d > 0$ and $\omega_\tau - \omega_p > 0$.

The absence of an experimental effect, a difference of zero, is represented as dotted vertical lines in each panel. The probabilities of the alternative hypotheses are given by the proportion of the experimental effects that are greater than zero. For the filler

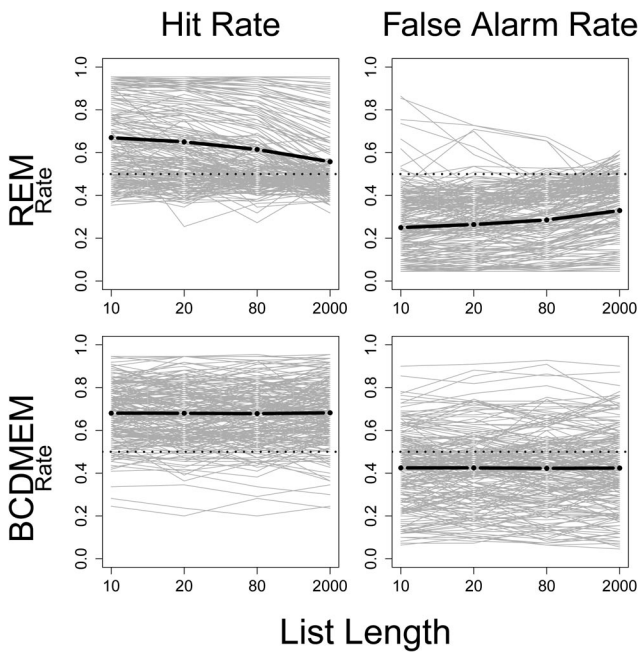

Figure 2. The prior predictive distributions for the hit (left panels) and false-alarm rates (right panels) as a function of list length for REM (top panels) and BCDMEM (bottom panels). The mean of the prior predictive distribution is represented as the black lines. Chance performance is represented as the dotted lines. BCDMEM = bind cue decide model of episodic memory; REM = retrieving effectively from memory model.

activity effect, the probability that $\omega_\delta - \omega_d$ is greater than zero is equal to 0.978. For the word-frequency effect, the probability that $\omega_\tau - \omega_p$ is greater than zero is equal to one. This is evidence that the experimental effects are present and in the direction consistent with the mechanisms proposed by BCDMEM.[7]

REM. As with BCDMEM, the most important parameters in REM are the parameters used to capture the effects of word frequency and filler activity. REM explains the effects of word frequency by a geometric distribution of feature values with a rate parameter *g*. For the geometric distribution, increases of the rate parameter will reduce both the mean and variability of the random variable that represents the feature values. This means that increasing *g* in REM will result in memory traces that contain primarily a few small values (ones and twos, say). Memory traces will then have more features in common as *g* increases.

To capture the differences in the distribution of feature values for high- and low-frequency words, we defined a separate rate parameter for each word-frequency condition. At the group level, we used the parameter ω_γ for high-frequency words and ω_g for low-frequency words. Memory traces of high-frequency words are assumed to have more features in common, and so, high-frequency words have a larger value of *g*. To compare these parameters, we examined the posterior distribution of $\omega_\gamma - \omega_g$, similar to the comparisons we made for BCDMEM.

We chose to model the effects of the filler task by adding a number η of spurious traces to the episodic image for a given subject after the study phase was complete. Adding spurious traces to the episodic image creates more interference in the recognition memory decision, which is akin to the inference induced by the filler activity. We modeled the number of spurious traces with two group-level parameters ω_η and ξ_η and 48 subject-level parameters η. Because η is a subject-level parameter, we can get an idea of what the effects of the filler task would be for an arbitrary subject by simulating values of η from the posterior distributions of the group-level parameters ω_η and ξ_η. We call the distribution of the simulated values $\tilde{\eta}$ the posterior predictive distribution of η.

The left panel of Figure 4 shows the effects of word frequency through the distribution of $\omega_\gamma - \omega_g$, whereas the right panel shows the effects of filler activity through the distribution of $\tilde{\eta}$. The null value of zero is represented by the dotted vertical lines. There is a strong effect for the word-frequency manipulation, and the probability that $\omega_\gamma > \omega_g$ is 1.0. This means that the model accounts for word-frequency effects with memory traces for low-frequency words that are more variable and hence contain more distinctive features than the memory traces for high-frequency words (e.g., Glanzer, Adams, Iverson, & Kim, 1993). The difference in the number of distinctive features results in greater discriminability of low-frequency over high-frequency words.

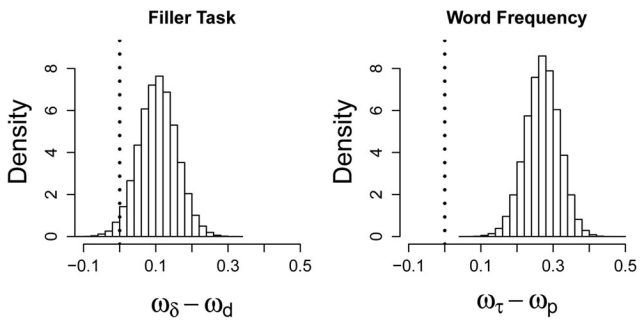

Figure 3. Estimated experimental effects for filler activity (left panel) and the word-frequency manipulation (right panel) at the group level. The null value of zero is represented by the dotted vertical line in both panels.

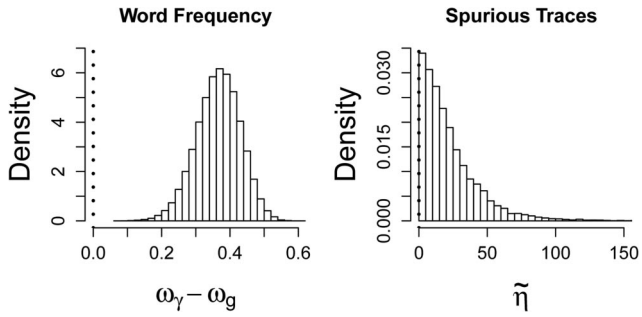

Figure 4. Estimated experimental effects for the word-frequency manipulation (left panel) and at the filler manipulation (right panel). The null value of zero is represented by the dotted vertical line in both panels.

The right panel of Figure 4 shows the distribution of $\tilde{\eta}$. Had there been no effects of the filler activity, the distribution of $\tilde{\eta}$ would be highly concentrated on the null value of zero, represented in the right panel of Figure 4 as the dotted vertical line. However, the mode of the distribution is six, showing some evidence that subjects are sensitive to the effects of filler conditions. Furthermore, this density has long tails, extending out to around 150, indicating that some subjects may be very sensitive to the effects of filler tasks.

Comparing the two models. Fitting the two models to the same data told us two things. First, the effects of the filler activity were strong. For BCDMEM, we came to this conclusion by examining the differences in the parameters ω_δ and ω_d corresponding to the presence and absence of filler activity. The distribution of the difference $\omega_\delta - \omega_d$ had most of its area above zero, suggesting that it was more difficult for subjects to reinstate the study-list context at test in the presence of filler activity. For REM, we drew a similar conclusion from the posterior predictive distribution for the number of spurious traces added to the episodic matrix. Because this distribution also had a great deal of area far from zero, we can conclude that the presence of filler activity created greater distortion in memory, resulting in worse performance.

Second, the effects of word frequency were strong, perhaps much stronger than the effects of filler activity. For both models, the word-frequency manipulation was captured by a comparison between two word-frequency parameters. For both models, at the group level, the posterior probability of a word-frequency effect being greater than zero was 1.0.

[7] An alternative approach to evaluating whether an effect of word frequency is present would have been to contrast the fits of the current model in which ω_δ and ω_d are free to vary against a null model in which $\omega_\delta = \omega_d$. Then, we could compute the Bayes factor, the likelihood of the full model over the likelihood of the null model. If the Bayes factor were larger than one, that would be evidence against the null model and for the presence of a filler effect. However, the filler effect has been demonstrated in a number of other studies (see Dennis et al., 2008), so there is little reason to believe in the null model in the first place.

While fitting the models to the data provides a clearer understanding of how experimental effects are captured by REM and BCDMEM, the primary question we want to answer centers on the list-length manipulation and how well each model captures the effects of list length. To assess model fit, we computed the DIC and the BPIC. For both of these statistics, a better fit is indicated by a smaller value (i.e., more negative). Table 1 shows relevant model fit statistics, including the number of effective model parameters *pD,* the average log likelihood value $\bar{D}$, and the best log likelihood value obtained during sampling $\hat{D}$. The statistics suggest that for these data, BCDMEM provided a better fit.

Summary

In this section, we have shown that the models provided convenient interpretations of the effects in the data in the form of posterior distributions of parameters. Furthermore, the distributions of the estimated parameters were consistent with the theoretical interpretations of the parameters provided by each model.

Through a relative comparison of DIC and BPIC values, we concluded that BCDMEM fit the data of Dennis et al. (2008) better than REM. To generalize these findings, we further investigate the list-length effect by fitting the models to data from Kinnell and Dennis (2012).

Study 2: Data From Kinnell and Dennis (2012)

Kinnell and Dennis (2012) examined the role of different stimulus types on the list-length effect, including pictures of faces, fractals, and photos of scenes. The experimental design was very similar to the design used in Dennis et al. (2008) but included neither word-frequency nor filler task manipulations. Subjects were assigned one stimulus type and then completed two study-list conditions: one short (consisting of 20 items) and one long (consisting of 80 items). Study-list presentation order was counterbalanced across subjects, and the retention intervals were equated across conditions as in Dennis et al. Following the study list, subjects completed a recognition task on 40 items, 20 of which were targets. Forty subjects participated in each of the stimulus conditions.

Kinnell and Dennis (2012) performed conventional (i.e., frequentist) analyses on their data. They estimated d' statistics from the hit and false-alarm rates. For faces, they found that there was a significant list-length effect on d' and the false-alarm rates in both the between- and within-subjects analyses, but there were no significant differences in the hit rates in either analyses. For fractals, the within-subjects analyses revealed a significant list-length effect only on the false-alarm rates, but the between-subjects analyses indicated significant list-length effects on both the false-alarm rates and the d's. For photos of scenes, no list-length effects were detected in any analysis, although they noted that both the mean d' and mean hit rate were higher in the long-list condition.

Table 1
Model Fit Statistics for BCDMEM and REM for the Data of Dennis, Lee, and Kinnell (2008)

Model	DIC	BPIC	*pD*	$\bar{D}$	$\hat{D}$
BCDMEM	−587.44	−509.24	78.21	−665.65	−743.86
REM	−501.68	−439.72	61.96	−563.64	−625.61

Note. BCDMEM = bind cue decide model of episodic memory; BPIC = Bayesian predictive information criterion; DIC = deviance information criterion; REM = retrieving effectively from memory model.

Models

The models we fit to Kinnell and Dennis's (2012) data were very similar in structure to the models we fit to Dennis et al.'s (2008) data. A full description of the models appears in the online supplemental materials.

Results

As in the previous section, we begin by examining the results of the individual model fits and close by comparing and contrasting the fits obtained by both models. To account for the differences in memory performance across the different stimulus materials, we needed to focus on additional model parameters. For BCDMEM, this was the learning rate parameter *r,* which determines the probability that input and context nodes are connected during study. For REM, these were the probabilities associated with copying item features into the memory trace. The probability that a feature is copied is *u,* and the probability that a feature is copied correctly is *c.* For each individual, parameter values were sampled from a hyperdistribution of possible values with mean ω_r, ω_u, and ω_c.

BCDMEM. Figure 5 shows the estimated posterior distributions for each of the mean group-level parameters of the BCDMEM model: contextual reinstatement (ω_d; left panel), contextual noise (ω_p; middle panel), and the learning rate (ω_r; right panel). Each panel contains the estimate of the parameters for each stimulus type, where the solid black, gray, and dotted black lines represent the faces, fractals, and photos of scenes, respectively.

The right panel of Figure 5 shows that the learning rate for fractals takes on smaller values than the rates for both faces and photos of scenes. This learning rate difference suggests that it is harder to learn the features for fractal-type stimuli compared to either faces or photos of scenes. The left panel of Figure 5 shows that the decay parameter takes on larger values for fractals than for either faces or photos of scenes, suggesting poorer contextual reinstatement of fractals at test. The middle panel of Figure 5 shows that the contextual noise for photos of scenes takes on lower values than for either faces or fractals, suggesting that there was more contextual interference for the photos of scenes.

REM. Figure 6 shows the estimated posterior distributions for the group-level mean parameters for the probability of correct feature copying (ω_c; left panel), the feature rate parameter (ω_g; middle panel), and the probability of feature copying (ω_u; right panel) for each of the three stimulus types: faces (solid black lines), fractals (gray lines) and photos of scenes (dotted black lines).

Comparing the left and right panels of Figure 6, we see that the probabilities for copying correctly and feature copying have lower values for the fractals than for either faces or photos of scenes. Taken together, these results suggest that it is more difficult to encode fractals into memory, which leads to poorer performance at test.

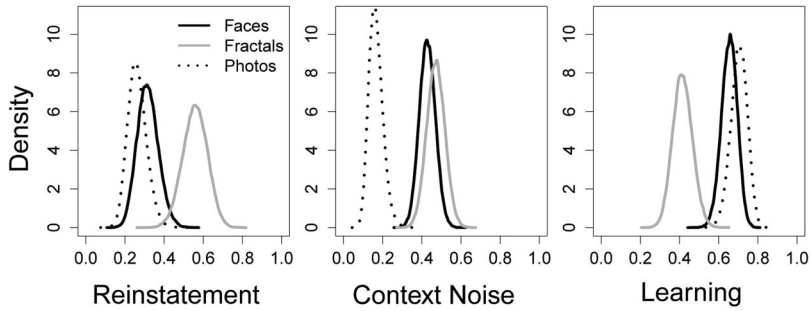

Figure 5. Estimated posterior distributions for each of the three stimuli types for the group-level mean parameters corresponding to contextual reinstatement (left panel), contextual noise (middle panel), and learning rate (right panel). Stimulus types are faces, fractals, and photos of scenes and are represented as solid black, gray, and dotted black lines, respectively.

The middle panel of Figure 6 shows that the values of the feature rate parameter for photos of scenes are smaller than the values for faces and fractals. Recall that a higher feature rate parameter produces less diagnostic feature values, producing smaller memory strength at test, which ultimately leads to lower hit rates and worse recognition performance. Thus, the information perceived in the photos of scenes is more diagnostic of previous exposure than the information in either faces or fractals.

Comparing the models. BCDMEM and REM provide similar interpretations of recognition performance for photos of scenes, faces, and fractals (Figures 5 and 6), despite making different assumptions about how information is stored and retrieved. In particular, the posterior distributions of contextual noise in BCDMEM (ω_p) for each stimulus type closely resemble the estimates for feature discriminability in REM (ω_g). The middle panels of Figures 5 and 6 suggest that the quality of stored information for photos of scenes is appreciably greater than for the other stimulus types. In addition, the left and right panels of Figures 5 and 6 show that the feature copying process assumed by both models is more error prone for fractals than it is for either faces or photos of scenes.

We can also compare the models in terms of how well they fit the data, which is determined by how well each model can accommodate the pattern of list-length effects across the different stimulus conditions. Table 2 shows the DIC (third column) and BPIC (forth column) statistics for each model by stimulus condition, along with the other relevant model fit statistics. The table shows that the DIC and BPIC statistics for BCDMEM are lower than those of REM for all stimulus types, indicating that BCDMEM fit these data best.

Summary

We used hierarchical versions of REM and BCDMEM to fit the data presented in Kinnell and Dennis (2012). We again made use of ABC techniques to obtain estimates of the posterior distributions for all parameters. We showed that the models provided convenient and consistent interpretations of the experimental effects in the form of posterior distributions. Quantitative measures of goodness of fit indicated that BCDMEM provided a better fit to Kinnell and Dennis's data than did REM.

General Discussion

Most memory models like REM and BCDMEM that fit memory data well across a range of domains are simulation based and consequently have not been able to take advantage of Bayesian techniques for model comparison and parameter interpretation. Often, such models do not have a tractable closed-form expression that relates the parameters of the model to the data that were

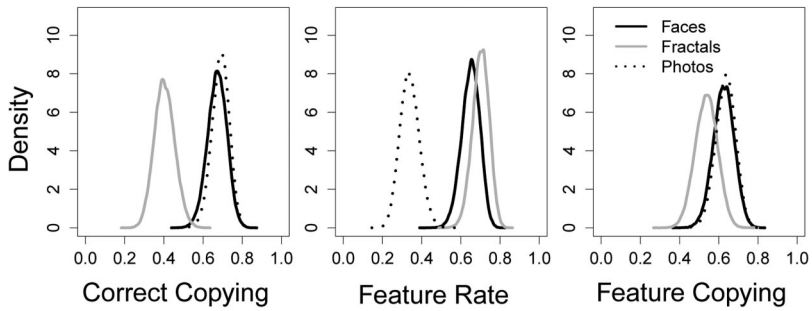

Figure 6. Estimated posterior distributions for each of the three stimulus types for the group-level mean parameters corresponding to the probability of correct feature copying (left panel), the feature rate parameter (middle panel), and the probability of feature copying (right panel). Stimulus types are faces, fractals, and photos of scenes and are represented as solid black, gray, and dotted black lines, respectively.

Table 2
Model Fit Statistics for BCDMEM and REM for the Data of Kinnell and Dennis (2012)

Stimuli	Model	DIC	BPIC	*pD*	$\bar{D}$	$\hat{D}$
Faces	BCDMEM	−249.15	−194.47	54.68	−303.84	−358.52
	REM	−52.80	−7.12	45.67	−98.47	−144.15
Fractals	BCDMEM	−58.43	−8.41	50.02	−108.44	−158.46
	REM	−42.68	4.40	47.08	−89.76	−136.83
Photos of scenes	BCDMEM	−98.92	−62.24	36.68	−135.60	−172.27
	REM	−93.23	−59.42	33.80	−127.03	−160.83

Note. BCDMEM = bind cue decide model of episodic memory; BPIC = Bayesian predictive information criterion; DIC = deviance information criterion; REM = retrieving effectively from memory model.

observed (i.e., a likelihood function). In this note, have we demonstrated how a Bayesian approach to model fitting and testing can improve the interpretation of model parameters. We employed a new technique, ABC, to estimate the posterior distributions of the models' parameters. The ABC technique is important because, without it or something like it, there is no principled way to estimate the parameters of simulation-based models or explore the contributions that these parameters make to different patterns of data.

To test the two models, we fit them to data from two experiments. We chose the list-length data from Dennis et al. (2008) and Kinnell and Dennis (2012) because they used experimental variables that have been shown to be important determinants of the size of the list-length effect (see Dennis et al., 2008; Dennis & Humphreys, 2001). We fit hierarchical versions of both models to the data and found that BCDMEM consistently outperformed REM on both DIC and BPIC metrics. However, the particular paradigm used in these studies has been the subject of some controversy. We make no attempt to resolve this issue here; the model evaluations in this article are meant for illustrative purposes only.

There are three critical lessons to be gained from the exercises in this article. First, ABC allows Bayesian methods to be efficiently extended to simulation-based models. This is a major step forward in modeling technology with important ramifications for cognitive modeling. Second, BCDMEM provides a better explanation for the list-length data of Dennis et al. (2008) and Kinnell and Dennis (2012) than does REM. Given that REM predicts a list-length effect and that not all of the data that we fit show such an effect, this might seem to be a foregone conclusion. However, under some parameter settings, the amount of item interference in REM is small, which can produce a negligible list-length effect. If the models are fit to the data using an approximate least squares method, the advantage for BCDMEM may be similarly small: Both models will seem to fit the data well. However, using a full Bayesian analysis, REM is penalized for its ability to capture both weak and strong list-length effects. The analysis thus exposed a critical distinction between the models to which currently employed modeling technologies are completely insensitive.

One could argue that the versions of BCDMEM and REM that we employed were limited because they did not include parameters for individual item effects (e.g., DeCarlo, 2011; Pratte & Rouder, 2011). Adding such effects could provide a better account of the data and perhaps allow for a more accurate analysis of the particular type of interference present in recognition memory. However, there is no straightforward way to incorporate such effects into either REM or BCDMEM without considerable additional theoretical overhead and increased computational complexity.

The third lesson concerns the role of experimental design on model complexity. Although they are not reported here, some of our preliminary studies showed an interesting interaction between the flexibility of the two models and the experimental design (list length). As the length of the study list increases, the predictions of REM change, whereas the predictions of BCDMEM remain fixed (see Figures 1 and 2). REM's predicted hit rates either stay the same or decrease, while its predicted false-alarm rates tend to stay the same or (mostly) increase as list length increases. Therefore, although REM is better able to predict the list-length effect, it is also capable of producing the null list-length effect. By contrast, BCDMEM makes the same predictions across all study-list lengths and is *only* capable of producing the null list-length effect.

The range of data that REM explains decreases as list length increases, while, for BCDMEM, it remains the same. In other words, the complexity of REM depends on list length, while BCDMEM's does not. Model complexity is often viewed as an intrinsic property of a model independent from any experimental considerations (but see, e.g., Pitt et al., 2002, for a discussion). Although this may seem obvious in hindsight, our Bayesian analysis demonstrated that complexity is a property of the interaction between a model and the experimental design that produced the data to which it is applied. This interactive component of model complexity is naturally incorporated in our hierarchical Bayesian approach.

Conclusion

Both REM and BCDMEM are mechanistic in the sense that they can be programmed and simulated to perform recognition memory tasks. Watkins (1990) argued against the use of such models on the grounds that they are not falsifiable using experimental methods. Instead, he suggested that memory theorists should propose general principles that describe the operation of human memory, and as an example, he promoted the cue overload principle.

It is unclear to us how a general verbal theory like the cue overload principle could be more falsifiable than the constraints provided by a mathematical framework such as those embodied in REM and BCDMEM. General principles must be tested, as are mathematical models, empirically, by collecting data and subjecting those data to statistical tests, which may be Bayesian or may be something else. By quantifying the cue overload principle using

the mechanisms in REM and BCDMEM, ABC gives us the machinery by which the boundary conditions of the cue overload principle can be better appreciated.

The present article has demonstrated the usefulness of the ABC approach. For the first time, we have investigated two models of episodic memory—BCDMEM and REM—in a Bayesian framework. We have presented the first hierarchical fits for both of these models to empirical data, fits that provided posterior estimates of the parameters on every level of the hierarchy. In addition, using our ABC approach, we were able to quantitatively compare the fits of hierarchical versions of both models, and we concluded that, for the data of Dennis et al. (2008) and Kinnell and Dennis (2012), BCDMEM provided a better fit.

References

Ando, T. (2007). Bayesian predictive information criterion for the evaluation of hierarchical Bayesian and empirical Bayes models. *Biometrika, 94,* 443–458. doi:10.1093/biomet/asm017

Batchelder, W. H., & Riefer, D. M. (1990). Multinomial processing models of source monitoring. *Psychological Review, 97,* 548–564. doi:10.1037/0033-295X.97.4.548

Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial processing tree modeling. *Psychonomic Bulletin & Review, 6,* 57–86. doi:10.3758/BF03210812

Craigmile, P., Peruggia, M., & Van Zandt, T. (2010). Hierarchical Bayes models for response time data. *Psychometrika, 75,* 613–632. doi:10.1007/s11336-010-9172-6

Criss, A. H., & McClelland, J. L. (2006). Differentiating the differentiation models: A comparison of the retrieving effectively from memory model (REM) and the subjective likelihood model (SLIM). *Journal of Memory and Language, 55,* 447–460. doi:10.1016/j.jml.2006.06.003

DeCarlo, L. T. (2011). Signal detection theory with item effects. *Journal of Mathematical Psychology, 55,* 229–239. doi:10.1016/j.jmp.2011.01.002

Dennis, S., & Humphreys, M. S. (2001). A context noise model of episodic word recognition. *Psychological Review, 108,* 452–478. doi:10.1037/0033-295X.108.2.452

Dennis, S., Lee, M., & Kinnell, A. (2008). Bayesian analysis of recognition memory: The case of the list-length effect. *Journal of Memory and Language, 59,* 361–376.

Estes, W. K., & Maddox, W. T. (2005). Risks of drawing inferences about cognitive processes from model fits to individual versus average performance. *Psychonomic Bulletin & Review, 12,* 403–408. doi:10.3758/BF03193784

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis*. New York, NY: Chapman & Hall.

Glanzer, M., Adams, J. K., Iverson, G. J., & Kim, K. (1993). The regularities of recognition memory. *Psychological Review, 100,* 546–567. doi:10.1037/0033-295X.100.3.546

Hu, B., & Tsui, K.-W. (2005). *Distributed evolutionary Monte Carlo with applications to Bayesian analysis* (Technical Report No. 1112). Madison: University of Wisconsin—Madison, Department of Statistics.

Kinnell, A., & Dennis, S. (2012). The role of stimulus type in list length effects in recognition memory. *Memory & Cognition, 40,* 311–325. doi:10.3758/s13421-011-0164-2

Klugkist, I., Laudy, O., & Hoijtink, H. (2010). Bayesian evaluation of inequality constrained hypotheses for contingency tables. *Psychological Methods, 15,* 281–299. doi:10.1037/a0020137

Kruschke, J. K. (2011). *Doing Bayesian data analysis: A tutorial with R and BUGS*. Burlington, MA: Academic Press.

Lee, M. D. (2004). A Bayesian analysis of retention functions. *Journal of Mathematical Psychology, 48,* 310–321. doi:10.1016/j.jmp.2004.06.002

Lee, M. D. (2008). Three case studies in the Bayesian analysis of cognitive models. *Psychonomic Bulletin & Review, 15,* 1–15. doi:10.3758/PBR.15.1.1

Lee, M. D. (2011). How cognitive modeling can benefit from hierarchical Bayesian models. *Journal of Mathematical Psychology, 55,* 1–7. doi:10.1016/j.jmp.2010.08.013

Lee, M. D., Fuss, I. G., & Navarro, D. J. (2006). A Bayesian approach to diffusion models of decision-making and response time. In B. Scholkopf, J. Platt, & T. Hoffman (Eds.), *Advances in neural information processing* (Vol. 19, pp. 809–815). Cambridge, MA: MIT Press.

Lee, M. D., & Vanpaemel, W. (2008). Exemplars, prototypes, similarities and rules in category representation: An example of hierarchical Bayesian analysis. *Cognitive Science, 32,* 1403–1424. doi:10.1080/03640210802073697

Lee, M. D., & Wagenmakers, E.-J. (2010). *A course in Bayesian graphical modeling for cognitive science*. Retrieved from http://www.ejwagenmakers.com/BayesCourse/BayesBookWeb.pdf

Liu, C. C., & Aitkin, M. (2008). Bayes factors: Prior sensitivity and model generalizability. *Journal of Mathematical Psychology, 52,* 362–375. doi:10.1016/j.jmp.2008.03.002

Malmberg, K. J., Zeelenberg, R., & Shiffrin, R. (2004). Turning up the noise or turning down the volume? On the nature of the impairment of episodic recognition memory by Midazolam. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 540–549. doi:10.1037/0278-7393.30.2.540

Montenegro, M., Myung, J. I., & Pitt, M. A. (2011). *REM integral expressions*. Unpublished manuscript.

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology, 47,* 90–100. doi:10.1016/S0022-2496(02)00028-7

Myung, J. I., Montenegro, M., & Pitt, M. A. (2007). Analytic expressions for the BCDMEM model of recognition memory. *Journal of Mathematical Psychology, 51,* 198–204. doi:10.1016/j.jmp.2007.02.001

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review, 4,* 79–95. doi:10.3758/BF03210778

Oravecz, Z., Tuerlinckx, F., & Vandekerckhove, J. (2009). A hierarchical Ornstein-Uhlenbeck model for continuous repeated measurement data. *Psychometrika, 74,* 395–418. doi:10.1007/s11336-008-9106-8

O'Reilly, R. C. (2001). Generalization in interactive networks: The benefits of inhibitory competition and Hebbian learning. *Neural Computation, 13,* 1199–1241. doi:10.1162/08997660152002834

O'Reilly, R. C. (2006, October 6). Biologically based computational models of cortical cognition. *Science, 314,* 91–94. doi:10.1126/science.1127242

O'Reilly, R. C., & Munakata, Y. (Eds.). (2000). *Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain*. Cambridge, MA: MIT Press.

Pitt, M. A., Kim, W., Navarro, D. J., & Myung, J. I. (2006). Global model analysis by parameter space partitioning. *Psychological Review, 113,* 57–83. doi:10.1037/0033-295X.113.1.57

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review, 109,* 472–491. doi:10.1037/0033-295X.109.3.472

Pratte, M. S., & Rouder, J. N. (2011). Hierarchical single- and dual-process models of recognition memory. *Journal of Mathematical Psychology, 55,* 36–46. doi:10.1016/j.jmp.2010.08.007

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. doi:10.3758/BF03196750

Rouder, J. N., Lu, J., Speckman, P., Sun, D., & Jiang, Y. (2005). A hierarchical model for estimating response time distributions. *Psychonomic Bulletin & Review, 12,* 195–223. doi:10.3758/BF03257252

Rouder, J. N., Sun, D., Speckman, P., Lu, J., & Zhou, D. (2003). A hierarchical Bayesian statistical framework for response time distributions. *Psychometrika, 68,* 589–606. doi:10.1007/BF02295614

Shadlen, M. N., & Newsome, W. T. (2001). Neural basis of a perceptual decision in the parietal cortex (Area LIP) of the rhesus monkey. *Journal of Neurophysiology, 86,* 1916–1936.

Shiffrin, R. M., Lee, M. D., Kim, W., & Wagenmakers, E.-J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science, 32,* 1248–1284. doi:10.1080/03640210802414826

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—retrieving effectively from memory. *Psychonomic Bulletin & Review, 4,* 145–166. doi:10.3758/BF03209391

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B*. Statistical Methodology, *64,* 583–639. doi:10.1111/1467-9868.00353

Steyvers, M., Lee, M. D., & Wagenmakers, E.-J. (2009). A Bayesian analysis of human decision-making on bandit problems. *Journal of Mathematical Psychology, 53,* 168–179. doi:10.1016/j.jmp.2008.11.002

Storn, R., & Price, K. (1997). Differential evolution: A simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization, 11,* 341–359. doi:10.1023/A:1008202821328

ter Braak, C. J. F. (2006). A Markov chain Monte Carlo version of the genetic algorithm differential evolution: Easy Bayesian computing for real parameter spaces. *Statistics and Computing, 16,* 239–249. doi:10.1007/s11222-006-8769-1

Turner, B. M., & Sederberg, P. B. (2012). Approximate Bayesian computation with differential evolution. *Journal of Mathematical Psychology, 56,* 375–385. doi:10.1016/j.jmp.2012.06.004

Turner, B. M., Sederberg, P. B., Brown, S., & Steyvers, M. (in press). A method for efficiently sampling from distributions with correlated dimensions. *Psychological Methods*.

Turner, B. M., & Van Zandt, T. (2011). *Hierarchical approximate Bayesian computation*. Manuscript submitted for publication.

Turner, B. M., & Van Zandt, T. (2012). A tutorial on approximate Bayesian computation. *Journal of Mathematical Psychology, 56,* 69–85. doi:10.1016/j.jmp.2012.02.005

Usher, M., & McClelland, J. L. (2001). On the time course of perceptual choice: The leaky competing accumulator model. *Psychological Review, 108,* 550–592. doi:10.1037/0033-295X.108.3.550

Vandekerckhove, J., Tuerlinckx, F., & Lee, M. D. (2011). Hierarchical diffusion models for two-choice response time. *Psychological Methods, 16,* 44–62. doi:10.1037/a0021765

Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology, 54,* 491–498. doi:10.1016/j.jmp.2010.07.003

Van Zandt, T. (2000). How to fit a response time distribution. *Psychonomic Bulletin & Review, 7,* 424–465. doi:10.3758/BF03214357

Vrugt, J. A., ter Braak, C. J. F., Diks, C. G. H., Robinson, B. A., Hyman, J. M., & Higdon, D. (2009). Accelerating Markov chain Monte Carlo simulation by differential evolution with self-adaptive randomized subspace sampling. *International Journal of Nonlinear Sciences and Numerical Simulation, 10,* 273–290.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review, 14,* 779–804. doi:10.3758/BF03194105

Watkins, M. J. (1990). Mediationism and the obfuscation of memory. *American Psychologist, 45,* 328–335. doi:10.1037/0003-066X.45.3.328

Wilkinson, R. D. (2011). *Approximate Bayesian computation (ABC) gives exact results under the assumption of model error*. Manuscript submitted for publication.

Received October 18, 2012
Revision received February 11, 2013
Accepted February 13, 2013 ■