Advanced Review

Bayesian statistical approaches to evaluate cognitive models

Jeffrey Annis* and Thomas J. Palmeri

Cognitive models aim to explain complex human behavior in terms of hypothesized mechanisms of the mind. These mechanisms can be formalized in terms of mathematical structures containing parameters that are theoretically meaningful. For example, in the case of perceptual decision making, model parameters might correspond to theoretical constructs like response bias, evidence quality, response caution, and the like. Formal cognitive models go beyond verbal models in that cognitive mechanisms are instantiated in terms of mathematics and they go beyond statistical models in that cognitive model parameters are psychologically interpretable. We explore three key elements used to formally evaluate cognitive models: parameter estimation, model prediction, and model selection. We compare and contrast traditional approaches with Bayesian statistical approaches to performing each of these three elements. Traditional approaches rely on an array of seemingly *ad hoc* techniques, whereas Bayesian statistical approaches rely on a single, principled, internally consistent system. We illustrate the Bayesian statistical approach to evaluating cognitive models using a running example of the Linear Ballistic Accumulator model of decision making (Brown and Heathcote 2008).

How to cite this article:
WIREs Cogn Sci 2017, e1458. doi: 10.1002/wcs.1458

INTRODUCTION

Cognitive models aim to explain behavior by positing mechanisms that underlie perception, memory, decision making, and other fundamental aspects of cognition. Formal cognitive models instantiate these hypothesized mechanisms in terms of mathematics, computations, and simulations, and models are fitted, evaluated, and compared based on tools and techniques from statistics. Formal cognitive models go beyond verbal theories in that they are precisely defined and make explicit predictions. They go beyond statistical models that describe patterns of behavior in that they attempt to explain patterns of behavior in terms of hypothesized mechanisms of the mind. While statistical models make parametric assumptions about observed data, such as linearity or a particular distributional form, cognitive models make further assumptions about the underlying cognitive processes hypothesized to cause observed behavior, allowing differences across stimuli, conditions, groups, or individuals to be quantitatively characterized (e.g., Ref 1).

Such differences are often reflected in the values of free parameters in cognitive models. In the case of decision making, model parameters might describe the quality of evidence driving the decision process, how decisions are made more or less cautiously, and whether decisions are biased or not (e.g., Ref 2). In the case of object categorization, parameters might describe attention weights to relevant or irrelevant features, the relative strength of stored category representations, and biases to chose particular categorization responses over others (e.g., Refs 3,4). And in the case of memory recall, parameters might reflect how well items are stored in memory, how strongly stored items are integrated with current context, and when recall will terminate.[5]

Given a particular cognitive model and a set of observed data, a first step is often *parameter*

*Correspondence to: jeff.annis@vanderbilt.edu

Department of Psychology, Vanderbilt University, Nashville, TN, USA

Conflict of interest: The authors have declared no conflicts of interest for this article.

estimation, fitting a model to observed data, estimating model parameters that maximize the correspondence between model predictions and observed data. Once parameters are estimated, they might be related to brain measures,[6] used to explain individual differences,[7,8] or characterize psychiatric disorders.[9] Best-fitting model parameters can be used to generate *model predictions*. Does a model actually provide an adequate account of observed data and does it generate novel predictions about yet-unseen data?

Often there are multiple competing models of any particular aspect of cognition, and researchers might do *model selection* based on how well or how poorly different models account for observed data. Parameter estimation and model prediction are often repeated for competing models to determine which model best explains the data. 'Best' usually means a model that provides the best fit to data with the lowest model complexity, a formal application of Occam's razor.[10] This model comparison, which leads to model selection, might pit a more general model against a simpler special case of that general model, or pit one model against a different model making different mechanistic assumptions, or pit a mechanistic model against a 'saturated' statistical model to ask whether the mechanistic model leaves any observed variance unexplained.

This article focuses on these three key elements of formally testing a cognitive model: parameter estimation, model prediction, and model selection. To maintain a tight focus, we assume throughout that a researcher has a formal cognitive model already in hand, and is ready to fit that model to observed data and evaluate its quality. There are many interesting and important aspect of cognitive modeling that we will not address, such as a discussion of why formal cognitive models are developed and how they are used to advance theory (e.g., Refs 11–13), how cognitive models are initially developed or extended (e.g., Ref 14), a survey different kinds of cognitive models (e.g., Refs 1,15) and how they have been applied to particular aspects of cognition (e.g., Refs 2,16,17), or how cognitive models can be applied to neural data (e.g., Ref 18). This article first describes how parameter estimation, model prediction, and model selection have been traditionally carried out in the cognitive modeling literature.[1,11] The bulk of the article describes how *Bayesian statistics* can provide an alternative, coherent, and principled approach to these elements of modeling.

To be clear, Bayesian principles have made inroads into cognitive science and cognitive modeling in two different ways: One way is assuming that the mind and brain are inherently Bayesian, that human learning and cognition follow the principles of Bayesian probabilistic inference (e.g., Refs 19–21 but see Ref 22). Such Bayes-in-the-head models are often referred to as *Bayesian cognitive models*. Instead, here we are using Bayesian statistics not as a principle to explain cognition, but as a tool to evaluate models of cognition. Bayesian statistics can be used to evaluate cognitive models, whether those cognitive models are themselves Bayesian or not. In fact, while there are indeed many successful Bayesian cognitive models, the majority of cognitive models are non-Bayesian, in the sense that learning and cognition are not governed by Bayesian probabilistic inference. We can test quantitatively the adequacy of non-Bayesian models using Bayesian statistics.

The ideal intended audience for this article is someone who is familiar with cognitive models and how those models are traditionally fitted and evaluated, and wants to understand how Bayesian statistics might be used as a tool to do that fitting and evaluating. This could be because they want to use Bayesian statistics in their own work or merely want to better understand how other researchers are using Bayesian statistics in published articles. Some familiarity with Bayes rule would be beneficial, but we have tried to craft this article without requiring much background knowledge of Bayesian statistics. Because this article reflects the intersection of cognitive modeling and Bayesian statistics, we recommend readers new to both of these topics to read some of the more introductory textbooks and articles we reference (e.g., Refs 1,11) for cognitive modeling, and (e.g., Refs 23–25) for Bayesian statistics.

TRADITIONAL APPROACHES TO EVALUATE COGNITIVE MODELS

Traditionally, parameter estimation, model prediction, and model selection in cognitive modeling have been carried out using various optimization and statistical techniques, each with their own advantages and disadvantages.[1,11]

Estimating best-fitting parameters of a cognitive model involves finding the values of model parameters that minimize or maximize some objective function measuring how much the model predictions mismatch the observed data. The earliest published modeling work often minimized sum of squared error or root mean squared error (RMSE), or maximized correlation (r) between model and data or percent variance accounted for by the model; while these objective functions may be fine for linear statistical and mechanistic models, they are often inappropriate

for nonlinear models, which is a property of many cognitive models.

A better approach, used in both nonlinear statistical modeling and cognitive modeling, is maximum likelihood estimation (MLE[26];). Formally, and abstractly, the goal is to find the parameter vector, $\boldsymbol{\theta}$, that maximizes the likelihood of the data, $\boldsymbol{D}$, given the parameters, or maximizing $p(\boldsymbol{D}|\boldsymbol{\theta})$; practically speaking, you will often see work minimizing the negative of the log likelihood (minimizing the negative only because many optimization routines are set up to minimize by default, and taking the log because calculating likelihoods often involve numerous multiplications of small numbers that would lead to numeric underflow without taking logs). Of course, MLE assumes that this likelihood exists or can be approximated, which we will assume to be the case throughout the rest of this article.

Numerous methods exist for optimizing parameters given some objective function. For the majority of cognitive models, direct application of calculus for optimization is so unwieldy as to render it largely impractical. For instance, a model with n parameters requires n partial derivatives, a solution to a system of n (often nonlinear) equations, and a series of tests to rule out local minima. Use of optimization algorithms that require the first or second derivatives can be somewhat less cumbersome, but they quickly become tedious to implement, computationally expensive, and error-prone as dimensionality increases (but see Ref 27).

So instead, optimization is often done using techniques such as hill-climbing, such as the well-known Simplex method;[28] Simplex is the default optimization algorithm in Matlab and Python when derivative are unspecified or unavailable. Given a starting point for the hill climb (often many different starting points are used), maximizing likelihood or minimizing RMSE yields a point estimate for the best-fitting parameter vector. This omits any measure of uncertainty we might want to know in the parameter estimate; for example, if we obtain a maximum likelihood estimate of 1.03 for a given parameter, is there any chance that the parameter could be 1.04 or 1.02 instead? To obtain estimates of uncertainty, other procedure must be carried out such as parametric or nonparametric bootstrap sampling,[29] which gives an estimate of a confidence interval around a maximum likelihood point estimate.

Best-fitting parameters generated by MLE or other methods are often then used to generate model predictions. Especially for models with stochastic elements, this may involve simulating the model hundreds or thousands of times, aggregating the predictions in some way, and then comparing predictions against the observed data. Typical of a traditional approach, and a potential disadvantage as well, is that these predictions are based on point estimates alone, ignoring any uncertainty of the parameter estimates.

Models are compared quantitatively according to how well they fit the observed data. But a mere comparison of quantitative fit, declaring the 'winner' as the model with the largest likelihood or with the smallest RMSE, is nonsensical because doing so does not take into account the relative complexity or flexibility of the competing models. A more general model is mathematically guaranteed to fit at least as well, if not better, than a special case of that general model. The question is not whether the special case fits worse—it generally will—but whether it fits significantly worse than the more general model. Also in the case where one model is not a special case of another model, a model with more free parameters or with more flexibility in terms of the range of predictions it can make will undoubtedly fit better than a model with fewer parameters and greater restrictions. The question is not whether the more complex model fits better—it often will—but whether it fits better even when that model is appropriately penalized for its greater complexity. Traditional methods that penalize likelihood measures of fit based on complexity include the Bayesian Information Criterion (BIC;[30]) and the Akaike Information Criterion (AIC;[31]). While these methods are easy to implement—they simply involve adding a penalty term based on the number of free parameters—they largely ignore other important aspects of model complexity such as the functional form of the model and the size of space of possible predictions (e.g.,[10]).

INTRODUCING THE BAYESIAN STATISTICAL APPROACH

In this article, we outline a cognitive modeling approach to parameter estimation, model prediction, and model selection that uses *Bayesian Statistics*. The Bayesian approach answers some of the limitations of the traditional approaches outlined above. Interest in Bayes has exploded over the past decade or more. As noted earlier, we do not discuss ideas of how human cognition might be based on Bayesian principles.[19,20,32] We will also not discuss general aspects of Bayesian data analysis,[23,33] although there are certainly parallels to what we discuss here. Rather, we will discuss how the Bayesian approach can be applied to cognitive models in order to perform

parameter estimation, model prediction, and model selection in a manner that is arguably both logically consistent and principled.[34] Therefore, this article is aimed at readers who are familiar with cognitive modeling, but are less familiar with how the Bayesian statistical approach can be applied to parameter estimation, prediction, or model selection in the context of cognitive modeling.

To state the obvious, Bayesian analysis is based on Bayes' rule. For example, we can use Bayes' rule to compute the full posterior probability distribution of the parameters given the data, $p(\boldsymbol{\theta}|\boldsymbol{D})$:

$$p(\boldsymbol{\theta}|\boldsymbol{D}) = \frac{p(\boldsymbol{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{D})} \quad (1)$$

The posterior probability of the parameters given the data is the product of the likelihood, $p(\boldsymbol{D}|\boldsymbol{\theta})$, and the prior probability of the parameters $p(\boldsymbol{\theta})$, normalized by the probability of the data, $p(\boldsymbol{D})$. The likelihood of the data given the parameters is defined by the cognitive model in exactly the same way it would be specified for MLE. The prior distribution reflects a subjective belief about parameter values before observing the data.[35] These beliefs could be vague in the form of flat or relatively flat priors, they could be based on previous data used to estimate the parameters, or they could be based on values of the parameters that are meaningful by some theoretical, objective, or subjective criterion. Because Bayes' rule yields a posterior joint probability distribution of parameter values, it provides naturally a measure of parameter uncertainty, and hence an elegant solution to model prediction that takes into account that uncertainty. The Bayesian approach also allows a means for model comparison and model selection that takes into account model complexity in a natural, coherent, and comprehensive manner.

While traditional techniques of parameter estimation require a search of parameter space to minimize or maximize some objective function, Bayes' rule tells us directly the probability of the parameter values for a model given the observed data—at least in principle. As we will see, despite its apparent simplicity, even for models with only modest complexity, Bayes' rule cannot be solved analytically and requires computational estimation. One reason for the relatively recent explosion of interest in an idea first suggested in the 1700s,[36] and characterized and formalized mathematically many decades ago, has been the joint development of computational techniques and the availability of powerful computer hardware to make Bayesian analysis tractable. As such, interest in the Bayesian approach to evaluating cognitive models has grown significantly over the past few years. Some applications include evaluating variants of signal detection theory (e.g., Refs 37,38), multinomial processing trees (e.g., Ref 39), individual differences (e.g., Refs 7,40–42), decision making (e.g., Refs 43,44), multidimensional scaling (e.g., Refs 37,42), choice response time (e.g., Refs 45–48), memory (e.g., Refs 49–52), and joint modeling of neural and behavioral data (e.g., Refs 6,53).

What follows is a combination of a review and tutorial of Bayesian approaches to evaluating cognitive models, with an added aim of pointing the reader to emerging new developments. We structure the remainder of the article around the three key elements of cognitive modeling outlined earlier—parameter estimation, model prediction, and model selection. For each, we describe the underlying concepts and mathematics, followed by the computational techniques used in practice. Throughout, we use a cognitive model called the *Linear Ballistic Accumulator* model (LBA;[54]) as a running example. We chose the LBA because it is a general model of choice response time that can be applied to a wide variety of tasks; we also recently published a companion paper[45] that outlines how to implement the LBA in a Bayesian statistical language called Stan.[27]

There are many excellent reviews and textbooks on Bayesian approaches to evaluating cognitive models and Bayesian statistics more generally. See Lee and Wagenmakers[55] for a practical introduction to Bayesian approaches to cognitive modeling, Rouder and Lu[38] for a more mathematically rich tutorial, Lee[37,56] for illustrative examples of how the Bayesian framework can be applied to cognitive models, and Shiffrin et al.[57] for a tutorial on Bayesian model selection. For textbooks on Bayesian statistics in general, see, for example, Kruschke[23] for the uninitiated and Gelman et al.[25] for the more heroic; Jackman[24] and Lynch[58] are texts specifically geared toward social scientists.

Example: The LBA Model

LBA belongs to a class of models known as *sequential sampling models*[2] that describes an evidence accumulation process of decision making over time and can be used to predict both response probabilities and response times. LBA assumes that after a stimulus is presented, its representation is perceptually encoded and compared to some form of knowledge that will drive a decision. The time to complete this encoding process is given by the parameter τ (this parameter also includes motor execution time to

produce an overt response). After encoding, evidence begins to accrue in independent *accumulators* that each correspond to one response alternative, *i*. Figure 1 shows an example of LBA with two accumulators, one for Response A and one for Response B. The rate at which evidence accumulates for response *i* is given by its corresponding drift rate, d_i. Drift rates are assumed to vary across trials, and are sampled from a normal distribution with mean v_i and standard deviation *s*. The starting point of evidence accumulation is also assumed to vary over trials, sampled from a uniform distribution, U(0, *a*), where *a* is the maximum (note that in some articles that use LBA, the maximum is denoted by an uppercase *A*; here we chose to instead use the lowercase *a* to ensure no confusion with the response alternative A). The evidence accumulation terminates and a response is made when the first accumulator reaches its threshold $a + k$, where *k* is the *relative threshold*.

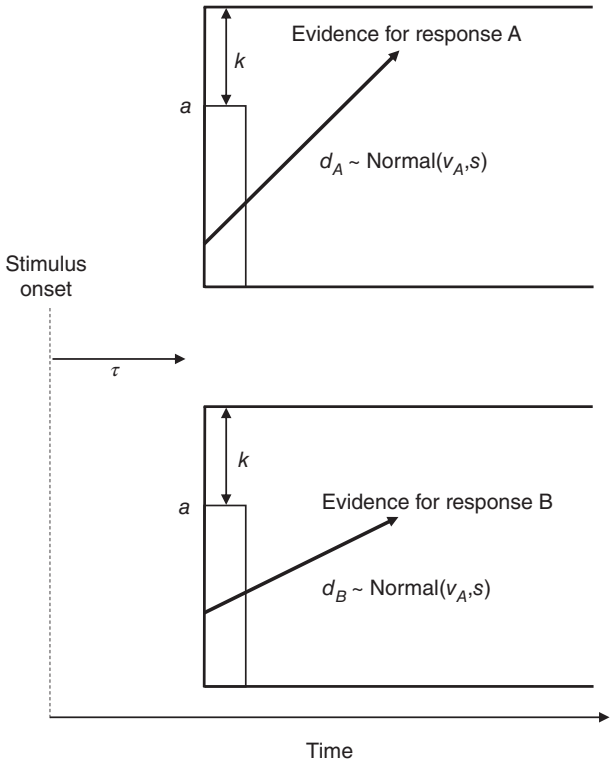


FIGURE 1 | The Linear Ballistic Accumulator (LBA) model is an example of a formal cognitive model that predicts response probabilities and distributions of response times. LBA can be used to decompose response time and accuracy into core cognitive parameters: evidence accumulation, response caution, and perceptual encoding. LBA assumes that after the stimulus is perceptually encoded after time τ, evidence toward each response alternative, *i*, accumulates with drift rate d_i. Drift rates across trials are sampled from normal distributions with mean v_i and standard deviation *s*. In our examples, we constrain mean drift rate for the Response B accumulator to be 1 minus mean drift rate for the Response A accumulator. The starting point of the evidence accumulation process on each is sampled from a uniform distribution between 0 and *a*. The response is determined by the first accumulator to reach threshold $a + k$.

Compared to other accumulator models that assume noisy accumulation or lateral inhibition (e.g., Ref 59), LBA assumes a linear rise to threshold, which significantly simplifies its mathematical formulation. Brown and Heathcote[54] derived the likelihood function for the LBA. If $\boldsymbol{D^A}$ is a vector of observed (Data) response times for Response A, and $\boldsymbol{D^B}$ is a vector of observed (Data) response times for Response B, the likelihood for the combined observed data vector, $\boldsymbol{D}$, is the product of the two likelihoods:

$$p(\boldsymbol{D}|\boldsymbol{\theta}) = p\left(\boldsymbol{D^A}|\boldsymbol{\theta}\right)p\left(\boldsymbol{D^B}|\boldsymbol{\theta}\right). \quad (2)$$

For the full mathematical description of the likelihoods we refer the reader to Brown and Heathcote.[54]

Bayesian Parameter Estimation

One key difference between traditional and Bayesian approaches, is that Bayesian statistics treats data as well as unknown parameters as random variables. This allows us to write a joint distribution of the data, $\boldsymbol{D}$, and the parameter(s), $\boldsymbol{\theta}$:

$$p(\boldsymbol{D},\boldsymbol{\theta}). \quad (3)$$

We can re-express the joint distribution via the definition of conditional probability:

$$p(\boldsymbol{D},\boldsymbol{\theta}) = p(\boldsymbol{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\boldsymbol{D})p(\boldsymbol{D}). \quad (4)$$

Rearranging, we obtain Bayes' rule:

$$p(\boldsymbol{\theta}|\boldsymbol{D}) = \frac{p(\boldsymbol{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{D})} = \frac{p(\boldsymbol{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\boldsymbol{D}|\boldsymbol{\theta'})p(\boldsymbol{\theta'})d\boldsymbol{\theta'}}, \quad (5)$$

where $p(\boldsymbol{\theta}|\boldsymbol{D})$ is the posterior distribution, $p(\boldsymbol{D}|\boldsymbol{\theta})$ is the likelihood function (the same as that found in MLE), $p(\boldsymbol{\theta})$ is the prior distribution, and $p(\boldsymbol{D})$ is a normalization constant, which ensures the posterior integrates to 1, and is referred to as the *marginal likelihood*, or sometimes as the *evidence*. Here, the parameter vector $\boldsymbol{\theta'}$ has a prime within the integral to make clear that $\boldsymbol{\theta'}$ is different from $\boldsymbol{\theta}$; for simplicity, from now on, we will drop the superscript on parameters appearing within the integral.

Note that this form of Bayes' rule can be thought of as being implicitly conditioned on a given model. For example, $p(\boldsymbol{\theta}|\boldsymbol{D})$ under the LBA model will be different than $p(\boldsymbol{\theta}|\boldsymbol{D})$ under the diffusion model. To make the model we are working with explicit, sometimes Bayes' rule will be written including a model notation ($\mathcal{M}$) explicitly:

$$p(\boldsymbol{\theta}|\boldsymbol{D},\mathcal{M}) = \frac{p(\boldsymbol{D}|\boldsymbol{\theta},\mathcal{M})p(\boldsymbol{\theta}|\mathcal{M})}{p(\boldsymbol{D}|\mathcal{M})}. \quad (6)$$

This more complete formulation will be important later when we discuss model selection. For now, we will omit the explicit model notation.

Unlike traditional parameter estimation approaches that result in point estimates, the Bayesian approach results in a full posterior probability distribution of the parameters. Summary statistics like the mean, mode,[a] and standard deviation of a parameter can be computed, as can the correlations between parameters to detect potential parameter trade-offs. One statistic often used to summarize the amount of uncertainty in a Bayesian parameter estimate is called the *95% Highest Density Interval* (HDI); this is the shortest interval that contains 95% of the mass of the posterior distribution.[60] Smaller HDIs indicate a posterior distribution with concentrated mass over a smaller amount of parameter space (and thus higher probability) than larger HDIs. Although many loosely equate the HDI with confidence intervals in traditional methods,[61,62] Bayesian HDIs are not the same thing.[53,55]

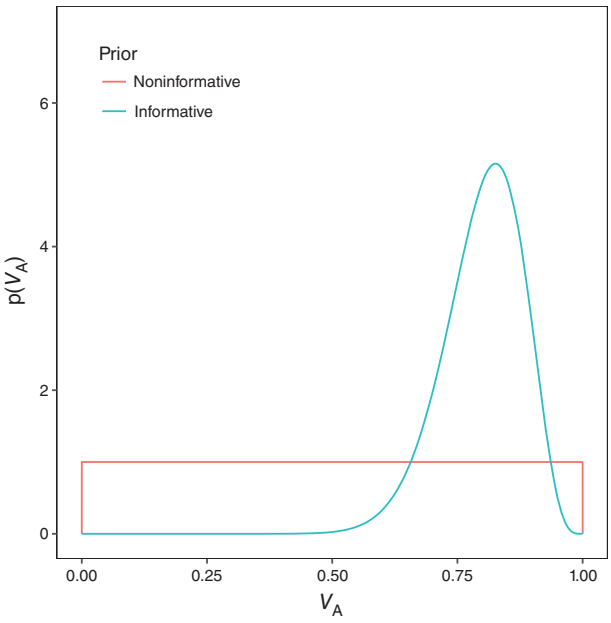


FIGURE 2 | Prior distributions represent our subjective *a priori* beliefs about parameter values. This figure illustrates two different prior distributions chosen for v_A in the LBA model, which in one common instantiation is constrained to fall between 0 and 1. Both priors are beta distributions with different shapes (controlled by the rate parameter α and the shape parameter β of the beta). When $\alpha = \beta$, the beta distribution is mathematically equivalent to the uniform distribution, depicted in red. This prior is *noninformative* because it represents the belief that all values of v_A are equally likely. Depicted in blue is a beta distribution with $\alpha = 5$ and $\beta = 20$ and represents the prior belief that relatively large values of v_A are more likely than relatively small values. This prior is *informative* because we have concentrated a large amount of mass over a relatively small range of parameter values.

As noted earlier, in addition to specifying the likelihood of the data given the parameters, $p(\boldsymbol{D}|\boldsymbol{\theta})$, the Bayesian approach also requires us to specify priors, $p(\boldsymbol{\theta})$. Broadly speaking, there are two types of priors: *informative* and *noninformative*. An informative prior represents a strong *a priori* belief about the parameter values. These prior beliefs might represent expert knowledge about the parameters, previous fits of the model to other data, biological or other constraints on possible parameter values, or theoretical information regarding allowable parameter values. A noninformative prior represents a similar degree of belief across all possible parameter values (or a very wide range of parameter values). For example, if a parameter can only exist within a bounded range (not including positive or negative infinity), a useful noninformative prior might be a uniform distribution over that bounded range.

How to set priors is hotly debated in the Bayesian literature (e.g., Ref 63). For some, any informative prior is seen as too subjective, and has been part of a general critique of Bayesian statistics.[64] For others, an informative prior is viewed as an integral part of the model, forcing the theorist to formalize assumptions about model parameters.[65,66] In the case of cognitive psychology, we often have too little prior information to set informative priors in a way that would be universally accepted, so we often use noninformative priors. There is a large body of research devoted to developing noninformative priors which can be used as (at least arguably) a reasonable default.[67–69] Figure 2 shows examples of informative and noninformative priors based on the beta distributions.

The last part of Bayes' rule is its denominator (Eq. (6)), $p(\boldsymbol{D})$. This marginal likelihood is obtained by integrating the product of the likelihood and prior over the entire parameter space. When, as usually is the case for a cognitive model, $\boldsymbol{\theta}$ is a vector of parameters, this will be a multivariate integral. For example, the simplest version of the LBA posterior distribution (fixing s for identifiability) expands to:

$$p(v_1,v_2,a,\tau,b|\boldsymbol{D}) = \frac{p(\boldsymbol{D}|v_1,v_2,a,\tau,k)p(v_1,v_2,a,\tau,k)}{\int_{v_1}^{b}\int_{v_2}^{b}\int_{a}^{b}\int_{\tau}^{b}\int_{b}^{b} p(\boldsymbol{D}|v_1,v_2,a,\tau,k)p(v_1,v_2,a,\tau,k)dv_1\,dv_2\,da\,d\tau\,dk}. \quad (7)$$

This is LBA applied to a single subject and a single experimental condition. Multiple conditions manipulating difficulty, bias, and speed–accuracy trade-off will significantly increase the number of model parameters, and hence significantly increase the complexity of the integral. Fitting a hierarchical model with many subjects simultaneously will further increase the number of model parameters and the complexity of the integral by orders of magnitude.

Multivariate integrals of complex nonlinear functions (such as likelihoods defined by cognitive models multiplied by the priors on parameters) nearly always defy any closed form solution. While standard numeric integration techniques such quadrature exist, they can only be applied to numeric integrals over a handful of variables. These techniques are impractical for models with dozens or hundreds of parameters, which are often the case with hierarchical cognitive models; in fact it is not hard to reach a level of complexity that would require longer than the lifetime of the universe to solve the resulting multivariate integral using standard numerical methods.

Thankfully, modern computer hardware permits the use of *Monte Carlo techniques* (e.g., Ref 70). These allow one to draw random samples, $\boldsymbol{\theta}_i$, directly from the posterior, $p(\boldsymbol{\theta}|\boldsymbol{D})$, without having to explicitly solve an intractable integral; from these samples, summaries, and inferences about the underlying parameter distribution are possible.

Computing the Posterior: Markov Chain Monte Carlo

Markov Chain Monte Carlo (MCMC) methods[55,70–72] can efficiently sample from high-dimensional probability density functions. Using random numbers to solve (hard) problems in generally are referred to as *Monte Carlo* techniques. With a traditional random number generator—think rand() or randn() in Matlab—sequential random samples (random numbers) are statistically independent of one another.[b] In MCMC, sequential samples are not statistically independent, but depend on the previous sample. Such dependency makes the process *Markov*. The sequence of such random numbers forms the chain in *Markov Chain*.

Perhaps the first MCMC method was the *Metropolis algorithm*.[73] In its simplest form, for the case of a single parameter, it begins by picking a random initial value of the chain, θ_0. Each step of the chain represents the next (potential) random sample from the probability density function. On each step i of the chain, a proposed random sample, θ_*, is generated by adding random noise (often from a zero-centered normal distribution), ϵ_i, to the previous random sample, θ_{i-1}. The proposed sample, θ_*, is always accepted if it has a higher probability density than that of the previous sample, θ_{i-1}. If the proposed sample has a lower probability density than the previous sample, then the proposed sampled is accepted probabilistically, with an acceptance probability equal to the ratio of the probability density of the proposed sample versus the probability density of the previous sample.

The Metropolis algorithm is completely generic and can be applied to any probability density function. For the case of sampling from posterior probabilities in a Bayesian analysis, we can formalize the probability of acceptance of the next random sample in the chain as:

$$p(\text{accept}) = \min\left(1, \frac{p(\theta_*|\boldsymbol{D})}{p(\theta_{i-1}|\boldsymbol{D})}\right). \quad (8)$$

If the sample is accepted then $\theta_i = \theta_*$, otherwise $\theta_i = \theta_{i-1}$. The chain of θ_i values represents random samples drawn from $p(\theta|\boldsymbol{D})$. Those random samples from the posterior can be used to calculate quantities like the mean, MAP, or HDI of parameter θ. As noted earlier, being a chain of samples, one random number is not independent of the previous random number, unlike standard, non-MCMC, random number generators; such autocorrelation is not always an issue in practice, but techniques like *thinning*, whereby only every 10th or 50th or 100th sample in the chain are kept as true samples, are sometimes used.

At first blush, it appears as if we have done nothing to make the problem any more tractable. After all, calculating the posterior distributions still requires calculating an integral in the denominator of Bayes' rule. But note that the denominator is the same whether calculating $p(\theta_*|\boldsymbol{D})$ or calculating $p(\theta_{i-1}|\boldsymbol{D})$. Those two denominators cancel each other in the ratio. Therefore, the acceptance formula simplifies to:

$$p(\text{accept}) = \min\left(1, \frac{p(\boldsymbol{D}|\theta_*)p(\theta_*)}{p(\boldsymbol{D}|\theta_{i-1})p(\theta_{i-1})}\right), \quad (9)$$

with the integrals eliminated entirely. This is form of the Metropolis acceptance ratio most commonly seen

in the literature. Here we have illustrated the algorithm assuming a single model parameter, but the method can be extended to posteriors on many parameters. This makes an intractable problem tractable, albeit at the computational cost of calculating long chains of sampled random numbers (or sampled vectors of random numbers in the case of multidimensional posteriors).

The Metropolis algorithm was later generalized to arbitrary proposal distributions, including those that are asymmetric, in the *Metropolis–Hastings algorithm*;[74] asymmetric proposal distributions are more efficient, for example, in the case of parameters defined within bounded regions. MCMC was further extended with the development of the *Gibbs sampler*,[75,76] a special case of Metropolis–Hastings where the acceptance probability is always 1, where proposals are drawn from the full conditional distributions for each parameter one at a time.[70] In MCMC, it is valid to take Gibbs steps on some parameters and Metropolis steps on others as in the *Metropolis-within-Gibbs Sampler*.[77] While some modelers program their own MCMC algorithms by hand, there are a variety of software toolboxes that function as black box inference engines performing 'automatic Bayesian inference' via built-in Metropolis–Hastings, Gibbs, and other samplers. These include *WinBUGS*,[78] *JAGS*,[79] and *Stan*[27] with its advanced MCMC algorithm based on *Hamiltonian Monte Carlo*.[80] These toolboxes only require the user to specify the model (statistical or cognitive) in a probabilistic programming language and then let an automated inference engine generate samples from the posterior distribution.

These toolkits have many built-in probability distributions that make programming Bayesian statistical models fairly straightforward. However, cognitive models, such as LBA, require specialized likelihood functions that are not pre-packaged with any toolbox. For many, if not most, cognitive models, it is necessary to implement custom probability distributions. WinBUGS and JAGS allow this, but require relatively low-level programming in C++[81]; Stan allows this within the same Stan programming language directly.[45]

LBA Example

Here we illustrate component-wise Metropolis with the LBA model, in a simple example assuming data from a single subject in a single condition. The MCMC chain begins by initializing model parameters: $v_A^0, v_B^0, s^0, \tau^0, b^0$, with a fixed at 1 for identifiability.[82] On each step i of a chain of length N, $i = \{1,\ldots, N\}$, we do a Metropolis step for each parameter of the model in random order. For each parameter, random noise (here normally distributed with mean zero and a small standard deviation of .05) is added to the previous sample to produce a proposed sample. The variance of the proposal distribution was found after some experimentation by running the chain for a short time and observing whether the chain was efficiently exploring the parameter space.

To give a concrete example of how the algorithm works, suppose we update v_A first:

$$v_A^* = v_A^{i-1} + \epsilon_{v_A}^i \tag{10}$$

The probability of accepting v_A^* is given by:

$$p(accept) = \min\left(1, \frac{p(D|v_A^*, \boldsymbol{\theta}')p(v_A^*)}{p(D|v_A^{i-1}, \boldsymbol{\theta}')p(v_A^{i-1})}\right), \tag{11}$$

where $\boldsymbol{\theta}'$ contains all the most recently updated parameters except for v_A (not shown in the prior calculation because these terms cancel out due to an assumption of independent priors for different parameters); in this case, $\boldsymbol{\theta}' = (v_B^0, s^0, \tau^0, b^0)$. If the proposal is accepted, we set $v_A^i = v_A^*$, otherwise $v_A^i = v_A^{i-1}$. We repeat for all parameters of the model to constitute a single MCMC step. We continue sampling, creating an MCMC of length N or until the algorithm converges on the full posterior distribution by some criterion (for best practices on diagnosing MCMC chain convergence, e.g., Ref 25).

Figure 3 shows an example of an MCMC chain and the resulting samples for two of the five LBA parameters; while we plot only two variables, all five parameters are sampled. Panel (a) shows that the chain stepped toward a higher value of τ and a slightly lower value of v_A. Panel (b) shows the distribution of samples that were obtained after letting the chain run for 3000 iterations and discarding the first 100 samples as *burn-in*; 'burn-in' represents samples at the beginning of the MCMC chain that may not be representative of the posterior distribution because they are sampled from an extremely low density region and simply reflect the initialization point for a chain. Panel (c) shows the individual chain for v_A over the course of the iterations and these samples are plotted as a histogram in Panel (d); panels (e) and (f) show the same for τ.

These Bayesian approaches require a mathematically specified likelihood function, $p(D|\boldsymbol{\theta})$, for the cognitive model. Unfortunately, predictions from many interesting and important models in cognitive psychology are based on computer simulation, they

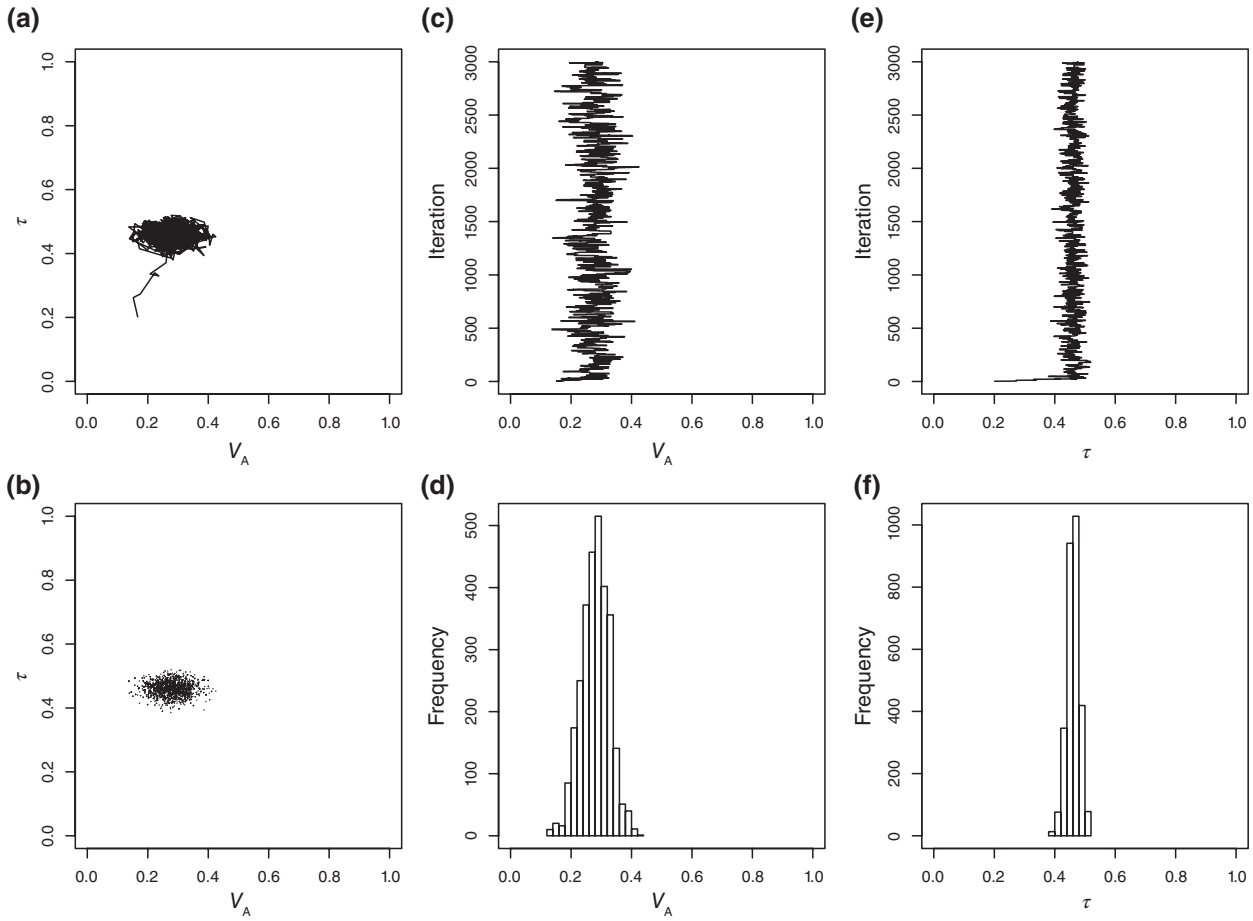


FIGURE 3 | (a) The path of the Markov chain for τ and v_A. The chain begins in low density region around $\tau = .2$ and $v_A = .2$ and quickly moves to a higher density region as per the Metropolis acceptance probability ratio. (b) (below (a)) The resulting samples drawn from the joint posterior distribution of τ and v_A, excluding the first 100 samples as burn-in. (c) The path the chain took over the marginal distribution for v_A at each iteration of the algorithm. The resulting marginal distribution is plotted below in (d). The path of the chain over the marginal distribution of τ and the resulting samples are shown in (e) and (f), respectively.

are not defined by an explicit closed-form likelihood function. Traditionally, such models are fitted by simulating the model thousands of times for a given set of parameters, computing the discrepancy between model predictions and observed data, adjusting the parameters to minimize or maximize some objective function by hill-climbing or some other optimization technique. Fortunately, new methods are being proposed to allow Bayesian approaches to be applied to simulation-based cognitive models without any explicitly specified likelihood function (e.g., Refs 83–87).

BAYESIAN MODEL PREDICTION

Once a cognitive model's parameters have been estimated, a common next step is to generate model predictions and compare those qualitatively and quantitatively with observed data. In a traditional modeling approach, one obtains a point estimate of the parameters, $\hat{\boldsymbol{\theta}}$, that maximize likelihood or minimize RMSE, and then uses that point estimate to generate model predictions.

Bayesian prediction is different because Bayesian analysis produces a full joint posterior distribution of parameter values, $p(\boldsymbol{\theta}|\boldsymbol{D})$. So cognitive model predictions should be based on that full parameter distribution, not a point estimate. The probability of a prediction, $\hat{\boldsymbol{D}}$, is conditionalized on the model parameters, $\boldsymbol{\theta}$, which in turn are conditionalized on the observed data, $\boldsymbol{D}$. By the law of total probability, we can characterize model prediction as:

$$p\left(\hat{\boldsymbol{D}}|\boldsymbol{D}\right) = \int p\left(\hat{\boldsymbol{D}}|\boldsymbol{\theta}\right) p(\boldsymbol{\theta}|\boldsymbol{D}) d\boldsymbol{\theta}. \qquad (12)$$

This is known as the *posterior predictive distribution*. Consider an extreme case where the posterior, $p(\boldsymbol{\theta}|$

$\boldsymbol{D}$), had all of its mass at a single point rather than a distribution; after observing the data, only a single vector of parameters, $\hat{\boldsymbol{\theta}}$, is possible. In that case, the posterior predictive would reduce to the likelihood, $p\left(\hat{\boldsymbol{D}}|\hat{\boldsymbol{\theta}}\right)$. For all other cases, the posterior predictive distribution is effectively a weighted average of the likelihood, $p\left(\hat{\boldsymbol{D}}|\boldsymbol{\theta}\right)$, with weights determined by the posterior, $p(\boldsymbol{\theta}|\boldsymbol{D})$.

COMPUTING THE POSTERIOR PREDICTIVE

On the surface, calculating the posterior predictive in Eq. (13) looks daunting. After all, it requires solving a multivariate integral, which could well be composed of hundreds or thousands of variables for a complex hierarchical model. But it turns out this is a fairly straightforward because a commonly used Monte Carlo integration technique can be applied. Since this technique might not be familiar to all readers, we provide a brief introduction.

Consider first the definition of expected value:

$$E[\boldsymbol{x}] = \int \boldsymbol{x} p(\boldsymbol{x}) d\boldsymbol{x} \approx \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i, \qquad (13)$$

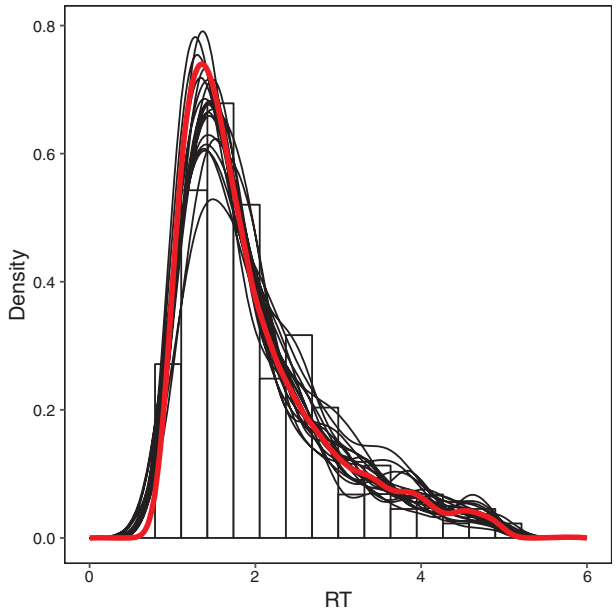


FIGURE 4 | Shows the predictive distribution for each of several posterior samples (black lines) and the overall posterior predictive distribution (red line) plotted against the response time distribution simulated from the LBA (histogram bars).

where $p(\boldsymbol{x})$ is the probability of $\boldsymbol{x}$. This can be generalized to the expected value of a function g applied to $\boldsymbol{x}$ as:

$$E[g(\boldsymbol{x})] = \int g(\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x} \approx \frac{1}{N}\sum_{i=1}^{N} g(\boldsymbol{x}_i). \qquad (14)$$

The *expected value* is, as the name suggests, the long-run, expected theoretical mean for random variables $\boldsymbol{x}$ having probability $p(\boldsymbol{x})$, whereas the *average* is the empirical mean of observed data points $\boldsymbol{x}_i$ that have been sampled from $p(\boldsymbol{x})$.

Monte Carlo integration turns these formulae on their heads. Imagine instead of trying to estimate the theoretical mean or the empirical mean per the definition of expected value, you are trying to solve an integral. Suppose you need to solve an integral that has the functional form of Eqs. (13) or (14), where $p(\boldsymbol{x})$ happens to be in the form of a probability density function from which samples can be generated. This integral can be approximated by drawing samples $\boldsymbol{x}_i$ from $p(\boldsymbol{x})$ and averaging all the $\boldsymbol{x}_i$ (Eq. (13)) or averaging all the $g(\boldsymbol{x}_i)$ (Eq. (14)). By this Monte Carlo method, integrals that could be difficult or impossible to solve analytically are approximated computationally by sampling lots of random numbers.

In the present case,

$$\int p\left(\hat{\boldsymbol{D}}|\boldsymbol{\theta}\right) p(\boldsymbol{\theta}|\boldsymbol{D}) d\boldsymbol{\theta} \approx \frac{1}{N}\sum_{i=1}^{N} p\left(\hat{\boldsymbol{D}}|\boldsymbol{\theta}_i\right) \qquad (15)$$

where $\boldsymbol{\theta}_i$ is randomly sampled from $p(\boldsymbol{\theta}|\boldsymbol{D})$. This is a form of Eq. (14). How do we generate those random samples? Well, $p(\boldsymbol{\theta}|\boldsymbol{D})$ is produced by the MCMC procedures described in the previous section, which in fact instantiates a random number generator for $p(\boldsymbol{\theta}|\boldsymbol{D})$.

To generate predictions, we can simply simulate data from the model given each posterior sample, $\boldsymbol{\theta}_i$, of the Markov chain and take the average. Figure 4 shows the histogram of response time data to which the LBA model was fit. Black lines plot a subset of the distributions, $p\left(\hat{\boldsymbol{D}}|\boldsymbol{\theta}_i\right)$, and the red line plots the posterior predictive distribution, $p\left(\hat{\boldsymbol{D}}|\boldsymbol{D}\right)$. The variability across $p\left(\hat{\boldsymbol{D}}|\boldsymbol{\theta}_i\right)$ is reflective of the uncertainty associated with the parameters in the model. The posterior predictive distribution takes into account all of this uncertainty by averaging over all possible $p\left(\hat{\boldsymbol{D}}|\boldsymbol{\theta}_i\right)$. The posterior predictive distribution can

then be plotted over the relative frequency distribution given by the observed data to perform a *posterior predictive check*. If the predictive distribution aligns with the empirical distribution, then we can make a qualitative judgment as to whether the model adequately accounts for the data.

It is often the case that there are visual discrepancies between the observed data and the model predictions. Sometimes these discrepancies may be large enough to warrant rejection of the model. What is 'large enough'? Sometimes it may be visually obvious when to reject the model and start over and other times it is not. One way to quantify these discrepancies is to construct a so-called discrepancy function between the observed and predicted data, denoted $T\left(\hat{D},D\right)$. The discrepancy function measures how different the predicted values are from the observed. Using the same discrepancy function, one then measures the difference between the predictions, $\hat{D}$, and replications of the predictions $\hat{D}_{rep}$ with $T\left(\hat{D},\hat{D}_{rep}\right)$. The probability that $T\left(\hat{D},\hat{D}_{rep}\right)$ is less than $T\left(\hat{D},D\right)$ is referred to as the Bayesian p value.[88,89] The outcome of the Bayesian p value heavily relies on the choice of the discrepancy function, which is largely an arbitrary choice.[90] While this is problematic for the Bayesian p value, it might serve as a more objective alternative to a visual posterior predictive check.

BAYESIAN MODEL SELECTION

Often, there are several competing cognitive models that the researcher would like to contrast, picking the model that 'best' explains the observed data. In a sense, while in parameter estimation we are interested in the probabilities of parameter values given data, in model selection we are interested in probabilities of models given data. In both the traditional and Bayesian approaches, to determine the 'best' fitting model, a trade-off must be made between overall goodness-of-fit and model complexity.[91]

As noted earlier, one traditional approach to model selection among non-nested models involves computing the maximum likelihood and then penalizing the model based on its number of parameters. The AIC[31] and the BIC[30] are both based on this type of rule. The BIC is given by:

$$BIC = -2\ln p\left(\boldsymbol{D}|\hat{\boldsymbol{\theta}}\right) + k\ln(n), \tag{16}$$

where k is the number of parameters in the model and n is the number of data points (the AIC penalty term is similarly additive, but only involves $2k$). These approaches are computationally simple to implement, but they ignore both parameter uncertainty and the functional forms of the models (e.g., Ref 10).

By contrast, the Bayesian framework promises to provide a principled—if computationally more challenging—approach to select among competing models while taking into account both parameter uncertainty and the functional form of the model. Bayesian model selection can be thought of as weighing the evidence provided by the data in favor of alternative models.[67] In Bayesian terms, we are interested in the probability of model k, $\mathcal{M}_k$, given data, $\boldsymbol{D}$. That probability can be found by a simple application of Bayes' rule:

$$p(\mathcal{M}_k|\boldsymbol{D}) = \frac{p(\boldsymbol{D}|\mathcal{M}_k)p(\mathcal{M}_k)}{\sum_{j=1}^{M} p(\boldsymbol{D}|\mathcal{M}_j)p(\mathcal{M}_j)}. \tag{17}$$

where $p(\mathcal{M}_k)$ is the prior probability of model k, $p(D|\mathcal{M}_k)$ is the *marginal likelihood* for model k, and the sum in the denominator is a normalizing constant over all M possible models under consideration.

In the case of comparing two models, $\mathcal{M}_1$ and $\mathcal{M}_2$, we consider the ratio of the posterior probabilities:

$$\frac{p(\mathcal{M}_1|\boldsymbol{D})}{p(\mathcal{M}_2|\boldsymbol{D})} = \frac{p(\boldsymbol{D}|\mathcal{M}_1)p(\mathcal{M}_1)}{\sum_{j=1}^{M} p(\boldsymbol{D}|\mathcal{M}_j)p(\mathcal{M}_j)} \Big/ \frac{p(\boldsymbol{D}|\mathcal{M}_2)p(\mathcal{M}_2)}{\sum_{j=1}^{M} p(\boldsymbol{D}|\mathcal{M}_j)p(\mathcal{M}_j)}. \tag{18}$$

The normalizing constants in the denominators drop out and we can rewrite the posterior odds as:

$$\frac{p(\mathcal{M}_1|\boldsymbol{D})}{p(\mathcal{M}_2|\boldsymbol{D})} = \frac{p(\boldsymbol{D}|\mathcal{M}_1)p(\mathcal{M}_1)}{p(\boldsymbol{D}|\mathcal{M}_2)p(\mathcal{M}_2)}. \tag{19}$$

The transformation from prior to posterior odds is determined by the ratio of the marginal likelihoods for each model. This transformation is the weight of the evidence provided by the data and is called the *Bayes Factor*,[10,92–94] denoted B_{12}:

$$B_{12} = \frac{p(\boldsymbol{D}|\mathcal{M}_1)}{p(\boldsymbol{D}|\mathcal{M}_2)}. \tag{20}$$

Note that the Bayes factor does not depend on the prior odds of the models. Arguably, this is convenient because there might be disagreement among theorists

as to the prior probability of the alternative models. If it is principled, one could set the prior odds to give each model equal footing.[95–97] In that case, the $p(\mathcal{M}_k)$ terms in Eq. (19) cancel out, and the posterior odds equal the Bayes factor. As a rule of thumb, a B_{12} greater than 3 is generally considered to be positive evidence for Model 1, while a B_{12} greater than 10 is generally considered to be strong evidence for Model 1 (e.g., Ref 93). Of course, reciprocals of these Bayes factors provide corresponding levels of evidence for Model 2; indeed, one of the great strengths of Bayesian statistics in general is that it lets us evaluate both sides a comparison, not just one.

While simple to write out, how do we calculate $p(\boldsymbol{D}|\mathcal{M}_k)$? Recall from earlier that we can rewrite Bayes' rule for parameter estimation in a form that makes the assumed model, $\mathcal{M}$, explicit:

$$p(\boldsymbol{\theta}|\boldsymbol{D},\mathcal{M}) = \frac{p(\boldsymbol{D}|\boldsymbol{\theta},\mathcal{M})p(\boldsymbol{\theta}|\mathcal{M})}{p(\boldsymbol{D}|\mathcal{M})}. \tag{21}$$

Now it is the marginal likelihood in the denominator, $p(\boldsymbol{D}|\mathcal{M})$, that is the focus of attention. But recall that we needed the development of MCMC techniques because the marginal likelihood is nearly always impossible to solve analytically and very difficult to estimate computationally. The marginal likelihoods, $p(\boldsymbol{D}|\mathcal{M}_k)$, integrate over the entire parameter space. While computationally challenging to estimate, the marginal likelihood takes into account both uncertainty associated with the parameters and the functional form of the model, providing a natural and principled penalty for model complexity.

Computing the Bayes Factor

As previously shown, the marginal likelihood is obtained by integrating over the parameters:

$$p(\boldsymbol{D}|\mathcal{M}) = \int p(\boldsymbol{D}|\boldsymbol{\theta},\mathcal{M})p(\boldsymbol{\theta}|\mathcal{M})d\boldsymbol{\theta}. \tag{22}$$

So the computation of the Bayes factor can be written as:

$$B_{12} = \frac{\int p(\boldsymbol{D}|\boldsymbol{\theta},\mathcal{M}_1)p(\boldsymbol{\theta}|\mathcal{M}_1)d\boldsymbol{\theta}}{\int p(\boldsymbol{D}|\boldsymbol{\theta},\mathcal{M}_2)p(\boldsymbol{\theta}|\mathcal{M}_2)d\boldsymbol{\theta}} \tag{23}$$

These integrals are intractable for complex models with many parameters; standard numerical methods for solving integrals may work for models with relatively few parameters, but are not scalable to complex nonlinear models with many parameters. What about the Monte Carlo method we discussed earlier for Bayesian prediction? Well, if we sample random numbers from the prior, it is *theoretically* possible to estimate marginal likelihoods thus:

$$\int p(D|\theta)p(\theta)d\theta \approx \frac{1}{N}\sum_{i=1}^{N} p(D|\theta_i), \tag{24}$$

where θ_i represent random samples from the prior. Unfortunately, estimating marginal likelihood this way turns out to be so highly inefficient as to be rendered impractical; this is in part because priors are often relatively flat, covering jointly a huge expanse of possible parameter space.[93]

That said, recent advancements in specialized computer hardware, referred to as graphical processing units (GPUs), have begun to make this task more feasible for some models.[98] GPUs were originally designed for the efficient control of computer graphics. Unlike CPUs, which process a single instruction at a time, GPUs implement massively parallel architectures and are now applied to range of scientific computing problems (e.g., Ref 99). The Monte Carlo marginal likelihood estimator in Eq. (24) is a perfect example of such a problem, in which the sampling procedure and likelihood computation can both be parallelized on a GPU. Evans and Brown,[98] using the LBA as an example, showed that a GPU can produce a marginal likelihood estimate containing 100,000,000 samples in mere minutes. For a typical CPU (around 2 GHz, *circa* 2017), that same computation might take a couple of days. Although the GPU method is promising, it is unknown whether it reasonably scales with increases in the dimensionality of the model. Evans and Brown showed the marginal likelihood estimates were highly variable for moderately sized models with multiple subjects. The GPU method also requires specialized hardware, which might be prohibitively expensive for some. However, there are other methods available that provide ways of estimating the Bayes factor with standard CPUs (for a tutorial, see Ref 100).

Nested Model Comparison

Model comparison can be broadly classified into *nested model comparison* and *non-nested model comparison*. In nested model comparison, the two models being compared have the same parameters, but restriction is placed on the parameters of one of the models. For example, we might want to compare a model in which a particular parameter θ is equal to some specific value θ_1 (e.g., in a model where θ acts additively, the specific value might be $\theta_1 = 0$, and in

a model where θ acts multiplicatively, the specific value might be $\theta_1 = 1$) versus a model in which the parameter θ is allowed to vary freely. For this nested model situation, we can estimate the Bayes factor by simply computing the ratio of the density of the posterior and prior at θ_1. This is known as the *Savage–Dickey Ratio Test.*[101]

Consider two models, one nested within the other, where a restriction is placed on one of the parameters in one of the models, $\theta = \theta_1$. We refer to the restricted model as $\mathcal{M}_R$ and the full model as $\mathcal{M}_F$. It can be shown (e.g., Ref 102) that the Bayes factor is the ratio of the posterior density at θ_1 and the prior density at θ_1:

$$B_{\mathrm{RF}} = \frac{p(\boldsymbol{D}|\mathcal{M}_R)}{p(\boldsymbol{D}|\mathcal{M}_F)} = \frac{p(\theta = \theta_1|D)}{p(\theta = \theta_1)}. \quad (25)$$

This entails that the Bayes factor can be computed by sampling from the posterior as is done via MCMC and determining the density at the desired point, here θ_1. Values of B_{RF} greater than 1 indicate evidence for the restricted model, while values less than 1 indicates evidence for the full model, with strength of evidence given by the size of those ratios.

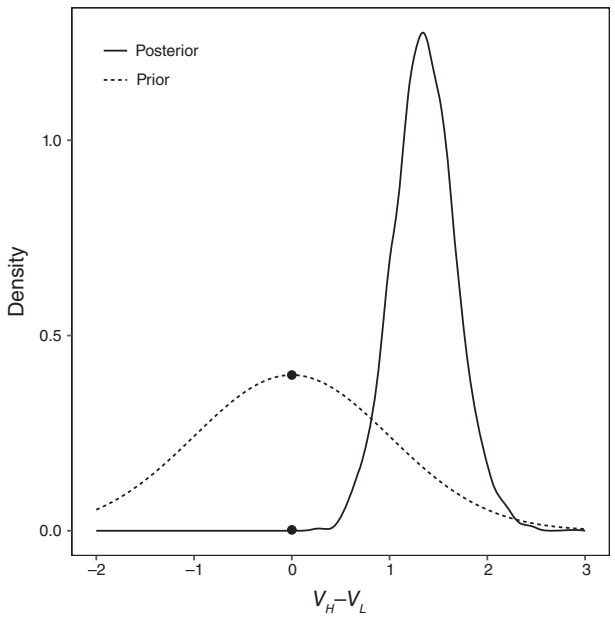


FIGURE 5 | Graphical depiction of the Savage-Dickey ratio test. The dotted line is the prior placed on the effect size and solid line is posterior. The black dots represent the height of the prior and posterior when the effect size is 0. The ratio of these heights is the Bayes factor, the weight of the evidence. The height of the posterior at zero is 178 times less than the prior at zero, indicating the data have decreased our belief in the effect size being zero by a factor of 178.

As an example, consider a perceptual decision task with two conditions in which the discriminability of the target stimulus was manipulated. The hypothesis is that drift rate should vary across conditions. Let ν_{L} and ν_{H} denote the drift rate for correct responses for the condition in which discriminability is low and high, respectively. To test the hypothesis, we must determine whether the difference between ν_{L} and ν_{H} is equal to zero. One way to test this via the Savage–Dickey ratio would be to determine the following quantity:

$$B_{RF} = \frac{p(\nu_H - \nu_L = 0|D)}{p(\nu_H - \nu_L = 0)}. \quad (26)$$

Values greater than 1 would indicate evidence for the restricted model in which ν_{L} equals ν_{H}, while evidence less than 1 indicates evidence for the full model in which ν_{L} and ν_{H} are free to vary. Visually, the numerator is the height of the marginal posterior distribution $p(\nu_H - \nu_L = 0|D)$ when ν_{L} equals ν_{H} and the denominator is the height of the prior when ν_{L} equals ν_{H}. Figure 5 shows a graphical example of what a Savage–Dickey test might look like after obtaining MCMC samples. The dotted line shows the density of the prior distribution $p(\nu_H - \nu_L)$. The black points show the height of the posterior and prior when $\nu_H - \nu_L = 0$. The ratio of these heights is equal to B_{RF}. In this example, the height of the marginal posterior at zero is $p(\nu_H - \nu_L = 0|D) \approx .002$. For the prior, the density is roughly $p(\nu_H - \nu_L) = .40$. Thus, $B_{\mathrm{RF}} \approx \frac{.002}{.40} = .005$. This indicates that the data are roughly 200 times (i.e., 1/.005) more likely under the full model than they are under the restricted model. There are several packages in R and Python that can compute the height of the estimated density from MCMC samples. In this example, we use a nonparametric density estimator suggested by Wagenmakers et al.[101]

Non-nested Model Comparison

With non-nested model comparison, the models are completely different with different parameters, so computational short-cuts like Savage–Dickey are not available. For example, the LBA and the Leaky Competing Accumulator model[59] are both used to model choice response times, but make different assumptions and have different sets of parameters. These two models are therefore non-nested and if we wanted to compare their performance using the Bayes factor we would not be able to use the Savage–Dickey test.

However, there are some general computational methods that have been devised to compute the marginal likelihoods for any model (for a review, see Ref 103). Most of these methods are quite complex and we will not detail them here, and only provide pointers for interested readers. One class of methods computes an estimate of the marginal likelihoods via Monte Carlo sampling. These include importance sampling,[104] reciprocal importance sampling,[105] annealed importance sampling,[106] bridge sampling,[107] Chib's method,[108,109] nested sampling,[110] and thermodynamic integration.[111,112] The thermodynamic approach has received significant attention in fields like biology[112] and ecology (e.g., Ref 113) in part because it is a general method that can be applied to any model with little modification to existing model code.

Another class of methods is called *transdimensional MCMC*, in which the competing models are placed within one 'supermodel.' On each step of the algorithm, a model index variable indicates one of the two models. The ratio of the proportion of times each model is visited equals the Bayes factor. Examples of transdimensional MCMC algorithms include *reversible-jump MCMC*[114] and the *product space method*.[115,116]

Lastly, there are information criterion approaches that are similar to BIC and AIC, but takes into account the uncertainty in the parameter estimates by considering the entirety of the MCMC sample. Examples of such information criteria include the *Bayesian Predictive Information Criterion* ([117]), the *Widely Applicable Bayesian Information Criterion* (WBIC;[118,119]), the *Widely Applicable Information Criterion* (WAIC;[120,121]), and the *Deviance Information Criterion* (DIC;[122]).[c]

Sensitivity to the Prior

Although the Bayes factor has gained significant traction in model selection in psychology,[10,57,94,123] and has been applied to model selection problems in a variety of domains (e.g., Refs 41,124,125), one potentially contentious issue is the Bayes factor's sensitivity to the priors on $\boldsymbol{\theta}$.[93,96,97] This sensitivity contrasts with Bayesian parameter estimation, where any influence of priors is largely overwhelmed by the likelihood given a sufficient amount of data.

In Bayesian model selection, the average predictive performance of the model over the entire parameter space is assessed. Any prior that assigns low weight to high likelihood areas or high weight to low likelihood areas will penalize the model. And informative priors that end up in line with the likelihood will reward a model more so than vague priors or one with informative priors that are not in line with the likelihood. Unlike parameter estimation, effects of the prior are not significantly diminished by having large amounts of data. Some argue that subjective prior belief about parameters should not significantly affect model selection.

In response, others have argued that priors placed on model parameters are a vital component of the theory, and therefore model selection should be sensitive to the priors.[65,97,126] Theoretical development could reflect different theoretical assumptions of the priors on $\boldsymbol{\theta}$ and model selection could be performed over these different instantiations. For example, Vanpaemel and Lee[66] formalized different assumptions about optimal dimensional attention weight parameters in the generalized context model of categorization using different priors (GCM[127];); the Bayes factor was then used select among the alternatives. The Bayes factor's sensitivity to the prior can be seen as advantageous when testing different theoretical assumptions.

Other times, we do not have strong theoretical assumptions that we can instantiate in priors. For example, if we are just beginning to develop a new model, we usually do not know what sort of parameter values we should expect, or do not have a theory of how those parameters might be set. In these cases, we are more concerned about the robustness of the inference made under a particular prior. To test the robustness of inference, we can conduct a *sensitivity analysis*[10,93,128] in which the Bayes factor is computed over a range of different priors. If the Bayes factor is qualitatively consistent across different prior settings, then we know that our inferences are robust under different assumptions about the prior.

CONCLUSION AND SUMMARY

Formal cognitive models describe psychological mechanisms in terms of mathematical structures. We contrasted traditional and Bayesian approaches to cognitive modeling, focusing on issues of parameter estimation, model prediction, and model selection. The goal of parameter estimation is to determine those parameters which provide the best fit of a cognitive model to the data. While traditional approaches, such as maximum likelihood, treat parameters as point estimates, the Bayesian approach quantifies uncertainty about the parameter estimates in terms of complete probability distributions. The goal of model prediction is to predict new data based on what has been observed. While traditional approaches use point estimate values to predict new

data, the Bayesian approach takes into account the uncertainty in the entire posterior distribution to generate model predictions. Lastly, the goal of model selection is to select the cognitive model which best explains the data. Traditional approaches use information theoretic measures, such as AIC and BIC, that do not take into account parameter uncertainty and function form. The Bayesian approach, based on the Bayes factor, takes into account the uncertainty over the entire parameter space and balances complexity versus fit to accomplish model selection.

An important topic we did not have space to discuss is Bayesian hierarchical approaches. The Bayesian approaches we discussed here could be applied to individual participants *i*, separately obtaining posterior distributions $\boldsymbol{\theta}_i|\boldsymbol{D}$ for each participant. Although this approach is useful for estimating parameters at the individual level, we are often interested in both individual-level and group-level performance. We can model both via a conceptually straightforward extension to the Bayesian approach that treats unknown parameters at the group level as random variables as well as those that describe each individual participant (e.g.,[25]).

By simultaneously estimating both group and individual-level parameters, such hierarchical Bayesian methods largely solve the problem of aggregation,[129] which has been a key issue in cognitive modeling for decades (e.g., Ref 130). When fitting a model to data, often data will be aggregated over trials or subjects depending on whether the interest lies in group-level or individual-level conclusions. However, the conclusions that one might draw from a model fitted to aggregated data must be drawn carefully because models can behave differently when fit to group and individual data.[8,131] Bayesian hierarchical methods have provided a solution to problems associated with aggregation in a wide range of different areas including recognition memory,[52] multidimensional scaling,[132,133] and category learning.[7] In addition, hierarchical Bayesian approaches offer various avenues for cognitive modeling including modeling multiple tasks within a single model, assigning subjects to latent classes, and modeling individual differences (for reviews, see Refs 8,56,129).

While Bayesian approaches to cognitive modeling clearly provide advantages over many traditional approaches, one thing that should be clear from this review is that Bayesian approaches carry the cost of being more computationally intensive. We discussed how MCMC algorithms have allowed Bayesian models with hundreds or thousands of parameters—especially in the case of hierarchical models—to be fitted to data. The Bayesian approach provides a coherent way to update beliefs in light of data and offers an extremely flexible framework to fit individual- and group-level parameters not only in theory but also in practice.

NOTES

[a] The mode of the posterior is often called the *maximum a posteriori* (MAP) estimate in Bayesian analysis.

[b] Technically, of course, the pseudo-random number generators used in nearly all programming environments are based on a completely deterministic algorithm; they produce a sequence of random samples that cannot be distinguished statistically from those produced by a true random process.

[c] While the WBIC requires samples from the posterior raised to the power of $1/\ln(n)$, where n is the number of data points, the WAIC and DIC only require samples from the posterior and are, therefore, easily computed. The DIC is known to have problems penalizing the model for complexity, as it is known to sometimes yield a negative estimate for the number of effective parameters in the model. The WAIC does not suffer from this and has many other advantages.[120] The WAIC is given by:

$$\mathrm{WAIC} = \sum_{i=1}^{n} \ln\left(\frac{1}{S}\sum_{s=1}^{S} p(\boldsymbol{D}_i|\boldsymbol{\theta}_s)\right) - \sum_{i=1}^{n} V_{s=1}^{S}(\ln p(\boldsymbol{D}_i|\boldsymbol{\theta}_s)),$$

where n is the number of data points, S is the number of posterior samples, and $V_{s=1}^{S}(\ln p(\boldsymbol{D}_i|\boldsymbol{\theta}_s))$ is the sample variance of the log likelihood of the data point, $\boldsymbol{D}_i$, under all the posterior samples. The first term acts as the goodness-of-fit measure, while the second term penalizes the model for its complexity.

ACKNOWLEDGMENTS

This work was supported by NSF SBE-1257098, NSF SBE-1640681, NEI R01 EY021833, the Temporal Dynamics of Learning Center (NSF SMA-1041755), the Vanderbilt Vision Research Center (NEI P30-EY008126), a Discovery Grant from Vanderbilt University, and a training grant from the NEI (T32-EY07135). The authors thank Brandon Turner and Gabriel Tillman for their helpful comments and criticisms.

REFERENCES

1. Lewandowsky S, Farrel S. *Computational Modeling in Cognition: Principles and Practice*. Thousand Oaks, CA: Sage Publications; 2011.
2. Ratcliff R, Smith PL. A comparison of sequential sampling models for two-choice reaction time. *Psychol Rev* 2004, 111:333–367.
3. Nosofsky RM. Exemplar-based approach to relating categorization, identification, and recognition. In: Ashby G, ed. *Multidimensional Models of Perception and Cognition*. Hillsdale, NJ: Lawrence Erlbaum; 1992, 363–394.
4. Richler JJ, Palmeri TJ. Visual category learning. *WIREs Cogn Sci* 2014, 5:75–94.
5. Polyn SM, Norman KA, Kahana MJ. A context maintenance and retrieval model of organizational processes in free recall. *Psychol Rev* 2009, 116:129–156.
6. Turner BM, Forstmann BU, Love BC, Palmeri TJ. Approaches to analysis in model-based cognitive neuroscience. *J Math Psychol* 2017, 76:65–79 https://doi.org/10.1016/j.jmp.2016.01.001.
7. Bartlema A, Lee MD, Wetzels R, Vanpaemel W. A Bayesian hierarchical mixture approach to individual differences: case studies in selective attention and representation in category learning. *J Math Psychol* 2014, 59:132–150.
8. Shen J, Palmeri T. Modelling individual difference in visual categorization. *Vis Cogn* 2016, 24:260–283 https://doi.org/10.1080/13506285.2016.1236053.
9. Wiecki TV, Poland J, Frank MJ. Model-based cognitive neuroscience approaches to computational psychiatry: Clustering and classification. *Clin Psychol Sci* 2015, 3:378–399.
10. Myung IJ, Pitt MA. Applying Occam's razor in modeling cognition: a Bayesian approach. *Psychon Bull Rev* 1997, 4:79–95.
11. Busemeyer JR, Diederich A. *Cognitive Modeling*. Thousand Oaks, CA: Sage Publications; 2010.
12. Newell A. *You Can't Play 20 Questions with Nature and Win : Projective Comments on the Papers of this Symposium*. New York: Academic Press; 1973.
13. Allport DA. Critical notice: The state of cognitive psychology. *Q J Exp Psychol* 1975, 27:141–152.
14. Shiffrin RM, Nobel PA. The art of model development and testing. *Behav Res Methods Instrum Comput* 1997, 29:6–14.
15. Polk TA, Seifert CM. *Cognitive Modeling*. Cambridge, MA: MIT Press; 2002.
16. Clark SE, Gronlund SD. Global matching models of recognition memory: how the models match the data. *Psychon Bull Rev* 1996, 3:37–60 https://doi.org/10.3758/BF03210740.
17. Pothos EM, Wills AJ. *Formal Approaches to Categorization*. Cambridge, UK: Cambridge University Press; 2011.
18. Palmeri TJ, Love BC, Turner BM. Model-based cognitive neuroscience. *J Math Psychol*. 2017, 76:59–64 https://doi.org/10.1016/j.jmp.2016.10.010.
19. Jacobs RA, Kruschke JK. Bayesian learning theory applied to human cognition. *WIREs Cogn Sci* 2011, 2:8–21.
20. Griffiths T, Kemp C, Tenenbaum J. Bayesian models of cognition. In: Sun R, ed. *The Cambridge Handbook of Expertise and Expert Performance*. New York: Cambridge University Press; 2008, 59–100.
21. Anderson JR. *The Adaptive Character of Thought*. Hillsdale NJ: Lawrence Erlbaum; 1990.
22. Jones M, Love BC. Bayesian fundamentalism or enlightenment ? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behav Brain Sci* 2011, 34:169–231.
23. Kruschke JK. *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. 2nd ed. Burlington, MA: Academic Press; 2014.
24. Jackman S. *Bayesian Analysis for the Social Sciences*. John Wiley & Sons; 2009.
25. Gelman A, Carlin JB, Stern HS, Rubin DB. *Bayesian Data Analysis*. Boca Raton, FL: Chapman & Hall/CRC; 2014.
26. Myung IJ. Tutorial on maximum likelihood estimation. *J Math Psychol* 2003, 47:90–100.
27. Carpenter B, Gelman A, Hoffman M, Lee D, Goodrich B, Betancourt M, et al. Stan: a probabilistic programming language. *J Stat Softw* 2016, 76:1–32.
28. Nelder JA, Mead R. A simplex method for function minimization. *Comput J* 1964, 7:308–313 https://doi.org/10.1093/comjnl/7.4.308.
29. Wagenmakers E-J, Ratcliff R, Gomez R, Iverson GJ. Assessing model mimicry using the parametric bootstrap. *J Math Psychol*. 2004, 48:28–50.
30. Schwarz G. Estimating the dimension of a model. *Ann Stat* 1978, 6:461–464 https://doi.org/10.1214/aos/1176344136.
31. Akaike H. A new look at the statistical model identification. *IEEE Trans Autom Control* 1974, 19:716–723 https://doi.org/10.1109/TAC.1974.1100705.
32. Chater N, Oaksford M, Hahn U, Heit E. Bayesian models of cognition. *WIREs Cogn Sci*. 2010, 1:811–823.
33. Kruschke JK. Bayesian data analysis. *WIREs Cogn Sci* 2010, 1:658–676.
34. Bernardo JM, Smith AFM. *Bayesian Theory*. New York: Wiley; 1994.

35. de Finetti B. Foresight: its logical laws in subjective sources. In: Kyburg H, Smokler H, eds. *Studies in Subjective Probability*. New York: Wiley; 1964, 93–158.
36. Bayes T. A letter from the late reverend Mr. Thomas Bayes, FRS to John Canton, MA and FRS. *Philos Trans* 1763, 53:269–271.
37. Lee MD. Three case studies in the Bayesian analysis of cognitive models. *Psychon Bull Rev* 2008, 15:1–15.
38. Rouder JN, Lu J. An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychon Bull Rev* 2005, 12:573–604.
39. Matzke D, Dolan CV, Batchelder WH, Wagenmakers E-J. Bayesian estimation of multinomial processing tree models with heterogeneity in participants and items. *Psychometrika* 2015, 80:205–235.
40. Katahira K. How hierarchical models improve point estimates of model parameters at the individual level. *J Math Psychol* 2016, 73:37–58 https://doi.org/10.1016/j.jmp.2016.03.007.
41. Lee MD, Webb MR. Modeling individual differences in cognition. *Psychon Bull Rev* 2005, 12:605–621.
42. Okada K, Lee MD. A Bayesian approach to modeling group and individual differences in multidimensional scaling. *J Math Psychol* 2016, 70:35–44.
43. Merkle EC, Smithson M, Verkuilen J. Hierarchical models of simple mechanisms underlying confidence in decision making. *J Math Psychol* 2011, 55:57–67 https://doi.org/10.1016/j.jmp.2010.08.011.
44. Ravenzwaaij D, Moore CP, Lee MD, Newell BR. A hierarchical Bayesian modeling approach to searching and stopping in multi-attribute judgment. *Cogn Sci* 2014, 38:1384–1405.
45. Annis J, Miller BJ, Palmeri TJ. Bayesian inference with Stan: a tutorial on adding custom distributions. *Behav Res Methods* 2016.
46. Ratcliff R, Childers R. Individual differences and fitting methods for the two-choice diffusion model of decision making. *Decision* 2015, 2:237–279.
47. Vandekerckhove J, Tuerlinckx F, Lee MD. Hierarchical diffusion models for two-choice response times. *Psychol Methods* 2011, 16:44–62.
48. Wiecki TV, Sofer I, Frank MJ. HDDM: hierarchical bayesian estimation of the drift-diffusion model in python. *Front Neuroinform* 2013, 7:14.
49. Annis J, Lenes JG, Westfall HA, Criss AH, Malmberg KJ. The list-length effect does not discriminate between models of recognition memory. *J Mem Lang* 2015, 85:27–41.
50. Dennis S, Lee MD, Kinnell A. Bayesian analysis of recognition memory: the case of the list-length effect. *J Mem Lang* 2008, 59:361–376.
51. Morey RD. A Bayesian hierarchical model for the measurement of working memory capacity. *J Math Psychol* 2011, 55:8–24 https://doi.org/10.1016/j.jmp.2010.08.008.
52. Pratte MS, Rouder JN. Hierarchical single- and dual-process models of recognition memory. *J Math Psychol* 2011, 55:36–46 https://doi.org/10.1016/j.jmp.2010.08.007.
53. Turner BM, Forstmann BU, Wagenmakers E-J, Brown SD, Sederberg PB, Steyvers M. A Bayesian framework for simultaneously modeling neural and behavioral data. *Neuroimage* 2013, 72:193–206 https://doi.org/10.1016/j.neuroimage.2013.01.048.
54. Brown SD, Heathcote A. The simplest complete model of choice response time: linear ballistic accumulation. *Cogn Psychol* 2008, 57:153–178.
55. Lee MD, Wagenmakers E-J. *Bayesian Cognitive Modeling: A Practical Course*. Cambridge, UK: Cambridge University Press; 2014.
56. Lee MD. How cognitive modeling can benefit from hierarchical Bayesian models. *J Math Psychol* 2011, 55:1–7.
57. Shiffrin RM, Lee MD, Kim W, Wagenmakers E-J. A survey of model evaluation approaches with a tutorial on hierarchical bayesian methods. *Cogn Sci* 2008, 32:1248–1284.
58. Lynch SM. *Introduction to Applied Bayesian Statistics and Estimation for Social Scientists*. New York: Springer Science & Business Media; 2007.
59. Usher M, McClelland JL. The time course of perceptual choice: the leaky, competing accumulator model. *Psychol Rev* 2001, 108:550–592.
60. Hyndman RJ. Computing and graphing highest density regions. *Am Stat* 1996, 50:120–126.
61. Jaynes ET, Kempthorne O. Confidence intervals vs Bayesian intervals. In: Harper WL, Hooker C, eds. *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*. Dordrecht, The Netherlands: Springer; 1976, 175–257.
62. Morey RD, Hoekstra R, Rouder JN, Lee MD, Wagenmakers E-J. The fallacy of placing confidence in confidence intervals. *Psychon Bull Rev* 2016, 23:103–123 https://doi.org/10.3758/s13423-015-0947-8.
63. Lindley D. That wretched prior. *Significance* 2004, 1:85–87.
64. Efron B. Why isn't everyone a Bayesian? *Am Stat*. 1986, 40:1–5.
65. Lee MD, Vanpaemel W. Determining informative priors for cognitive models. *Psychon Bull Rev* 2017.
66. Vanpaemel W, Lee MD. Using priors to formalize theory: optimal attention and the generalized context model. *Psychon Bull Rev* 2012, 19:1047–1056 https://doi.org/10.3758/s13423-012-0300-4.

67. Jeffreys H. *Theory of Probability*. 3rd ed. Oxford, UK: Oxford University Press; 1961.
68. Kass RE, Wasserman L. The selection of prior distributions by formal rules. *J Am Stat Assoc* 1996, 91:1343–1370.
69. Spiegelhalter DJ, Smith AFM. Exact and approximate posterior moments for a normal location. *J R Stat Soc Ser B Stat Methodol*. 1982, 54:793–804.
70. Brooks S, Gelman A, Jones G, Meng X-L. *Handbook of Markov Chain Monte Carlo*. Boca Raton, FL: Chapman & Hall/CRC Press; 2011.
71. Green PJ, Łatuszyński K, Pereyra M, Robert CP. Bayesian computation: a summary of the current state, and samples backwards and forwards. *Stat Comput* 2015, 25:835–862.
72. van Ravenzwaaij D, Cassey P, Brown SD. A simple introduction to Markov chain Monte Carlo sampling. *Psychon Bull Rev* 2016 https://doi.org/10.3758/s13423-016-1015-8.
73. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, Teller E. Equation of state calculations by fast computing machines. *J Chem Phys* 1953, 21:1087–1092 https://doi.org/10.1063/1.1699114.
74. Hastings WK. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 1970, 57:97–109.
75. Gelfand A, Smith A. Sampling-based approaches to calculating marginal densities. *J Am Stat Assoc* 1990, 85:398–409.
76. Geman S, Geman D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans Pattern Anal Mach Intell* 1984, PAMI-6:721–741.
77. Gilks WR, Best NG, Tan KKC. Adaptive rejection metropolis sampling within Gibbs sampling. *J R Stat Soc Ser C Appl Stat* 1995, 44:455–472.
78. Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Stat Comput*. 2000, 10:325–337.
79. Plummer M. JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling. In: *Proceedings of 3rd International Workshop on Distributed Statistical Computing* (DSC 2003), 2003, 20–2 doi:https://doi.org/10.1.1.13.3406.
80. Neal RM. MCMC using Hamiltonian dynamics. In: Brooks S, Gelman A, Jones G, Meng X-L, eds. *Handbook of Markov Chain Monte Carlo*. Boca Raton, FL: Chapman & Hall/CRC Press; 2011, 113–162.
81. Wabersich D, Vandekerckhove J, Extending JAGS. A tutorial on adding custom distributions to JAGS (with a diffusion model example). *Behav Res Methods* 2014, 46:15–28 https://doi.org/10.3758/s13428-013-0369-3.
82. Donkin C, Brown S, Heathcote A. Drawing conclusions from choice response time models: a tutorial using the linear ballistic accumulator. *J Math Psychol* 2011, 55:140–151.
83. Holmes WR. A practical guide to the probability density approximation (PDA) with improved implementation and error characterization. *J Math Psychol* 2015, 68–69:13–24.
84. Turner BM, Sederberg PB. Approximate Bayesian computation with differential evolution. *J Math Psychol* 2012, 56:375–385 https://doi.org/10.1016/j.jmp.2012.06.004.
85. Turner BM, Sederberg PB, McClelland JL. Bayesian analysis of simulation-based models. *J Math Psychol* 2014, 72:191–199 https://doi.org/10.1016/j.jmp.2014.10.001.
86. Turner BM, Van Zandt T. A tutorial on approximate Bayesian computation. *J Math Psychol* 2012, 56:69–85 https://doi.org/10.1016/j.jmp.2012.02.005.
87. Turner BM, Van Zandt T. Hierarchical approximate Bayesian computation. *Psychometrika* 2013, 79:185–209.
88. Gelman A. Two simple examples for understanding posterior p-values whose distributions are far from uniform. *Electron J Stat* 2013, 7:2595–2602.
89. Zhang JL. Comparative investigation of three Bayesian p values. *Comput Stat Data Anal* 2014, 79:277–291 https://doi.org/10.1016/j.csda.2014.05.012.
90. Kruschke JK. Posterior predictive checks can and should be Bayesian: comment on Gelman and Shalizi, "philosophy and the practice of Bayesian statistics". *Br J Math Stat Psychol* 2013, 66:45–56.
91. Myung IJ. The importance of complexity in model selection. *J Math Psychol* 2000, 44:190–204 https://doi.org/10.1006/jmps.1999.1283.
92. Good IJ. The interface between statistics and philosophy of science. *Stat Sci* 1988, 3:386–412 https://doi.org/10.1214/ss/1177013604.
93. Kass RE, Raftery AE. Bayes factors. *J Am Stat Assoc* 1995, 90:773–795.
94. Shiffrin RM, Chandramouli SH, Grunwald PD. Bayes factors, relations to minimum description length, and overlapping model classes. *J Math Psychol*. 2015, 72:56–77 https://doi.org/10.1016/j.jmp.2015.11.002.
95. Kruschke JK. Bayesian assessment of null values via parameter estimation and model comparison. *Perspect Psychol Sci* 2011, 6:299–312.
96. Liu CC, Aitkin M. Bayes factors: prior sensitivity and model generalizability. *J Math Psychol* 2008, 52:362–375 https://doi.org/10.1016/j.jmp.2008.03.002.
97. Vanpaemel W. Prior sensitivity in theory testing: an apologia for the Bayes factor. *J Math Psychol*

2010, 54:491–498 https://doi.org/10.1016/j.jmp.2010.07.003.

98. Evans NJ, Brown SD. Bayes factors for the Linear Ballistic Accumulator Model of Decision-Making. *Behav Res Methods* 2017 Advance online publication, https://doi.org/10.3758/s13428-017-0887-5.
99. Luebke D. CUDA: Scalable parallel programming for high-performance scientific computing. In: *5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, 2008 ISBI 2008*, 2008, 836–838.
100. Annis J, Evans NJ, Miller BJ, Palmeri TJ. Thermodynamic integration and steppingstone sampling methods for estimating Bayes factors: a tutorial for psychologists
101. Wagenmakers E-J, Lodewyckx T, Kuriyal H, Grasman R. Bayesian hypothesis testing for psychologists: a tutorial on the savage-dickey method. *Cogn Psychol* 2010, 60:158–189 https://doi.org/10.1016/j.cogpsych.2009.12.001.
102. Morey RD, Rouder JN, Pratte MS, Speckman PL. Using MCMC chain outputs to efficiently estimate Bayes factors. *J Math Psychol* 2011, 55:368–378.
103. Friel N, Wyse J. Estimating the evidence - a review. *Stat Neerl* 2012, 66:288–308.
104. Geweke BYJ. Bayesian inference in iconometric models using monte carlo integration. *Econometrica* 1989, 57:1317–1339.
105. Newton M, Raftery A. Approximate Bayesian inference with the weighted likelihood bootstrap. *J R Stat Soc Ser B Stat Methodol* 1994, 56:3–48.
106. Neal RM. Annealed importance sampling. *Stat Comput* 2001, 11:125–139.
107. Meng X-L, Wong HW. Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Stat Sin* 1996, 6:831–860.
108. Chib S, Jeliazkov I. Marginal likelihood from the metropolis-Hastings output. *J Am Stat Assoc* 2001, 96:270–281 https://doi.org/10.2307/2291521.
109. Chib S. Marginal likelihood from the Gibbs output. *J Am Stat Assoc* 1995, 90:1313–1321 https://doi.org/10.2307/2291521.
110. Skilling J. Nested sampling for Bayesian computations. *Bayesian Anal* 2006, 1:833–860 https://doi.org/10.1214/06-BA127.
111. Friel N, Pettitt AN. Marginal likelihood estimation via power posteriors. *J R Stat Soc Ser B Stat Methodol.* 2008, 70:589–607.
112. Lartillot N, Philippe H. Computing bayes factors using thermodynamic integration. *Syst Biol* 2006, 55:195–207 https://doi.org/10.1080/10635150500433722.
113. Liu P, Elshall AS, Ye M, Beerli P, Zeng X, Lu D, Tao Y. Evaluating marginal likelihood with thermodynamic integration method and comparison with several other numerical methods. *Water Resour Res* 2016, 52:734–758.
114. Green PJ. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* 1995, 82:711–732.
115. Carlin BP, Chib S. Bayesian model choice via Markov chain Monte Carlo methods. *J R Stat Soc Ser B (Stat Methodol)* 1995, 57:473–484.
116. Lodewyckx T, Kim W, Lee MD, Tuerlinckx F, Kuppens P, Wagenmakers E-J. A tutorial on Bayes factor estimation with the product space method. *J Math Psychol.* 2011, 55:331–347.
117. Ando T. Bayesian predictive information criterion for the evaluation of hierarchical Bayesian and empirical Bayes models. *Biometrika* 2017, 94:443–458.
118. Friel N, Mckeone JP, Oates CJ, Pettitt AN. Investigation of the widely applicable Bayesian information criterion. *Stat Comput* 2016, 27:833–844.
119. Watanabe S. A widely applicable Bayesian information criterion. *Mach Learn Res* 2013, 14:867–897.
120. Gelman A, Hwang J, Vehtari A. Understanding predictive information criteria for Bayesian models. *Stat Comput* 2014, 24:997–1016.
121. Watanabe S. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *J Mach Learn Res* 2010, 11:3571–3594.
122. Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. Bayesian measures of model complexity and fit. *J R Stat Soc Ser B Stat Methodol* 2002, 64:583–639.
123. Lee MD, Wagenmakers E-J. Bayesian statistical inference in psychology: comment on Trafimow (2003). *Psychol Rev.* 2005, 112:662–668.
124. Lee MD. Determining the dimensionality of multidimensional scaling representations for cognitive modeling. *J Math Psychol* 2001, 45:149–166 https://doi.org/10.1006/jmps.1999.1300.
125. Rouder JN, Lu J, Speckman P, Sun D. A hierarchical model for estimating response time distributions. *Psychon Bull Rev* 2005, 12:195–223.
126. Vanpaemel W. Constructing informative model priors using hierarchical methods. *J Math Psychol* 2011, 55:106–117 https://doi.org/10.1016/j.jmp.2010.08.005.
127. Nosofsky RM. Attention , similarity , and the identification-categorization relationship. *J Exp Psychol Gen* 1986, 115:39–57.
128. Sinharay S, Stern HS. On the sensitivity of Bayes factors to the prior distributions. *Am Stat* 2002, 56:196–201.
129. Rouder JN, Morey RD, Pratte MS. Hierarchical Bayesian models. In: Batchelder WH, Colonius H, Dzhafarov E, Myung JI, eds. *New Handbook of Mathematical Psychology*. Measurement and

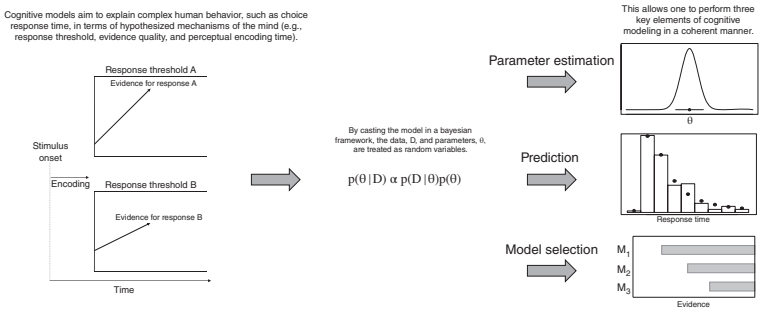
Methodology, vol. 1. London, UK: Cambridge University Press; 2013.

130. Estes WK. The problem of inference from curves based on group data. *Psychol Bull* 1956, 53:134–140.
131. Cohen AL, Sanborn AN, Shiffrin RM. Model evaluation using grouped or individual data. *Psychon Bull Rev* 2008, 15:692–712.
132. Ashby FG, Maddox WT, Lee WW. On the dangers of averaging across subjects when using multidimensional scaling or the similarity choice model. *Psychol Sci* 1994, 5:144–151.
133. Lee MD, Pope KJ. Avoiding the dangers of averaging across subjects when using multidimensional scaling. *J Math Psychol.* 2003, 47:32–46.

Graphical abstract

Bayesian statistical approaches to evaluate cognitive models

Jeffrey Annis[1], Thomas J. Palmeri[1]



Cognitive models aim to explain complex human behavior such as chance response time, in terms of hypothesized mechanisms of the mind (e.g., response threshold, evidence quality, and perceptual encoding time).