A General Computational Method for Estimating Bayes Factors

Tom Lodewyckx
Department of Psychology
University of Leuven

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

Eric-Jan Wagenmakers
Department of Psychology
University of Amsterdam

Abstract

The Bayes Factor is an averaged likelihood ratio for evaluating models using data. It automatically balances goodness-of-fit with complexity, and is among the most important and widely-used measures in Bayesian statistics. We argue that Bayes Factors are useful measures for the quantitative evaluation of statistical hypotheses and specific models in psychology, but have been under-used because they are often difficult to compute. We develop a largely automated method for estimating Bayes Factors based on the pioneering work of Carlin and Chib (1995), and provide all of the code needed for practical implementation. The method uses a conceptually simple approach, and is applicable to a very general range of situations, including the comparison of non-nested and hierarchical hypotheses and models. We demonstrate our method in two illustrative examples, and validate its accuracy where possible using an alternative method for estimating Bayes Factors. We conclude with a discussion of the strengths and limitations of our method, and argue it offers a relatively easy-to-use and general capability for the quantitative evaluation of hypotheses and models in psychology.

Keywords: Bayes factor, Bayesian statistics, graphical modeling, hierarchical modeling, hypothesis testing, model selection.

Introduction

A key to progress in psychology is the ability to evaluate theoretical ideas quantitatively against empirical observations. One common approach to quantitative evaluation involves the statistical analysis of data, which relies on general statistical models. Another common approach involves the evaluation of more specific psychological models, which provide detailed accounts of people's behavior in psychological tasks.

In both situations, one of the most basic challenges for quantitative evaluation is to be able to choose between competing models. For general statistical models, this comparison is usually called *hypothesis testing*. For more specific psychological models it is often called *model selection*. There are, of course, many formal and quantitative ways to compare and choose between models using data. Frequentist hypothesis testing relies on p-values, confidence intervals, and other devices developed within the sampling distribution statistical approach, despite well-known and well-documented problems (see Wagenmakers, 2007, for a recent overview focused on the issues as they are relevant to psychology). More recently, psychology followed the lead of modern statistics, and other empirical sciences, in adopting Bayesian methods to evaluate their models (e.g., Lee, 2008; Pitt, Myung, & Zhang, 2002; Shiffrin, Lee, Kim, & Wagenmakers, 2008). The Bayesian approach has the advantages of being a conceptually simple, theoretically coherent, and generally applicable way to make inferences about models from data (see Lee & Wagenmakers, 2005).

In this paper, we focus on the most popular—but certainly not the only—Bayesian method for choosing between competing models, known as the *Bayes Factor* (Jeffreys, 1961; Kass & Raftery, 1995). Intuitively, Bayes Factors simply measure the relative level of evidence data provide for one model over another, in the form of a likelihood ratio. Bayes Factors also automatically account for formal aspects of model complexity, rewarding simple models and penalizing complicated ones. This property is important to avoid choosing models that over-fit data (Pitt et al., 2002).

The psychological literature has a number of recent examples of the Bayes Factor being applied, including in general statistical settings (e.g., Hoijtink, 2001; Rouder, Speckman, Sun, Morey, & Iverson, 2009; Wetzels, Raaijmakers, Jakab, & Wagenmakers, 2009), and to specific psychological models (e.g., Gallistel, 2009; Kemp & Tenenbaum, 2008; Lee, 2002, 2004; Pitt et al., 2002; Steyvers, Lee, & Wagenmakers, 2009), but it could hardly be described as a widely used approach. There are a few possible reasons for the lack of application of Bayes Factors. Most obviously, there is a strong temptation to stay with known methods for analyzing data while they remain acceptable practice, whatever the limitations those methods impose or the errors they make.

More interestingly, even among those who accept the need to use the Bayesian approach, it is understood that it can be difficult to calculate Bayes Factors in practice. Sometimes, easily calculated but theoretically limited approximations to the Bayes Factor, such as those based on the Bayesian Information Criterion (BIC), have been used (e.g., Vickers, Lee, Dry, & Hughes, 2003). But, more often, Bayesian statistical methods have focused on parameter estimation rather than model selection. This has especially been the case when the models themselves have been complicated enough—often taking the form of hierarchical models—to warrant the use of modern computational approaches to Bayesian inference (e.g., Kuss, Jäkel, & Wichmann, 2005; Lee, 2006, 2008; Rouder & Lu, 2005; Rouder, Lu, et al., 2007; Rouder, Lu, Morey, Sun, & Speckman, 2008).

The aim of this paper is to develop and demonstrate a method for estimating Bayes Factors using a computational approach, based on the pioneering work of Carlin and Chib (1995). The method is general, in the sense that it can be applied to compare any models, including hierarchical models, and to any model comparison, including between non-nested models. We first provide a formal account of the Bayes Factor, and its estimation using the method developed by Carlin and Chib (1995). We then extend their method, and provide an easy-to-use and general implementation, which we demonstrate on two psychological examples. We conclude with a discussion of the strengths, weaknesses, and niche of application for

our method.

Understanding and Estimating Bayes Factors

Understanding Bayes Factors

A Bayes Factor compares two probabilistic models, each of which may have some number of parameters. Parameters are variables that index the predictions a model can make. In a general statistical model, a parameter might correspond to the size of the difference between means in experimental groups, or the size of an interaction effect. In a more specific model, a parameter might correspond to a psychological variable, like a learning rate, or a memory capacity, or a response threshold, that is assumed to help control a psychological process.

The Bayes Factor compares two models by considering *on average* how well each can fit the observed data, where the average is taken with respect to all of the possible values of the parameters. It is this averaging that accounts for differences in model complexity, because more complicated models (i.e., those that can fit many data patterns by changing their parameter values) will have lower average levels of fit than simple models which can only predict the observed data.

Formally, if Model A with parameters θ_a is being compared to Model B with parameters θ_b using data D, the Bayes Factor is defined as

$$B_{AB} = \frac{p\left(D \mid M_A\right)}{p\left(D \mid M_B\right)} = \frac{\int p\left(D \mid \theta_a, M_A\right) p\left(\theta_a \mid M_A\right) \, \mathrm{d}\theta_a}{\int p\left(D \mid \theta_b, M_B\right) p\left(\theta_b \mid M_B\right) \, \mathrm{d}\theta_b}. \tag{1}$$

Equation 1 shows that the Bayes Factor is the ratio of two likelihoods, $p\left(D \mid M_A\right)$ and $p\left(D \mid M_B\right)$, representing how likely the data are under each model, and that these likelihoods are found by *averaging* or *marginalizing* the likelihood across the prior parameter space for each model. Intuitively, the Bayes Factor is the likelihood ratio for the two models, averaged across all of the predictions they make. For the marginal likelihood to be high, a model must not only be able to fit the observed data well, but also must not predict data different from those observed.

The standard definition of the Bayes Factor in Equation 1 can be rewritten as

$$\frac{p(M_A \mid D)}{p(M_B \mid D)} = B_{AB} \times \frac{p(M_A)}{p(M_B)}, \tag{2}$$

which can be read as: "Posterior model odds = $B_{AB} \times$Prior model odds." This gives a second interpretation of the Bayes Factor as the change in the model odds resulting from observing data. That is, whatever the prior odds in favor of Model A, the Bayes Factor B_{AB} is the multiple that describes the increase or decrease in those odds following from the new evidence provided by the data D.

Because Bayes Factors can be considered as likelihood ratios, it is easy to interpret and calibrate them in the context of betting. Raftery (1995) proposed a useful interpretation scheme for values of the Bayes factor, as presented in Table 1. This table includes a verbal expression of the strength of evidence, and corresponding ranges for the Bayes Factor B_{AB} itself, for its logarithmic rescaling $\log BF_{AB}$, and for the posterior probability of Model A $\Pr(M_A \mid D)$, assuming equal prior probabilities for the models. Expressing Bayes Factors on the logarithmic scale has the advantages of making zero the point of indifference between the two models being compared (i.e., the point at which the Bayes Factor is 1, and the

data provide no more evidence for one model than the other), and making equal increments correspond to equal changes in the relative probabilities (i.e., $\log BF_{AB} = +2$ is the same level of evidence in favor of Model A as $\log BF_{AB} = -2$ is in favor of model B). The posterior probability is a convenient and easily interpreted value in cases where the two models being compared are the only ones of theoretical interest.

Table 1: Interpretation for values of the Bayes factor, the logarithm of the Bayes Factor, and the corresponding posterior model probability, according to Raftery (1995).

Interpretation	B_{AB}	$\log(B_{AB})$	$\Pr(M_A \mid D)$
Very strong support for M_B	$< .0067$	$<$ -5	$< .01$
Strong support for M_B	.0067 to .05	-5 to -3	.01 to .05
Positive support for M_B	.05 to .33	-3 to -1	.05 to .25
Weak support for M_B	.33 to 1	-1 to 0	.25 to .50
No support for either model	1	0	.50
Weak support for M_A	1 to 3	0 to 1	.50 to .75
Positive support for M_A	3 to 20	1 to 3	.75 to .95
Strong support for M_A	20 to 150	3 to 5	.95 to .99
Very strong support for M_A	> 150	> 5	$> .99$

Estimating Bayes Factors

For all but the simplest models, the integrations required to calculate Bayes Factors are analytically intractable. Accordingly, a large number of methods have been developed to approximate Bayes Factors. The earliest methods focused on analytic approximations to the required integration (see Kass & Raftery, 1995, for a review). Many of these approaches continue to be refined (e.g., Myung, Balasubramanian, & Pitt, 2000), and remain useful and applicable methods for many simple statistical and psychological models.

More recently, Bayes Factor estimation has been approached within a computational sampling-based framework for inference, mirroring the shift in inferences about parameters from analytic to computational methods. Within the computational framework, there are at least two quite different approaches for estimating Bayes Factors. The first approach is based on estimating the marginal model likelihoods for both models separately, as per Equation 1. This approach includes methods such as prior simulation (Kass & Raftery, 1995), importance sampling (DiCiccio, Kass, Raftery, & Wasserman, 1997; Geweke, 1989), candidate estimation (Chib, 1995), and the Laplace (Tierney & Kadane, 1986), and Laplace-Metropolis methods (Lewis & Raftery, 1997).

The second computational approach to Bayes Factor approximation is rooted in trans-dimensional Markov Chain Monte Carlo (MCMC) methods. It involves estimating posterior model odds for chosen prior model odds, as per Equation 2. Reversible jump MCMC (Green, 1995) is one widely used transdimensional MCMC method. A less widely used method is one developed by Carlin and Chib (1995), known as the product space method. Both methods are conceptually very simple, and rely on combining the two models to be compared within one hierarchical "supermodel."

Figure 1 presents the basic framework of this approach graphically. The hierarchical combination of Model A and Model B is achieved using a single binary 'model indicator' variable M that controls which model generates the observed data D. The prior of the model

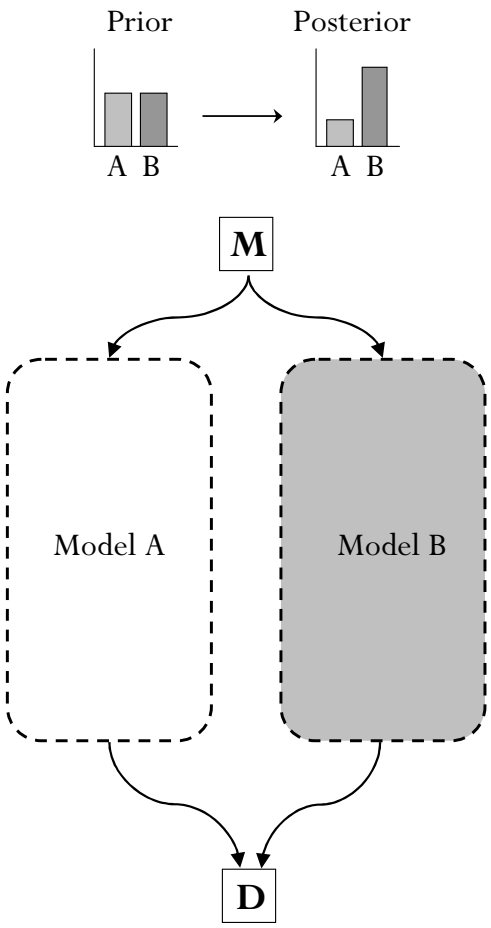


Figure 1. The general framework for the transdimensional MCMC method of estimating Bayes Factors. The two models to be compared—Model A and Model B—are hierarchically joined by a model indicator variable M that controls which model generates the observed data D. The Bayes Factor can then be estimated by comparing the prior and posterior distribution of the M.

indicator variable corresponds to the prior model odds, and is specified in defining the combined hierarchical model. The posterior of the model indicator variable corresponds to the posterior odds, and can be estimated by MCMC posterior sampling methods. Combining these two odds (the first exact, the second estimated) according to Equation 2 then automatically gives an estimate of the Bayes Factor. In the schematic demonstration in Figure 1, for example, both models are equally likely in the prior, but Model B is about three times more likely in the posterior. This change from prior to posterior odds corresponds to a Bayes Factor of about 3.

An Implementation of the Product Space Method

The conceptual simplicity of the product space approach is unfortunately not matched by a practical simplicity in its implementation. There are two important computational issues—involving pseudopriors for parameters, and priors for models—that need to be ad-

dressed before the product space approach provides a working estimation method for many interesting statistical and psychological models. It is by addressing these two problems that our approach extends and makes more accessible the pioneering work of Carlin and Chib (1995).

Parameter Pseudopriors

The first problem our method addresses involves the "linking densities" or "pseudopriors" that have to be specified to guarantee continuous sampling for all parameters in the supermodel. When the model indicator takes the value $M = 0$, and so Model A is being used to model the data, posterior values are sampled for θ_a conditional on the data, as is normally the case using MCMC. At the same time, however, values are drawn from the pseudoprior for θ_b. That is, even though Model B is not being used to model the data for the current sample, values for its parameters must be sampled, and this sampling is done from a pseudoprior distribution. Similarly, for samples with $M = 1$, posterior values for θ_b are sampled according to the data, and values are drawn from the pseudoprior of θ_a.

As Carlin and Chib (1995) note, for the entire product space approach to be efficient, it is important to choose pseudopriors that approximate the true posterior distributions of θ_a and θ_b. In previous practical implementations of the product space method, these choices have been made using separate analyses of the individual models against the data (e.g., Carlin & Chib, 1995; Spiegelhalter, Thomas, Best, & Gilks, 1996). For example, the pseudoprior of a parameter might be assumed to have a Gaussian form, and then the mean and variance of the Gaussian is estimated from the posterior of the parameter when only that model is applied to data. This approach is sensible, but can be inefficient or problematic if good choices for pseudopriors are not well described by the chosen distribution (e.g., if a pseudoprior should be highly skewed, but a Gaussian form is assumed).

Our method solves this problem by approximating the pseudoprior distributions directly using posterior sampling. We conduct independent inferences for each of the two models *simultaneously* with the estimation of the supermodel, and use the posterior samples of model parameters from these independent runs as the required samples from the pseudopriors. This direct computational approach avoids the need to choose a parametric distribution, and provides an efficient characterization of the pseudopriors.

Model Priors

The second problem our method addresses involves choosing the prior distribution on the model indicator variable. The prior is a Bernoulli distribution, parameterized by the prior probability $\hat{\pi}$ that Model A is true. It is the choice of the probability $\hat{\pi}$ that is important. Ideally, for the efficiency of computational estimation, this prior should lead to both models being considered equally often, so that the posterior probabilities for both models are 0.5. A strong asymmetry in the proportion of times each model is considered (e.g., if one of the models is considered on fewer than 10% of the samples) will result in poor estimation of the parameters of that model. This failing, in turn, affects the quality of the Bayes Factor estimate.

To address this problem, our method iteratively tunes the prior distribution on the model indicator variable. A first Bayes Factor estimate $\hat{B}_{AB}$ is obtained for some initial choice of prior, usually given by the $\hat{\pi} = 1/2$. Then, given the goal of obtaining equal posterior probabilities, the prior model probabilities can be updated according to $\hat{\pi}_{new} =$

$\hat{B}_{AB}/(1+\hat{B}_{AB})$. Our method repeats this updating as often as required for the posterior probabilities to be approximately equal. As the prior is updated, the quality of $\hat{B}_{AB}$ improves.

Implementation

Our method implements these extensions of the Carlin and Chib (1995) approach in a unified and largely automated way within the widely used WinBUGS software (Lunn, Thomas, Best, & Spiegelhalter, 2000; Lunn, Spiegelhalter, Thomas, & Best, in press). In this way, our method potentially provides a general way of estimating Bayes Factors for the wide array of statistical and psychological models that can be expressed as probabilistic graphical models (e.g., Donkin, Averell, Brown, & Heathcote, in press; Kemp, Shafto, Berke, & Tenenbaum, 2007; Lee, 2008; Lee & Wagenmakers, 2009; Shiffrin et al., 2008; Vandekerckhove, Tuerlinckx, & Lee, 2008; Wetzels, Vandekerckhove, Tuerlinckx, & Wagenmakers, in press). An online appendix that provides all of the technical details, together with WinBUGS and R code, is available at `http://sites.google.com/site/tomlodewyckx`.

Two Illustrative Applications

We present two applications of the product space method for hypothesis testing in the domain of subliminal priming (Rouder, Morey, Speckman, & Pratte, 2007; Zeelenberg, Wagenmakers, & Raaijmakers, 2002). For each application, we first describe the research question, experimental setup and data, and discuss the original statistical analyses. We then show our method can estimate Bayes Factors for an appropriate hypothesis test or model selection problem, to address the same research question.

Illustration 1: Is Priming Subliminal?

Research Question. In a standard subliminal priming task, the participant is presented a prime stimulus for a very short time interval. The research question is whether and how some characteristics of the prime influence subsequent behavior. It is often indirectly assumed that the presentation time of the prime stimulus is so short that the participant does not perceive stimuli consciously, implying that any influence on the overt behavior is due to unconscious processes. However, it is not clear to what extent primes are perceived on a subliminal level, or whether there are important individual differences.

Experimental Design and Data. Rouder, Morey, et al. (2007) conducted an experiment to test the validity of the assumption of subliminality. The visually presented stimulus material consisted of the numbers 2, 3, 4, 6, 7 and 8. In each trial, one of these numbers was presented on the computer screen as a 22 ms prime stimulus, followed by a 66 ms mask and another number as a 200 ms target stimulus. In the prime-identification task, the participant had to indicate whether the prime stimulus in the current trial was higher or lower than 5. The dependent measure was the accuracy of the answer, so that the experiment resulted in K successes out of N trials. All 27 participants were presented 288 trials.[1]

The proportions of successes and a non-parametric density estimate of the proportions are presented in Figure 2. Under the assumption of subliminal perception, a correct identification of the higher-than-five status of the prime stimulus is expected in about half of the trials (i.e., "performance at chance"). Assuming supraliminal perception, performance is expected to be better than chance level.

[1] For some of the participants, the data are incomplete such that $N < 288$.

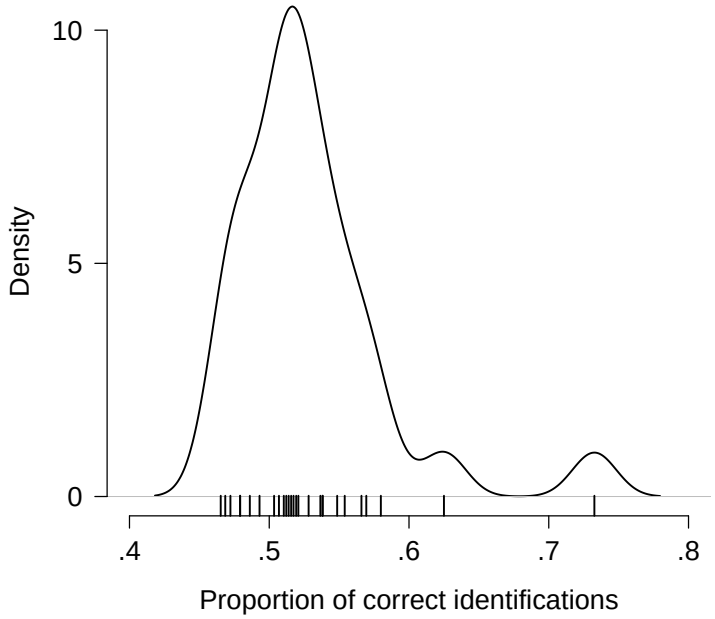
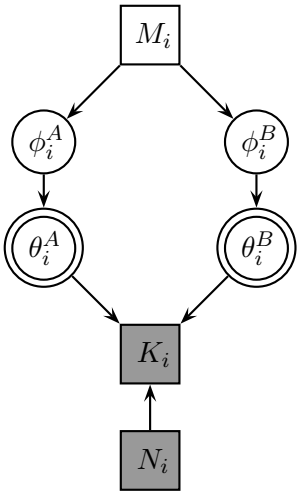


Figure 2. Proportions of successes for the 27 participants in the prime-identification task reported by Rouder et al. (2007), marked as ticks on the horizontal axis. The curve is a non-parametric density estimate of these proportions.

Previous Analysis. The "mass at chance model" (Rouder, Morey, et al., 2007) is an advanced model to analyze counts of successes in subliminal priming experiments. The essence of the model is that a binomial rate parameter can be higher than 0.5, implying supraliminality, whereas the value of 0.5 corresponds to subliminal perception, which is the theoretical lower limit for that rate parameter. Therefore, the mass below the value of 0.5 is squeezed to the value of 0.5. Using this model Rouder, Morey, et al. (2007) concluded that there was evidence in favor of subliminal perception for only 3 of the 27 participants, with marginal evidence for another 2 participants. For the remaining 22 participants, they concluded: "Although many of these participants may be truly at chance, we do not have sufficient evidence from the data to conclude this."

Bayes Factor Analysis. One way to test the subliminality assumption is to find the Bayes Factors comparing the supraliminal and subliminal models for each participant. Both competing models are formally described in Figure 3, using the notation provided by graphical models. Graphical models are a standard language for representing probabilistic models, widely used in statistics and machine learning (e.g., Gilks, Thomas, & Spiegelhalter, 1994; Jordan, 2004; Koller, Friedman, Getoor, & Taskar, 2007), and recently finding some popularity in psychological modeling (e.g., Kemp et al., 2007; Lee, 2008; Shiffrin et al., 2008).

The graphical model presented in Figure 3 uses the same notation as Lee (2008). Nodes in the graph correspond to variables, and the graphical structure is used to indicate depen-



$$
\begin{aligned}
M_i &\sim \text{Bernoulli}(\hat{\pi}) \\
\phi_i^A &= 0 \\
\phi_i^B &\sim \text{Normal}_{(0,+\infty)}(0,1) \\
\theta_i^A &= \Phi(\phi_i^A) \\
\theta_i^B &= \Phi(\phi_i^B) \\
K_i &\sim \begin{cases} \text{Binomial}(\theta_i^A, N_i) & \text{if } M_i = 0 \\ \text{Binomial}(\theta_i^B, N_i) & \text{if } M_i = 1 \end{cases}
\end{aligned}
$$

Figure 3. Graphical model for the individual participant Bayes Factor analysis of the Rouder et al. (2007) data.

dencies between the variables, with children depending on their parents. Continuous variables are represented with circular nodes and discrete variables with square nodes. Observed variables (i.e., usually data) are shaded and unobserved variables (i.e., usually model parameters) are not shaded. Deterministic variables—that is, variables that are simply functions of other nodes, and included for conceptual clarity—are shown as double-bordered nodes.

In Figure 3, K_i is the (observed, discrete) number of successes out of the (observed, discrete) number of trials N_i for the ith participant. These data are modeled as following a Binomial distribution, using either a rate parameter θ_i^A, generated according to Model A, or a rate parameter θ_i^B, generated according to Model B. Which rate generates the data is controled by the (unobserved, discrete) model indicator variable M.

The supraliminal and subliminal models correspond to different assumptions about the prior distribution on the rate of success. These different priors are most easily expressed on a reparameterization of θ, given by $\phi = \Phi^{-1}(\theta)$. This is the probit transformation, or the inverse cumulative standard normal distribution function. It is useful to model rate parameters this way, because it facilitates the construction of priors (Rouder & Lu, 2005). Specifically, in Figure 3 the subliminal model, with a chance level of performance, needs a prior $\theta_i^A = 0.5$ which corresponds to $\phi_i^A = \Phi^{-1}(0.5) = 0$. The supraliminal model, in contrast, uses truncated Normal prior that allows only positive values, $\phi_i^B \sim \text{Normal}_{(0,+\infty)}$, corresponding to a uniform prior over rates with performance better than chance.

It is straightforward to interpret the combined hierarchical graphical model in Figure 3 in terms of the schematic framework provided by Figure 1. The model indicator M controls whether the observed data K_i are generated according to the subliminal Model A on the left, or the supraliminal Model B on the right. For the more elaborate models we will consider later, however, it becomes unwieldy to show the full combined hierarchical graphical model. Instead, we will reply on a graphical model characterization like that shown in Figure 4, which is a compact representation of the same model information. It simply shows the generating

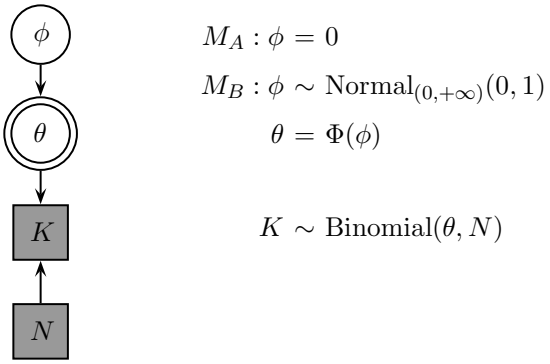


Figure 4. A compact graphical model representation for the individual participant Bayes Factor analysis of the Rouder et al. (2007) data.

process that is common to both models, and indicates the different choices of priors that distinguish the models. We also do not explicitly denote each of the participants in Figure 4, since they all are defined identically and completely by the generating process shown.

Returning to the main aim of comparing the models, we estimated the log Bayes factor comparing the subliminal and supraliminal models for each participant using our product space method. To do this, we used three chains, which are independent runs of the MCMC sampler. Each chain contained 100,000 samples after removing a burn in of 1000 samples. Random draws from the joint prior distribution were used as initial values. Convergence was evaluated with the $\hat{R}$ statistic, comparing within-chain variance and between-chain variance (Gelman, Carlin, Stern, & Rubin, 2004). We observed that the chains always converged, except for the data for one participant who had a proportion correct of 0.73, which is so extreme that it did not allow the prior model probability to be calibrated within the available computational precision. All of the retained samples from all three chains were used to estimate the log Bayes factor.

As a check of accuracy, we also estimated the same log Bayes Factors using an alternative method known as the Savage-Dickey method (Dickey & Lientz, 1970; Wetzels et al., 2009), which is applicable in this case, because the simpler subliminal model is nested within the more general supraliminal model. Figure 5a shows that the estimates for our product space method, $\log B^{PS}_{AB}$, are consistent with the validation estimates $\log B^{SD}_{AB}$ provided by the Savage-Dickey method.

Figure 5b shows $\log B^{PS}_{AB}$ as a function of the observed proportions of correct identifications K_i/N_i for all participants. As expected, support for the subliminal model is stronger for participants with a proportion close to or smaller than 0.5. There is at least strong support for this model when the participant answered fewer than 48% of the trials correctly. This is the case for only five participants. These are the same five participants identified by Rouder, Morey, et al. (2007), and so our basic conclusions match the results of their previous analyses.

Illustration 2: Does Previous Study Improve Visual Discriminability?

Research Question. Zeelenberg et al. (2002) report the results from a series of experiments examining whether previous exposure to a stimulus facilitated visual discriminabil-

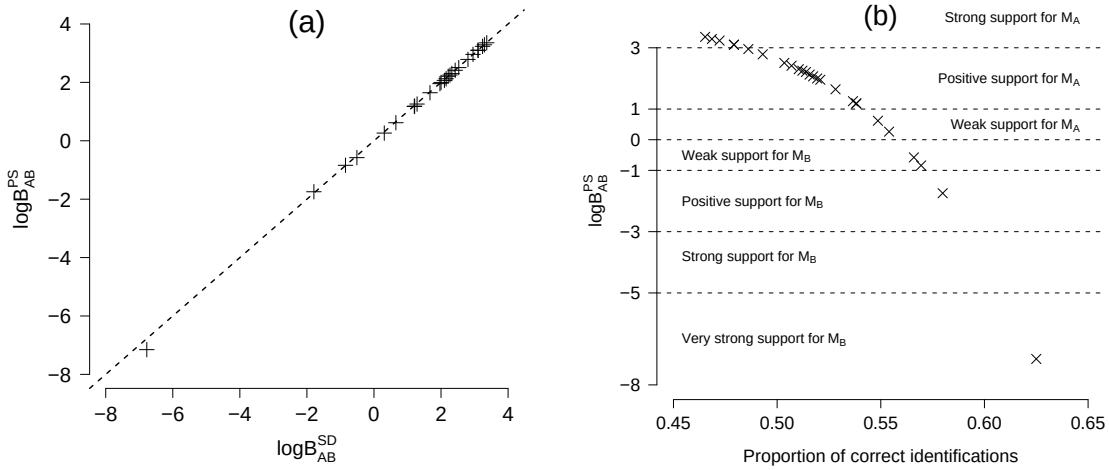


Figure 5. Bayes Factor results for the Rouder et al. (2007) data, showing (a) the comparison of our $\log B^{PS}_{AB}$ product space estimates with the $\log B^{SD}_{AB}$ Savage-Dickey validation estimates for the 26 participants, and (b) $\log B^{PS}_{AB}$ as a function of the proportions of correct identifications, with dashed lines indicating bounds for interpretation, as detailed in Table 1.

ity. A basic experimental trial involved presenting one of a pair of visually similar pictures (e.g., a clothes peg and a stapler) for 40ms, as a 'target', and then presenting both pictures. The participant's task is to identify the target, within a simple two-alternative forced-choice paradigm.

We focus on the third experiment presented by Zeelenberg et al. (2002), which included a study block and a test block. In the study block, participants were able to study half of the picture pairs. In the test block, participants were presented with all 42 picture pairs. This design naturally divides each test trial into 'studied both' (SB) and 'studied neither' (SN) experimental conditions, and the observed data are counts K_i^{SB} and K_i^{SN} of correct identifications in each condition for the ith participant. Figure 6 shows the relationship between the proportions of successes in the two experimental conditions for all 74 participants.

Previous Analysis. Zeelenberg et al. (2002) found a significant group effect, using a paired t-test: the percentage of correct trials was higher in the Study Both condition (74.7%) than in the Study Neither condition (71.5%), with $t(73) = 2.19$, $p < .05$. This result was taken to support the hypothesis that prior study leads to an improved visual discriminability.

Individual Participant Bayes Factor Analysis. The graphical model used in our analysis is presented in Figure 7, in the compact form discussed earlier. For both experimental conditions, we assume the counts of correct identifications are Binomially distributed with success rates θ^{SB} and θ^{SN}. As in the previous application, we use probit transformations of these rates ϕ^{SB} and ϕ^{SN} to help in setting prior distributions.

The key part of the analysis then concerns the difference between the success rates for the two conditions, formalized as the difference $\alpha = \phi^{SB} - \phi^{SN}$. To test whether a participant performed better in the Study Both condition than in the Study Neither condition,

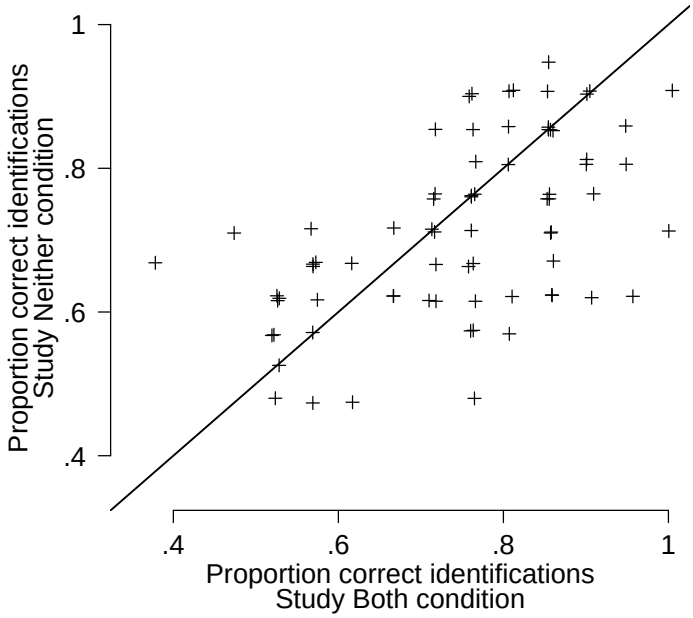


Figure 6. Proportions of correct identifications of the 74 participants in the Study Both and Study Neither conditions, using jitter to distinguish participants with the same proportions.

we compared the model that assumes no difference $\alpha = 0$ (M_A) with the model that assumes a positive difference $\alpha > 0$ (M_B). Specifically, we assumed a truncated normal prior with a mean of 0 and standard deviation equal to 0.85.[2]

For each participant, we again collected three chains using our product space method, each containing 100,000 samples after removing a burn in of 1000 samples. Once again, we checked convergence using the $\hat{R}$ statistic, and used all of the samples to obtain the log Bayes factor estimate. We also used the samples for the α to estimate the log Bayes Factor using the Savage-Dickey method. In Figure 8a, the consistency of our estimates with the Savage-Dickey estimates is shown. Figure 8b plots the log Bayes Factor estimates as a function of the difference of the proportions of correct identifications in both conditions $K_i^{SB}/N_i^{SB} - K_i^{SN}/N_i^{SN}$. As expected, they are positively related, showing that the greater the difference in proportions of correct identifications, the more support for an improved discriminability effect.

The results in Figure 8b support only cautious conclusions about the effect of priming. More than half of the participants have log Bayes Factors with magnitude less than one (i.e., they fall in the weak support region), suggesting that the experiment does not provide conclusive evidence for or against priming for many participants. There are only two participants whose behavior provides strong evidence in favor of an improvement in primed discrimination

[2]This standard deviation is obtained when simulating $\alpha = \phi^{SB} - \phi^{SN}$, assuming a truncated standard normal distribution for ϕ^{SB} and ϕ^{SN}.

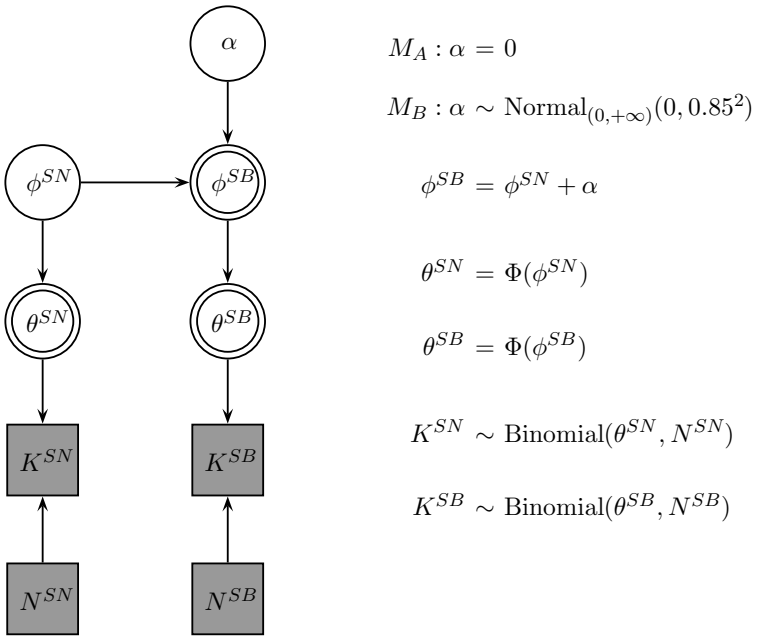


Figure 7. Graphical model for the individual participant Bayes Factor analysis of the Zeelenberg et al. (2002) data.

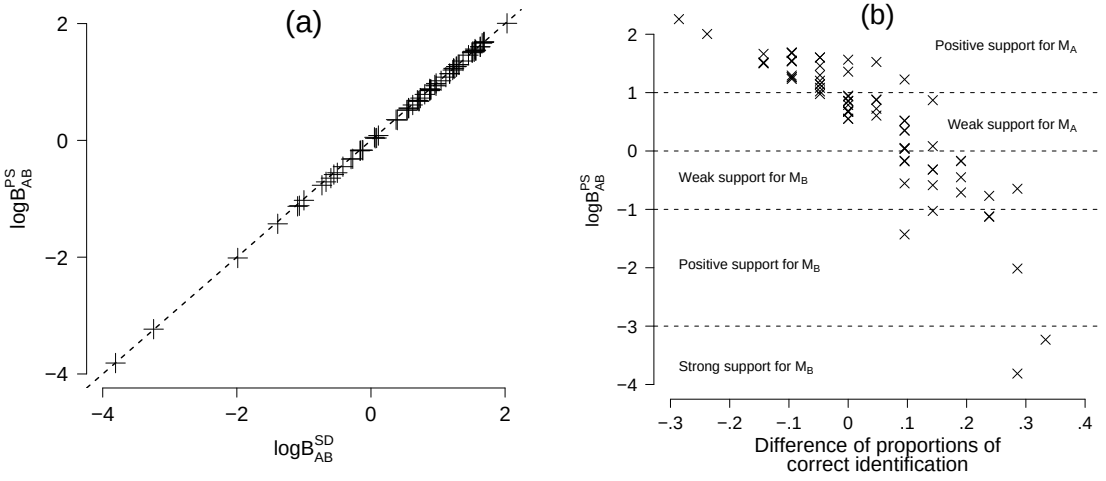


Figure 8. (a) Comparison of $\log B^{PS}_{AB}$ with the validation estimates $\log B^{SD}_{AB}$ for all 74 participants; (b) $\log B^{PS}_{AB}$ for the 74 participants as a function of the difference in proportions of correct identifications, with dashed lines indicating bounds for interpretation, as detailed in Table 1.

performance. Thus, while the Bayes Factors are not inconsistent with the original conclusion of Zeelenberg et al. (2002), they suggest a more nuanced understanding, in which there are individual differences between participants, and the experiment itself does not provide a decisive test of the research question.

Group Bayes Factor Analysis

Our first analysis focused on comparing the possible differences in discrimination at the level of the individual participants. In some ways, this is a different focus of analysis from that used by Zeelenberg et al. (2002), who focused on a group comparison. Accordingly, we undertook a second analysis using the Bayes Factor to test for a difference at the group level. This change in emphasis from the individual to group level requires the testing of hierarchical models within the Bayesian setting, and allows us to demonstrate the applicability of our method to this very general class of models.

The graphical model for this hierarchical group-level analysis is shown in Figure 9.[3] It assumes that the ϕ_i^{SN} and α_i parameters for the ith participant are drawn from a common group-level Normal distribution with means and standard deviation μ_ϕ, μ_α, σ_ϕ and σ_α. At this group level, the standardized mean of the proportion differences between the conditions is defined as $\delta = \mu_\alpha/\sigma_\alpha$. The group level Bayes Factor compares the models $\delta = 0$ (M_A) with $\delta > 0$ (M_B).

This Bayes Factor examines all of the participants simultaneously, and asks whether there is evidence for an increase in successful discrimination for stimuli that were previously studied over the group as a whole. In this sense, it is more consistent with the t-test reported by Zeelenberg et al. (2002). It is not simply a Bayesian surrogate for this test, however, because the hierarchical model in Figure 9 allows for individual participant differences in determining the group effect.

We estimated the log Bayes factor ten times using our product space method, and for the Savage-Dickey method, using the same MCMC settings as used previously. The mean of the estimated log Bayes Factor was 1.46 ($SD = .26$) for the product space method and 1.46 ($SD = .03$) for the Savage-Dickey method. We interpret this as positive support for the model that asserts an increased discrimination ability for primed stimuli, consistent with the general conclusion of Zeelenberg et al. (2002).

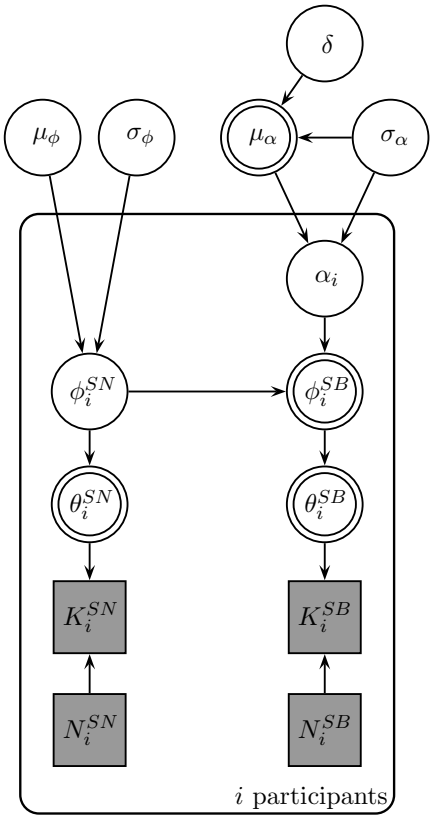
Bayes Factors For Other Model Comparisons

To this point, we have been able to validate our product space method for estimating Bayes Factors against the alternative Savage-Dickey method. This has been possible because all of the Bayes Factors have involved comparing a more complicated model, which allows for a parameter to be in a broad range, to a simpler model that requires the parameter to take a single value within that range. Under these circumstances, the Savage-Dickey method can be applied.

Many research questions, however, are best addressed by Bayes Factors involving models that do not have this special relationship. A key advantage of our product space approach is that it can be applied to more general model comparisons.

For example, one reasonable question might be whether or not the effect of priming on visual discriminability at the group level is large for the Zeelenberg et al. (2002) data. One

[3] This model is also presented in Wagenmakers, Lodewyckx, Kuriyal, and Grasman (submitted), where it is used to illustrate the Savage-Dickey method.



$$
\begin{aligned}
M_A : \delta &= 0 \\
M_B : \delta &\sim \text{Normal}_{(0,+\infty)}(0, 1) \\
\mu_\alpha &= \delta \times \sigma_\alpha \\
\sigma_\alpha &\sim \text{Uniform}(0, 10) \\
\mu_\phi &\sim \text{Normal}_{(0,+\infty)}(0, 1) \\
\sigma_\phi &\sim \text{Uniform}(0, 10) \\
\alpha_i &\sim \text{Normal}(\mu_\alpha, \sigma_\alpha^2) \\
\phi_i^{SB} &= \phi_i^{SN} + \alpha_i \\
\phi_i^{SN} &\sim \text{Normal}(\mu_\phi, \sigma_\phi^2) \\
\theta_i^{SB} &= \Phi(\phi_i^{SB}) \\
\theta_i^{SN} &= \Phi(\phi_i^{SN}) \\
K_i^{SB} &\sim \text{Binomial}(\theta_i^{SB}, N_i^{SB}) \\
K_i^{SN} &\sim \text{Binomial}(\theta_i^{SN}, N_i^{SN})
\end{aligned}
$$

Figure 9. Graphical model for the group Bayes Factor analysis of the Zeelenberg et al. (2002) data.

way of formalizing this question in terms of the Bayes Factor is to compare the model with prior $0 < \delta < 1$, corresponding to the assumption that there is a small effect, against the model with prior $\delta > 1$, corresponding to the assumption that there is a large effect. Note that these models are not nested, and so the Savage-Dickey method cannot be applied. [4] Applying our method, using the same MCMC settings as before, the log Bayes Factor is able to be estimated as 1.37 ($SD = 0.30$) in favor of the first model. This provides evidence that there is more likely a small than a large effect.

Another interesting model comparison tests the model with prior $\delta > 0$, corresponding to the assumption that there is some effect (of any magnitude), against the model with prior $\delta > 0.5$, corresponding to the assumption the effect is at least moderately large. While these hypotheses are nested, it is not the case that one is simply a point hypothesis within the other, and so the Savage-Dickey method again cannot be applied. Using our method, the log Bayes Factor is estimated as -0.15 ($SD = 0.14$), meaning that the data provide little evidence in favor of either model. It is just as likely an assumed effect has any magnitude as it has a magnitude greater than 0.5.

These two additional model comparisons are intended as illustrative examples, rather

[4] We do note, however, that this particular comparison could also be done using the encompassing priors method developed by Klugkist, Kato and Hoijtink (2005; see also Wetzels, Grasman, & Wagenmakers, submitted).

than conclusive analyses. They demonstrate the range of Bayes Factor comparisons that can be estimated using our method, including those involving non-nested hypotheses. It should be clear that, with this sort of flexibility in defining candidate models for comparison, carefully tailored research hypotheses can be evaluated against available data for almost any research question.

Discussion

In Bayesian statistics, the Bayes Factor is one of the most important and widely-used methods for the quantitative evaluation of hypotheses and models. Bayes Factors have an important role to play in the psychological sciences, which regularly seeks to test statistical hypotheses and substantive psychological models. We have developed, demonstrated, and validated where possible a general computational method, based on the original work of Carlin and Chib (1995), for estimating Bayes Factors. Our method can be applied to any statistical hypotheses or psychological model than can be expressed as a probabilistic graphical model, and can be applied to non-nested comparisons and hierarchical models.

An attractive feature of our method is its conceptual simplicity. Like all transdimensional MCMC methods, the basic approach is to estimate the posterior distribution of a model indicator variable that controls which model generates predictions about data. This variable directly corresponds to our intuitions about model selection, because it solves the problem of finding the model that best accounts for the available data. Each model begins with some prior probability of being true, and is updated to a posterior probability, based on the evidence provided by data. These prior and posterior probabilities are represented explicitly by the model indicator variable, and the change in their ratio from the prior to posterior captures the meaning of the Bayes Factor in a very direct way.

Like all practical methods, our implementation works better in some circumstances than others. In our worked illustrative examples, we noted that the method can be hard to apply when the data provide overwhelming evidence for one hypothesis or model over its competitor. This does not seem like a serious limitation, because in these circumstances formal model selection measures do not contribute much to our understanding about the appropriate answers to the motivating research questions.

Overall, we think our method occupies a useful niche between alternative approaches, based on a trade-off between ease-of-implementation and generality of application. The Savage-Dickey method we used for validation is relatively easy to implement, but, as we demonstrated in our final example, is only applicable to a restricted class of comparisons involving point-nested hypotheses or models. More powerful and general transdimensional MCMC methods like Reversible Jump, on the other hand, have the advantage of being able to compare more than two hypotheses or models simultaneously, but are much more difficult to implement. Very often in the psychological sciences, it suffices to compare only a few alternative formal models against available data, but these models will often not be point-nested. In these circumstances, we believe our method provides a relatively powerful and easy-to-use approach for estimating Bayes Factors, and quantifying the evidence the data provide for and against the competing hypotheses or models.

Acknowledgments

Correspondence should be addressed to: Tom Lodewyckx, Department of Psychology, Tiensestraat 102, 3000 Leuven, Belgium. Email: tom.lodewyckx@psy.kuleuven.be. Tel: +32

16 326052. MDL acknowledges the support of a Visiting Fellowship from the University of Leuven. EJW acknowledges the support of a Vidi grant from the Dutch Organization for Scientific Research (NWO).

References

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

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

DiCiccio, T. J., Kass, R. E., Raftery, A. E., & Wasserman, L. (1997). Computing Bayes factors by combining simulation and asymptotic approximations. *Journal of the American Statistical Association*, *92*, 903–915.

Dickey, J. M., & Lientz, B. P. (1970). The weighted likelihood ratio, sharp hypotheses about chances, the order of a Markov chain. *The Annals of Mathematical Statistics*, *41*, 214–226.

Donkin, C., Averell, L., Brown, S. D., & Heathcote, A. J. (in press). Getting more from accuracy and response time data: Methods for fitting the linear ballistic accumulator. *Behavior Research Methods*.

Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review*, *116*(2), 439–453.

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

Geweke, J. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica*, *57*, 1317–1339.

Gilks, W. R., Thomas, A., & Spiegelhalter, D. J. (1994). A language and program for complex Bayesian modelling. *The Statistician*, *43*, 169–177.

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

Hoijtink, H. (2001). Confirmatory latent class analysis: Model selection using Bayes factors and (pseudo) likelihood ratio statistics. *Multivariate Behavioral Research*, *36*, 563-588.

Jeffreys, H. (1961). *Theory of probability*. Oxford, UK: Oxford University Press.

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

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

Kemp, C., Shafto, P., Berke, A., & Tenenbaum, J. B. (2007). Combining causal and similarity-based reasoning. In *Advances in neural information processing systems 19*. Cambridge, MA: MIT Press.

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences*, *105*(31), 10687–10692.

Klugkist, I., Kato, B., & Hoijtink, H. (2005). Bayesian model selection using encompassing priors. *Statistica Neerlandica*, *59*, 57–69.

Koller, D., Friedman, N., Getoor, L., & Taskar, B. (2007). Graphical models in a nutshell. In L. Getoor & B. Taskar (Eds.), *Introduction to statistical relational learning*. Cambridge, MA: MIT Press.

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

Lee, M. D. (2002). Generating additive clustering models with limited stochastic complexity. *Journal of Classification*, *19*(1), 69–85.

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

Lee, M. D. (2006). A hierarchical Bayesian model of human decision-making on an optimal stopping problem. *Cognitive Science*, *30*, 555–580.

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

Lee, M. D., & Wagenmakers, E.-J. (2005). Bayesian statistical inference in psychology: Comment on Trafimow (2003). *Psychological Review*, *112*, 662–668.

Lee, M. D., & Wagenmakers, E.-J. (2009). *A Course in Bayesian Graphical Modeling for Cognitive Science.* Unpublished course notes, University of California Irvine. [http://www.socsci.uci.edu/∼mdlee/bgm].

Lewis, S. M., & Raftery, A. E. (1997). Estimating Bayes factors via posterior simulation with the Laplace-Metropolis estimator. *Journal of the American Statistical Association*, *92*, 648–655.

Lunn, D. J., Spiegelhalter, D., Thomas, A., & Best, N. (in press). The BUGS project: Evolution, critique and future directions. *Statistics in Medicine*.

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

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

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

Raftery, A. E. (1995). Bayesian model selection in social research. In P. V. Marsden (Ed.), *Sociological methodology* (pp. 111–196). Cambridge: Blackwells.

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

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

Rouder, J. N., Lu, J., Speckman, P. L., Sun, D., Morey, R. D., & Naveh-Benjamin, M. (2007). Signal detection models with random participant and item effects. *Psychometrika*, *72*, 621–642.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Pratte, M. S. (2007). Detecting chance: A solution to the null sensitivity problem in subliminal priming. *Psychonomic Bulletin & Review*, *14*, 597–605.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t-tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*, 225–237.

Shiffrin, R. M., Lee, M. D., Kim, W.-J., & Wagenmakers, E.-J. (2008). A survey of model evaluation approaches with a tutorial on hierarchical Bayesian methods. *Cognitive Science*, *32*(8), 1248–1284.

Spiegelhalter, D. J., Thomas, A., Best, N. G., & Gilks, W. R. (1996). *BUGS Examples Volume 1, Version 0.5.* Cambridge, UK: MRC Biostatistics Unit.

Steyvers, M., Lee, M. D., & Wagenmakers, E. (2009). A Bayesian analysis of human decision-making on bandit problems. *Journal of Mathematical Psychology*, *53*, 168–179.

Tierney, L., & Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, *81*, 82–86.

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

Vickers, D., Lee, M. D., Dry, M., & Hughes, P. (2003). The roles of the convex hull and number of intersections upon performanceon visually presented traveling salesperson problems. *Memory & Cognition*, *31*(7), 1094–1104.

Wagenmakers, E. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*, 779–804.

Wagenmakers, E., Lodewyckx, T., Kuriyal, H., & Grasman, R. (submitted). Bayesian hypothesis testing for psychologists: A tutorial on the Savage-Dickey procedure. *Manuscript submitted for publication.*

Wetzels, R., Grasman, R. P. P. P., & Wagenmakers, E. (submitted). An encompassing prior generalization of the Savage-Dickey density ratio test. *Manuscript submitted for publication.*

Wetzels, R., Raaijmakers, J. G. W., Jakab, E., & Wagenmakers, E.-J. (2009). How to quantify support for and against the null hypothesis: A flexible WinBUGS implementation of a default Bayesian t-test. *Psychological Bulletin & Review*, *16*, 752–760.

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

Zeelenberg, R., Wagenmakers, E.-J., & Raaijmakers, J. G. W. (2002). Priming in implicit memory tasks: Prior study causes enhanced discrimability, not only bias. *Journal of Experimental Psychology: General*, *131*, 38–47.