



Contents lists available at ScienceDirect

Cognitive Psychology

journal homepage: www.elsevier.com/locate/cogpsych



Bayesian hypothesis testing for psychologists: A tutorial on the Savage–Dickey method

Eric-Jan Wagenmakers^{a,*}, Tom Lodewyckx^b, Himanshu Kuriyal^c,
Raoul Grasman^a

^a University of Amsterdam, Department of Psychology, Roetersstraat 15, 1018 WB Amsterdam, The Netherlands

^b Leuven University, Department of Psychology, Tiensestraat 102, B-3000 Leuven, Belgium

^c Indian Institute of Technology, Kharagpur, India

ARTICLE INFO

Article history:

Accepted 14 December 2009

Available online 12 January 2010

Keywords:

Statistical evidence
Model selection
Bayes factor
Hierarchical modeling
Random effects
Order-restrictions

ABSTRACT

In the field of cognitive psychology, the p -value hypothesis test has established a stranglehold on statistical reporting. This is unfortunate, as the p -value provides at best a rough estimate of the evidence that the data provide for the presence of an experimental effect. An alternative and arguably more appropriate measure of evidence is conveyed by a Bayesian hypothesis test, which prefers the model with the highest average likelihood. One of the main problems with this Bayesian hypothesis test, however, is that it often requires relatively sophisticated numerical methods for its computation. Here we draw attention to the *Savage–Dickey density ratio* method, a method that can be used to compute the result of a Bayesian hypothesis test for nested models and under certain plausible restrictions on the parameter priors. Practical examples demonstrate the method's validity, generality, and flexibility.

© 2009 Elsevier Inc. All rights reserved.

1. Introduction

Inside every Non-Bayesian, there is a Bayesian struggling to get out – Dennis Lindley, as cited in Jaynes (2003).

How do cognitive psychologists analyze their data? Gert Gigerenzer answered this question by invoking the Freudian concept of unconscious conflict between the Superego, the Ego, and the Id (Gigerenzer, 1993, 2004; Gigerenzer, Krauss, & Vitouch, 2004). In Gigerenzer's analogy, the cognitive

* Corresponding author. Fax: +31 20 639 0279.

E-mail address: EJ.Wagenmakers@gmail.com (E.-J. Wagenmakers).

psychologist's Superego wants to follow the Neyman–Pearson tradition; it seeks to contrast two well-defined hypotheses (i.e., the null hypothesis and an alternative hypothesis), it operates using concepts of α -level and power, and it is generally concerned with procedures that will work well in the long run. In contrast, the cognitive psychologist's Ego follows the Fisherian tradition; it does not posit a specific alternative hypothesis, it ignores power, and it computes a p -value that is supposed to indicate the statistical evidence against the null hypothesis. Finally, the cognitive psychologist's Id is *Bayesian*, and it desperately wants to attach probabilities to hypotheses. However, this wish is suppressed by the Superego and Ego. In its continual struggle to obtain what it desires, the Id—although unable to change the statistical analysis procedures that are used—wields its influence to change and distort the interpretations that these analysis procedures afford.¹

The unconscious Freudian conflict has arguably resulted in widespread confusion. Researchers often assume that a small p -value means that the null hypothesis is likely to be false, that a large p -value means that the null hypothesis is likely to be true, and that a 95% confidence interval for a parameter μ means that there is a 95% chance that μ lies in the specified interval. All of these conclusions are false (Haller & Krauss, 2002)—this is because the conclusions are Bayesian, but the methodology that is used is not.

To resolve the unconscious Freudian conflict and bring the statistical procedures in line with their interpretation, two courses of action present themselves. First, one can try to suppress the Id even more strongly, perhaps by rigorous statistical education and repeated warnings such as “Never use the unfortunate expression ‘accept the null-hypothesis’.” (Wilkinson & the Task Force on Statistical Inference, 1999, p. 599). Second, one can explore Bayesian statistical procedures that provide exactly what the Id wants—probabilities for hypotheses. Using Bayesian procedures, one can quantify support both in favor of and against the null hypothesis (Gallistel, 2009; Rouder, Speckman, Sun, Morey, & Iverson, 2009; Wetzels, Raaijmakers, Jakab, & Wagenmakers, 2009), and one can state that the probability that a parameter μ lies in a 95% “credible interval” is, indeed, .95. In this article, we promote the second course of action.

In order to keep this article self-contained, we first provide a brief overview of the Bayesian paradigm, with special emphasis on the difference between parameter estimation and hypothesis testing. We then describe a method, known as the Savage–Dickey density ratio, to carry out a Bayesian hypothesis test with relative ease. Next we illustrate the practical value of the Savage–Dickey method by applying it to three data sets. The first data set is used to test the hypothesis that the sexual behavior of so-called virginity pledgers differs from that of non-pledgers (i.e., a hypothesis test for the equality of two rates, Brückner & Bearman, 2005); the second data set is used to test the hypothesis that prior study of both choice alternatives improves later performance in a two-choice perceptual identification task (i.e., a hypothesis test in a hierarchical within-subjects design, Zeelenberg, Wagenmakers, & Raaijmakers, 2002); and the third data set is used to test the hypothesis that typically developing children outperform children with ADHD on the Wisconsin card sorting test (i.e., a hypothesis test in a hierarchical between-subjects design, Geurts, Verté, Oosterlaan, Roeyers, & Sergeant, 2004).

In these examples, we show how the Bayesian hypothesis test can be adjusted to deal with random effects and order-restrictions, both for within-subjects and between-subjects designs. WinBUGS code is presented in Appendix B and R code is available online.²

2. Bayesian background

Before outlining the Savage–Dickey method, it is important to introduce some key concepts of Bayesian inference. More detailed information can be found in Bayesian articles and books that discuss philosophical foundations (Lindley, 2000; O'Hagan & Forster, 2004), computational innovations (Gelman & Lopes, 2006), and practical contributions (Congdon, 2003; Ntzoufras, 2009). An in-depth discussion on the advantages of Bayesian inference, especially when compared to p -value

¹ For more information about the difference between the three statistical paradigms, see for instance Christensen (2005), Hubbard and Bayarri (2003) and Royall (1997).

² All computer code is available from the first author's website, <http://users.fmg.uva.nl/ewagenmakers/papers.html>.

hypothesis testing, is beyond the scope of this article, and we instead refer the interested reader to Berger and Berry (1988b), Edwards, Lindman, and Savage (1963), Sellke, Bayarri, and Berger (2001), Wagenmakers (2007) and Wagenmakers et al. (2008). Those familiar with Bayesian inference can safely skip to the section that introduces the Savage–Dickey method.

2.1. Bayesian parameter estimation

As is customary, we introduce Bayesian parameter estimation by means of the binomial example. Assume we prepare for you a series of 10 factual true/false questions of equal difficulty. Interest centers on your latent probability θ of answering any one question correctly. In Bayesian inference, uncertainty with respect to parameters is—at any point in time—quantified by probability distributions. Thus, in order to get the Bayesian inference machine off the ground, we need to specify our uncertainty with respect to θ before seeing the data. Suppose you do not know anything about the topic or about the difficulty level of the questions. Then, a reasonable “prior distribution”, denoted by $p(\theta)$, is one that assigns equal probability to every value of θ . This uniform distribution is shown by the dotted horizontal line in Fig. 1.

Now we proceed with the test, and find that you answered 9 out of 10 questions correctly. After having seen these data, our updated knowledge about θ is described by a “posterior distribution”, denoted $p(\theta|s, n)$, where $s = 9$ and $n = 10$ indicate the number of successes and the number of questions, respectively. Assume that the probability of the data is given by the binomial distribution:

$$p(s|\theta, n) = \binom{n}{s} \theta^s (1 - \theta)^{n-s}. \quad (1)$$

The transition from prior $p(\theta)$ to posterior $p(\theta|s, n)$ is then given by Bayes’ rule,

$$p(\theta|s, n) = \frac{p(s|\theta, n)p(\theta)}{p(s|n)}. \quad (2)$$

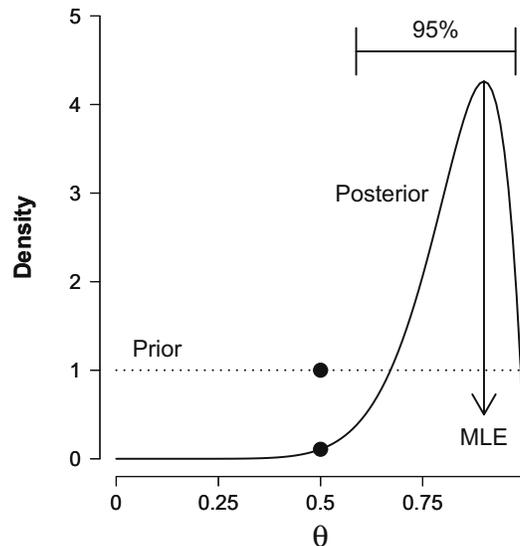


Fig. 1. Bayesian parameter estimation for binomial rate parameter θ , after observing nine correct responses and one incorrect response. The mode of the posterior distribution for θ is 0.9, equal to the maximum likelihood estimate, and the 95% confidence interval extends from 0.59 to 0.98. The two black circles positioned at $\theta = 0.5$ help to illustrate the Savage–Dickey density ratio discussed later.

This equation is often presented as

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{marginal likelihood}}. \quad (3)$$

Note that the marginal likelihood (i.e., the probability of the observed data) does not involve the parameter θ , and is given by a single number that ensures that the area under the posterior distribution equals 1. Therefore, Eq. (2) is often written as

$$p(\theta|s, n) \propto p(s|\theta, n)p(\theta), \quad (4)$$

which says that the posterior is proportional to the likelihood times the prior.

The solid line in Fig. 1 shows the posterior distribution for θ , which is obtained when the uniform prior is updated with data $s = 9$ and $n = 10$. The central tendency of a posterior distribution is often summarized by its mean, median, or mode. Note that with a uniform prior, the mode of a posterior distribution coincides with the classical maximum likelihood estimate or MLE, $\hat{\theta} = s/n = 0.9$ (Myung, 2003). The spread of a posterior distribution is most easily captured by a Bayesian $x\%$ confidence interval that extends from the $(x/2)$ th to the $(100 - x/2)$ th percentile of the posterior distribution. For the posterior distribution in Fig. 1, a 95% Bayesian confidence interval for θ extends from 0.59 to 0.98. In contrast to the classical or orthodox confidence interval, the Bayesian confidence interval has a direct and intuitive interpretation: after observing the data, we can be 95% confident that the true value of θ lies in between 0.59 and 0.98.

Now suppose we design a new set of five questions, of equal difficulty as before. How can we formalize our expectations about your performance on this new set? In other words, how can we use the posterior distribution $p(\theta|n = 10, s = 9)$ —which after all represents everything that we know about θ from the old set—to predict the number of correct responses out of the new set of $n^{\text{new}} = 5$ questions? The mathematical solution is to integrate over the posterior,

$$p(s^{\text{new}}|n^{\text{new}} = 5) = \int_0^1 p(s^{\text{new}}|\theta, n^{\text{new}} = 5)p(\theta|n = 10, s = 9)d\theta, \quad (5)$$

where s^{new} is the predicted number of correct responses out of the additional set of five questions. Computationally, one may think of this procedure as repeatedly drawing a random value θ_i from the posterior, and using that value to every time determine a single s_i^{new} by means of Eq. (1). The end result, $p(s^{\text{new}}|n^{\text{new}} = 5)$, is the predictive density of the possible number of correct responses in the additional set of five questions. The important point is that by integrating over the posterior, all predictive uncertainty is taken into account. In contrast, much of classical inference relies on the “plug-in principle” that in this case would lead us to predict $p(s^{\text{new}}|n^{\text{new}} = 5)$ solely based on $\hat{\theta}$, the maximum likelihood estimate. Plug-in procedures ignore uncertainty in θ , and hence lead to predictions that are overconfident, that is, predictions that are less variable than they should be (Aitchison & Dunsmore, 1975).³

You are now presented with the new set of five questions. You answer 3 out of 5 correctly. How do we combine this new information with the old? Or, in other words, how do we update our knowledge of θ ? Consistent with intuition, Bayes’ rule entails that the prior that should be updated based on your performance for the new set is the posterior that was obtained based on your performance for the old set. Or, as Lindley put it, “today’s posterior is tomorrow’s prior” (Lindley, 1972, p. 2). When all the data have been collected, however, the precise order in which this was done is irrelevant; the results from the 15 questions could have been analyzed as a single batch, they could have been analyzed sequentially, one-by-one, they could have been analyzed by first considering the set of 10 questions and next the set of 5, or vice versa. For all these cases, the end result, the final posterior distribution for θ , is identical. This again contrasts with classical inference, in which inference for sequential designs is different from that for non-sequential designs (for a discussion, see e.g., Anscombe, 1963).

³ It should be acknowledged that classical statisticians can account for uncertainty in the estimation of θ by repeatedly drawing a bootstrap sample from the data, calculating the associated bootstrap MLE, and finding the corresponding prediction for s^{new} (e.g., Wagenmakers, Ratcliff, Gomez, & Iverson, 2004).

Thus, a posterior distribution describes our uncertainty with respect to a parameter of interest, and the posterior is useful—or, as a Bayesian would have it, necessary—for probabilistic prediction and for sequential updating. Unfortunately, the posterior distribution or any of its summary measures can only be obtained in closed form for a restricted set of relatively simple models. To illustrate in the case of our binomial example, the uniform prior is a so-called beta distribution with parameters $\alpha = 1$ and $\beta = 1$, and when combined with the binomial likelihood this yields a posterior that is also a beta distribution, be it with parameters $\alpha + s$ and $\beta + n - s$. In simple *conjugate* cases such as these, where the prior and the posterior belong to the same distributional family, it is possible to obtain closed form solutions for the posterior distribution, but in other more interesting cases it is not.

For a long time, researchers did not know how to proceed with Bayesian inference when the posterior could not be obtained in closed form. As a result, practitioners interested in models of realistic complexity did not much use Bayesian inference. This situation changed with the advent of computer-driven sampling methodology generally known as Markov chain Monte Carlo (i.e., MCMC; e.g., Gamerman & Lopes, 2006; Gilks, Richardson, & Spiegelhalter, 1996). Using MCMC techniques such as Gibbs sampling or the Metropolis–Hastings algorithm, researchers can now directly sample sequences of values from the posterior distribution of interest, foregoing the need for closed form analytic solutions. At the time of writing, the adage is that Bayesian models are limited only by the user's imagination.

To provide a concrete and simple illustration of Bayesian inference using MCMC, we revisit our binomial example of 9 correct responses out of 10 questions, and the associated inference problem for θ , the probability of answering any one question correctly. Throughout this article, we use the general-purpose WinBUGS program (Lunn, Thomas, Best, & Spiegelhalter, 2000; Lunn, Spiegelhalter, Thomas, & Best, 2009; an introduction for psychologists is given by Sheu & O'Curry, 1998) that allows the user to specify and fit models without having to hand-code the MCMC algorithms. Although WinBUGS does not work for every application, it will work for most applications in the field of psychology. The WinBUGS program is easy to learn and is supported by a large community of active researchers.⁴

The WinBUGS program requires the user to construct a file that contains the model specification, a file that contains initial values for the model parameters, and a file that contains the data. The model specification file is most important. For our binomial example, we set out to obtain samples from the prior and the posterior of θ . The associated WinBUGS model specification code is three lines long:

```
model
{
  theta ~ dbeta(1,1) # the uniform prior for updating by the data
  k ~ dbin(theta,n) # the data; in our example, k = 9 and n = 10
  thetaprior ~ dbeta(1,1) # a uniform prior not for updating
}
```

In this code, the “~” or twiddle symbol denotes “is distributed as”, `dbeta(a,b)` indicates the beta distribution with parameters a and b , and `dbin(theta,n)` indicates the binomial distribution with rate θ and n observations. These and many other distributions are built in to the WinBUGS system. The “#” or hash sign is used for commenting out what should not be compiled. As WinBUGS is a declarative language, the order of the three lines is inconsequential.

When this code is executed, the user obtains a sequence of samples (i.e., an MCMC chain) from the posterior $p(\theta|D)$ and a sequence of samples from the prior $p(\theta)$. In more complex models, it may take some time before the chain converges from its starting value to what is called its stationary distribution. To make sure that we only use those samples that come from the stationary distribution (and are hence unaffected by the starting values) it is good practice to discard the first samples as “burn-in”, and to diagnose convergence by running multiple chains.

For instance, Fig. 2 shows the first 100 iterations for three chains that were set up to draw values from the posterior for θ . The three chains are almost indistinguishable, and they do not have slow

⁴ For more information on WinBUGS see <http://www.mrc-bsu.cam.ac.uk/bugs/>.

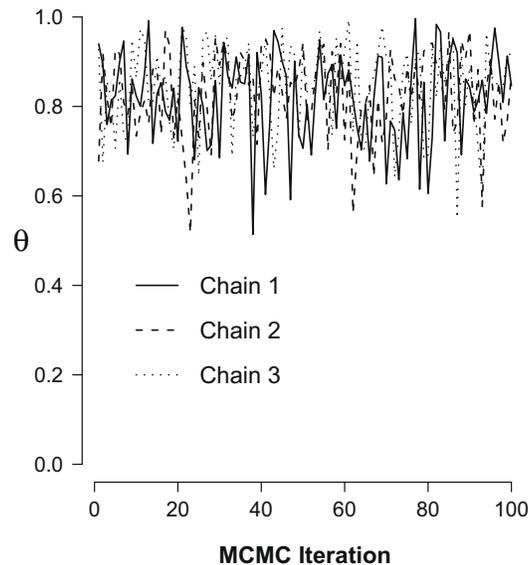


Fig. 2. Three MCMC chains for binomial rate parameter θ , after observing nine correct responses and one incorrect response.

upward or downward drift; these are two qualitative signs that the chains have converged and that samples are being drawn from the posterior distribution. Quantitative measures for diagnosing convergence are also available (e.g., the Gelman and Rubin (1992) \hat{R} statistic, that compares within-chain to between-chain variability; for more recommendations regarding convergence see e.g., Gamerman and Lopes (2006), Gelman (1996), and Gelman and Hill (2007)).

After assuring ourselves that the chains have converged, we can use the sampled values to plot a histogram, construct a density estimate, and compute values of interest. To illustrate, the three chains from Fig. 2 were run for 3000 iterations each, for a total of 9000 samples for the prior and the posterior of θ . Fig. 3 plots histograms⁵ for the prior (i.e., dotted line) and the posterior (i.e., thick solid line). In addition, the thin solid lines represent logspline nonparametric density estimates (Stone, Hansen, Kooperberg, & Truong, 1997). The mode of the logspline density estimate for the posterior of θ is 0.89, whereas the 95% confidence interval is (0.59, 0.98), matching the analytical result shown in Fig. 1.

Of course, this example represents an ideal scenario; in more complicated models, convergence might be obtained only after many MCMC iterations—that is, chains may move very slowly from their starting point to the stationary distribution. This problem is often easy to diagnose by running multiple chains with overdispersed starting values. Another problem is that, even when the chains have arrived at the posterior distribution, consecutive samples might be highly correlated. This is less worrisome than the problem of nonconvergence (after all, the samples are draws from the correct posterior distribution), but it does mean that more samples need to be collected before the entire posterior is adequately covered. This problem is easy to diagnose by computing the autocorrelation of the chains. A relatively high autocorrelation suggests that we need to draw relatively many samples. Thus, for complex models it is important to use MCMC algorithms that are efficient, reliable, and quick. This is currently an active area of research. Nevertheless, the fundamental theoretical obstacles for Bayesian parameter estimation have been overcome. In fields such as statistics, artificial intelligence, and machine learning, MCMC algorithms are now used on a routine basis.

⁵ These histograms were constructed such that the total area under each histogram equals one.

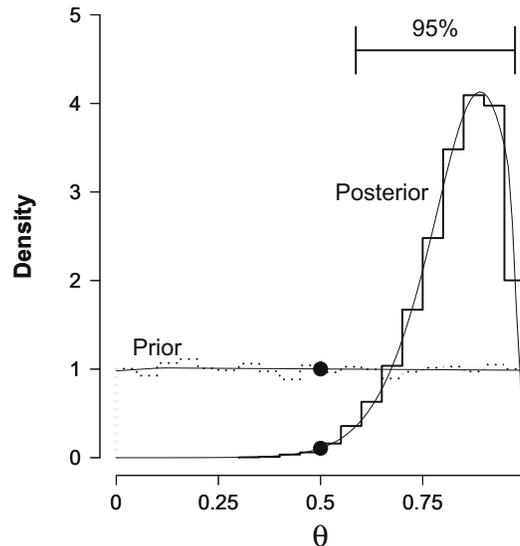


Fig. 3. MCMC-based Bayesian parameter estimation for binomial rate parameter θ , after observing nine correct responses and one incorrect response. The thin solid lines indicate the fit of a nonparametric logspline density estimator. Based on this density estimator, the mode of the posterior distribution for θ is approximately 0.89, and the 95% confidence interval extends from 0.59 to 0.98, closely matching the analytical results from Fig. 1. The two black circles positioned at $\theta = 0.5$ again help to illustrate the Savage–Dickey density ratio discussed later.

2.2. Bayesian hypothesis testing

Up to this point we have concerned ourselves with parameter estimation, implicitly taking the appropriateness of the underlying model for granted. In much of social science, however, researchers entertain more than just a single statistical model. In fact, the statistical models often represent competing theories or hypotheses, and the focus of interest is on which substantive theory or hypothesis is more correct, more plausible, and better supported by the data. For example, researchers might want to know whether the improvement of performance with practice follows a power function or an exponential function. As another example, we might want to know the extent to which your performance in our test (i.e., 9 correct answers out of 10 questions) is consistent with the hypothesis that you were just guessing. This may involve a test of $M_1 : \theta = 0.5$ versus $M_2 : \theta \neq 0.5$.

The fundamental and general Bayesian solution to the foregoing model selection of hypothesis testing problems is as follows. For simplicity, assume that you contemplate two alternative accounts of the data, M_1 and M_2 , and that you seek to quantify model uncertainty in terms of probability. Consider first M_1 . Bayes' rule dictates how your prior probability of M_1 , $p(M_1)$, is updated through the observed data D to give the posterior probability of M_1 , $p(M_1|D)$:

$$p(M_1|D) = \frac{p(M_1)p(D|M_1)}{p(M_1)p(D|M_1) + p(M_2)p(D|M_2)}. \quad (6)$$

In the same way, one can calculate the posterior probability of M_2 , $p(M_2|D)$. The ratio of these posterior probabilities is given by

$$\frac{p(M_1|D)}{p(M_2|D)} = \frac{p(M_1)}{p(M_2)} \frac{p(D|M_1)}{p(D|M_2)}. \quad (7)$$

This equation shows that the change from prior odds $p(M_1)/p(M_2)$ to posterior odds $p(M_1|D)/p(M_2|D)$ is determined entirely by the ratio of the marginal likelihoods $p(D|M_1)/p(D|M_2)$. This ratio is generally

known as the *Bayes factor* (Jeffreys, 1961), and the Bayes factor, or the log of it, is often interpreted as the *weight of evidence* coming from the data (Good, 1985).

A hypothesis test based on the Bayes factor supports the model under which the observed data are most likely (for details see Berger & Pericchi, 1996; Bernardo & Smith, 1994, chap. 6; Klugkist, Laudy, & Hoijtink, 2002, cha 7; Klugkist et al., 2005a; Kass & Raftery, 1995; MacKay, 2003; Myung & Pitt, 1997; O'Hagan, 1995). Therefore, the Bayes factor represents “the standard Bayesian solution to the hypothesis testing and model selection problems” (Lewis & Raftery, 1997, p. 648); in the following, we will use “Bayesian hypothesis test” as a shortcut for “a hypothesis test based on the Bayes factor”.

Thus, when the Bayes factor for M_1 versus M_2 equals 2 (i.e., $BF_{12} = 2$), this means that the data are twice as likely to have occurred under M_1 than under model M_2 . When the prior odds are equal, such that M_1 and M_2 are equally likely a priori, the Bayes factors can be converted to posterior probabilities: $p(M_1|D) = BF_{12}/(BF_{12} + 1)$. This means that $BF_{12} = 2$ translates to $p(M_1|D) = 2/3$.

To illustrate, consider again our binomial example of 9 correct responses out of 10 questions, and the test between two models for performance: guessing (i.e., $M_1 : \theta = 0.5$) versus not guessing (i.e., $M_2 : \theta \neq 0.5$). From Eq. (1), the marginal likelihood for M_1 , $p(D|M_1)$, is simply $\binom{10}{9} (\frac{1}{2})^{10}$. The marginal likelihood for model M_2 is more difficult to calculate, as θ is a free parameter. In general, the marginal likelihood is obtained by integrating out the model parameters in accordance with the law of total probability:

$$p(D|M_2) = \int p(D|\theta, M_2)p(\theta|M_2)d\theta. \quad (8)$$

This means that the marginal likelihood is computed by averaging the likelihood over the prior; conceptually, the likelihood is evaluated for every possible parameter value, weighted with its prior plausibility, and added to a summed total. When we again use the uniform distribution for θ as a prior, such that $p(\theta|M_2) \sim \text{Beta}(1, 1)$, then Eq. (8) famously simplifies to $p(D|M_2) = 1/(n + 1)$. Thus, in our binomial example, $BF_{12} = \binom{10}{9} (\frac{1}{2})^{10} (n + 1) \approx 0.107$. This means that the data are $1/0.107 \approx 9.3$ times more likely under M_2 than they are under M_1 . With unit prior odds, the posterior probability for M_1 is $0.107/(0.107 + 1) \approx .10$, which means that the complementary posterior probability for M_2 is approximately .90. These are probabilities assigned to hypotheses, and they are exactly what researchers (or, in Gigerenzer's Freudian analogy, their Ids) want to know about.

Posterior model probabilities are not just necessary to quantify our degree of belief or preference for the candidate models under consideration. They are also necessary for Bayesian model averaging (e.g., Draper, 1995; Hoeting, Madigan, Raftery, & Volinsky, 1999; Madigan & Raftery, 1994). For instance, in a regression context we might have one model, M_1 , that predicts a certain post-surgery survival rate by gender, age, weight, and history of smoking. A second model, M_2 , includes two additional predictors, namely body-mass index and fitness. We compute posterior model probabilities and find that $p(M_1|D) = .6$ and consequently $p(M_2|D) = .4$. For a given patient, M_1 predicts a survival rate of 90%, and M_2 predicts a survival rate of 80%. What is our best prediction for our patient's survival rate? It is tempting to base our prediction solely on M_1 , which is after all the preferred model. However, this would ignore the uncertainty inherent in the model selection procedure, and it would ignore the very real possibility that the best model is M_2 , according to which the survival rate is 10% lower than it is for M_1 . The Bayesian solution is to weight the two competing predictions with their associated posterior model probabilities, fully taking into account the uncertainty in the model selection procedure. In our example, the model-averaged prediction for survival rate would be $.6 \times 90\% + .4 \times 80\% = 86\%$.

2.2.1. Additional advantages of Bayesian hypothesis testing

We have seen how Bayes factors and posterior model probabilities describe the relative support or preference for a set of candidate models, and how they can be used for model averaged predictions. Other advantages of Bayesian hypothesis testing include the following (Wagenmakers, Lee, Lode-wyckx, & Iverson, 2008; see also Berger & Pericchi, 2001; Dennis, Lee, & Kinnell, 2008; Kass & Raftery, 1995):

1. *Coherence is guaranteed.* Suppose we have a set of three candidate models, M_1 , M_2 , and M_3 . As

$$\frac{p(D|M_1)}{p(D|M_3)} = \frac{p(D|M_1)}{p(D|M_2)} \frac{p(D|M_2)}{p(D|M_3)}, \quad (9)$$

this means that $BF_{13} = BF_{12} \times BF_{23}$. For instance, when the data are five times as likely to occur under M_1 than under M_2 , and seven times as likely under M_2 than under M_3 , it follows that the data are $5 \times 7 = 35$ times as likely under M_1 than under M_3 . No comparable result exists in classical statistics.

2. *Parsimony is automatically rewarded.* The main challenge of hypothesis testing or model selection is to identify the model with the best predictive performance (e.g., Myung, Forster, & Browne, 2000; Wagenmakers & Waldorp, 2006). However, it is not immediately obvious how this should be done; complex models will generally provide a better fit to the observed data than simple models, and therefore one cannot simply prefer the model with the best “goodness-of-fit”—such a strategy would lead to overfitting. Intuition suggests that this tendency for overfitting should be counteracted by putting a premium on simplicity. This intuition is consistent with the law of parsimony or “Ockham’s razor” which states that, when everything else is equal, simple models are to be preferred over complex models (Jaynes, 2003, chap. 20; Myung & Pitt, 1997).

Formal model selection methods try to quantify the tradeoff between goodness-of-fit and parsimony. Many of these methods measure a model’s overall performance by the sum of two components, one that measures descriptive accuracy and one that places a premium on parsimony. The latter component is also known as the Ockham factor (MacKay, 2003, chap. 28). For many model selection methods, the crucial issue is how to determine the Ockham factor. One of the attractive features of Bayesian hypothesis testing is that it automatically determines the model with the best predictive performance – Bayesian hypothesis testing therefore incorporates what is known as an automatic Ockham’s razor (Myung & Pitt, 1997).

To see why this is the case, consider that every statistical model makes *a priori* predictions. Complex models have a relatively large parameter space, and are therefore able to make many more predictions and cover many more eventualities than simple models. However, the drawback for complex models is that they need to spread out their prior probability across their entire parameter space. In the limit, a model that predicts almost everything has to spread out its prior probability so thinly that the occurrence of any particular event will not greatly add to that model’s credibility. As shown by Eq. (8), the marginal likelihood for a model M is calculated by averaging the likelihood $p(D|\theta, M)$ over the prior $p(\theta|M)$. When the prior is very spread out, it will occupy a relatively large part of the parameter space in which the likelihood is almost zero, and this greatly decreases the average or marginal likelihood. Consider for instance the situation shown in Fig. 1. The prior on the rate parameter θ was assumed to be uniform from 0 to 1, $\theta \sim U[0, 1]$. A different, more parsimonious model could take into account the prior knowledge that θ is unlikely to be lower than 0.5, as this would mean that your ability would be lower than chance (recall that the questions were true/false, such that the absence of any knowledge corresponds to $\theta = 0.5$). This more informed model could be instantiated as $\theta \sim U[0.5, 1]$, and we could then use the Bayes factor to compute the relative plausibility of model $M_1 : \theta \sim U[0, 1]$ versus $M_2 : \theta \sim U[0.5, 1]$. The more complex model M_1 kept its options open by assigning half of its prior mass to values for θ that are smaller than 0.5. This could have been advantageous when the data would have turned out differently (e.g., 1 correct answer instead of 9). As it is, the values of θ that are smaller than 0.5 are very unlikely; hence, the average likelihood is almost twice as high for the parsimonious model M_2 than it is for the more complex model M_1 .

3. *Evidence can be obtained in favor of the null hypothesis.* Bayesian hypothesis testing allows one to obtain evidence in favor of the null hypothesis. Because theories and models often predict the absence of a difference, it is vital for scientific progress to be able to quantify evidence in favor of the null hypothesis (e.g., Gallistel, 2009; Rouder et al., 2009; Wetzels et al., 2009). In the field of visual word recognition, for instance, the entry-opening theory (Forster, Mohan, & Hector, 2003) predicts that masked priming is absent for items that do not have a lexical representation;

Another example from that literature concerns the work by Bowers, Vigliocco, and Haan (1998), who have argued that priming effects are equally large for words that look the same in lower and upper case (e.g., kiss/KISS) or that look different (e.g., edge/EDGE), a finding supportive of the hypothesis that priming depends on abstract letter identities.

A final example comes from the field of recognition memory, where Dennis and Humphreys' *bind cue decide model of episodic memory* (BCDMEM) predicts the absence of a list-length effect and the absence of a list-strength effect (Dennis & Humphreys, 2001). This radical prediction of a null effect allows researchers to distinguish between context-noise and item-noise theories of inference in memory (Dennis et al., 2008). In Bayesian statistics, the null hypothesis has no special status, and evidence for it is quantified just as it is for any other hypothesis. In classical statistics, support for informative predictions from null hypothesis can only be indirect.

4. *Evidence may be monitored as it accumulates.* Bayesian hypothesis testing allows one to monitor the evidence as the data come in (Berger & Berry, 1988a). In contrast to frequentist inference, Bayesian inference does not require special corrections for “optional stopping” (Wagenmakers, 2007).

Consider, for instance, a hypothetical experiment on the neural substrate of dissociative identity disorder. In this experiment, the researcher Lisa has decided in advance to use functional magnetic resonance imaging (fMRI) to test 30 patients and 30 normal controls. Lisa inspects the data after 15 participants in each group have been tested, and finds that the results convincingly demonstrate the pattern she hoped to find. Unfortunately for Lisa, she cannot stop the experiment and claim a significant result, as she would be changing the sampling plan halfway through and be guilty of “optional stopping”. She has to continue the experiment, wasting not just her time and money, but also the time and efforts of the people who undergo needless testing.

In contrast, for Bayesian hypothesis testing there is nothing wrong with gathering more data, examining these data, and then deciding whether or not to stop collecting new data – no special corrections are needed. As stated by Edwards et al. (1963), “(· · ·) the rules governing when data collection stops are irrelevant to data interpretation. It is entirely appropriate to collect data until a point has been proven or disproven, or until the data collector runs out of time, money, or patience.” (Edwards et al., 1963, p. 193).

2.2.2. Challenges for Bayesian hypothesis testing

Bayesian hypothesis testing using Bayes factors (Eq. (7)) faces two main challenges, one conceptual and one computational. The conceptual challenge is that the Bayesian hypothesis test is acutely sensitive to the specification of the prior distributions for the model parameters (e.g., Bartlett, 1957; Liu & Aitkin, 2008). This distinguishes hypothesis testing from parameter estimation, in which the data quickly overwhelm the prior; the accumulation of data forces prior opinions that are very different to converge to posterior opinions that are very similar. For parameter estimation then, the choice of a prior distribution is not really all that critical unless there are very few data points.

In contrast, for Bayesian hypothesis testing the prior distributions are crucial and have a lasting impact. This occurs because the marginal likelihood is an average taken with respect to the prior. Consider for instance the prior for the mean μ of a Normal distribution with known variance. One might be tempted to use an “uninformative” prior, one that does not express much preference for one value of μ over the other. One such vague prior could be a Normal distribution with mean zero and variance 10,000. But, from a marginal likelihood perspective, this prior is consistent with almost any value of μ . When one hedges one's bets to such an extreme degree, the Bayes factor is likely to show a preference for a simple model (e.g., one in which $\mu = 0$), even when the data appear wildly inconsistent with it.

The main problem here is not that the Bayesian hypothesis test corrects for model complexity as manifested in the prior distribution. This is the automatic Ockam's razor that is an asset, not a liability, of the Bayesian hypothesis test. Instead, the problem seems to be that researchers have only a vague idea of the vagueness of their prior knowledge, or that researchers seek to use a prior that is “objective”, and uses as little prior knowledge as possible. When the vagueness of the prior is arbitrary, so are the results from the Bayesian hypothesis test. When the vagueness of the prior is purposefully

large, the results from the Bayesian hypothesis test tend to indicate a preference for the simple model, regardless of the data.

In order to increase the robustness of Bayesian hypothesis testing to the vagueness of the prior, several procedures have been proposed, including the local Bayes factor (Smith & Spiegelhalter, 1980), the intrinsic Bayes factor (Berger & Mortera, 1999; Berger & Pericchi, 1996), the fractional Bayes factor (O'Hagan, 1995), and the partial Bayes factor (O'Hagan, 1995; for a summary see Gill, 2002, chap. 7). The idea of the partial Bayes factor is to sacrifice a small part of the data to obtain a posterior that is robust to the various priors one might entertain. The Bayes factor is then calculated by integrating the likelihood over this posterior instead of over the original prior. Procedures such as these are still undergoing further development and deserve more study.

The problem of vague priors is particularly evident for parameters that can take on values across the entire real line, such as the mean μ of a Normal distribution. We believe that in such cases, whenever possible, the construction of a prior should be guided by the substantive knowledge in the domain of application. As Dennis Lindley has pointed out repeatedly, μ is only a Greek letter, an abstraction that may obscure the fact that it refers to something about which we have detailed prior knowledge. When μ stands for a person's weight, few rational people would assign μ an "uninformative" Normal prior distribution with mean zero and variance 10,000.

In this paper, we sidestep this conceptual challenge to some extent, as we focus completely on discrete data problems (i.e., those that involve a hit or a miss, a success or a failure, a yes or a no). In such cases, a perfectly plausible prior assigns equal mass to every value of the underlying rate parameter θ . In some cases, we use order-restrictions and assign equal mass to every value of θ greater than .5. We feel that in the absence of detailed prior knowledge, this assumption is reasonable. Note, however, that our approach in this paper is entirely general; when you are willing to defend and use a different prior, you are free to do so.

The second challenge for Bayesian hypothesis testing—the one that is the focus of this article—is that the marginal likelihood and the Bayes factor are often quite difficult to compute. Earlier, we saw that with a uniform prior on the binomial rate parameter θ (i.e., $p(\theta|M) \sim \text{Beta}(1, 1)$), the marginal likelihood $\int p(D|\theta, M)p(\theta|M)d\theta$ simplifies to $1/(1+n)$. However, in all but a few simple models, such simplifications are impossible. In order to be able to compute the marginal likelihood or the Bayes factor for more complex models, a series of different computational methods has been developed. A recent summary lists as many as 15 different methods (Gamerman & Lopes, 2006, chap. 7).

For instance, one method computes the marginal likelihood through what is called the *candidates' formula* (Besag, 1989) or the *basic marginal likelihood identity* (Chib, 1995; Chib & Jeliazkov, 2001). One simply exchanges the roles of posterior and marginal likelihood to obtain

$$p(D) = \frac{p(D|\theta)p(\theta)}{p(\theta|D)}, \quad (10)$$

which holds for any value of θ . When the posterior is available analytically, one only needs to plug in a single value of θ and obtain the marginal likelihood immediately. This method can however also be applied when the posterior is only available through MCMC output, either from the Gibbs sampler (Chib, 1995) or the Metropolis–Hastings algorithm (Chib & Jeliazkov, 2001).

Another method to compute the marginal likelihood is to repeatedly sample parameter values from the prior, calculate the associated likelihoods, and then take the likelihood average. When the posterior is highly peaked compared to the prior—as will happen with many data or with a medium-sized parameter space—it becomes necessary to employ more efficient sampling methods, with a concomitant increase in computational complexity.

Finally, it is also possible to compute the Bayes factor directly, without first calculating the constituent marginal likelihoods. The basic idea is to generalize the MCMC sampling routines for parameter estimation to incorporate a "model indicator" variable. In the case of two competing models, the model indicator variable k , say, can take on two values—for instance, $k = 1$ when the sampler is in model M_1 , and $k = 2$ when the sampler is in model M_2 . The Bayes factor is then estimated by the relative frequency with which $k = 1$ versus $k = 2$. This MCMC approach to model selection

is called transdimensional MCMC (e.g., [Sisson, 2005](#)), an approach that encompasses both reversible jump MCMC [Green, 1995](#) and the product space technique ([Carlin & Chib, 1995](#); [Lodewyckx et al., 2009](#)).

Almost all of these computational methods suffer from the fact that they become less efficient and more difficult to implement as the underlying models become more complex. We now turn to an alternative method, whose implementation is extremely straightforward. The methods' main limitation is that it applies only to nested models, a limitation that also holds for p -values.

3. The Savage–Dickey density ratio

In the simplest classical hypothesis testing framework, one contemplates two models: the null hypothesis, that fixes one of its parameters to a pre-specified value of substantive interest, say $H_0 : \phi = \phi_0$; and the alternative hypothesis, in which that parameter is free to vary, say $H_1 : \phi \neq \phi_0$. Hence, the null hypothesis is nested under the alternative hypothesis, that is, H_0 can be obtained from H_1 by setting ϕ equal to ϕ_0 . Note that in the classical framework, H_0 is generally a sharp null hypothesis, or a “point null”. That is, the null hypothesis states that ϕ is exactly equal to ϕ_0 .

For example, in the binomial example from [Fig. 1](#) you answered 9 out of 10 questions correctly. Were you guessing or not? The classical and the Bayesian framework define $H_0 : \theta = .5$ as the null hypothesis for chance performance. The alternative hypothesis under which H_0 is nested could be defined as $H_1 : \theta \neq .5$, or, more specifically, as $H_1 : \theta \sim \text{Beta}(1, 1)$, which states that θ is free to vary from 0 to 1, and that it has a uniform prior distribution as shown in [Fig. 1](#).

For the binomial example, the Bayes factor for H_0 versus H_1 could be obtained by analytically integrating out the model parameter θ . However, the Bayes factor may likewise be obtained by only considering H_1 , and dividing the height of the posterior for θ by the height of the prior for θ , at the point of interest. This surprising result was first published by [Dickey and Lientz \(1970\)](#), who attributed it to Leonard J. “Jimmie” Savage. The result is now generally known as the *Savage–Dickey density ratio* (e.g., [Dickey, 1971](#); [Gamerman & Lopes, 2006](#), pp. 72–74, pp. 79–80; [Kass & Raftery, 1995](#), p. 780–781; [O’Hagan & Forster, 2004](#), pp. 174–177; for extensions and generalizations see [Chen, 2005](#); [Verdinelli & Wasserman, 1995](#)). Mathematically, the Savage–Dickey density ratio says that

$$BF_{01} = \frac{p(D|H_0)}{p(D|H_1)} = \frac{p(\theta = .5|D, H_1)}{p(\theta = .5|H_1)}. \quad (11)$$

A straightforward mathematical proof is presented in [Appendix A](#) (see also [O’Hagan & Forster, 2004](#), pp. 174–177).

In [Fig. 1](#), the two thick dots located at $\theta = .5$ provide the required information. It is evident from the figure that after observing 9 out of 10 correct responses, the height of the density at $\theta = .5$ has decreased, so that one would expect these data to cast doubt on the null hypothesis and support the alternative hypothesis. Specifically, the height of the prior distribution at $\theta = .5$ equals 1, and the height of the posterior distribution at $\theta = .5$ equals 0.107. From [Eq. \(11\)](#) the corresponding Bayes factor is $BF_{01} = 0.107/1 = 0.107$, and this corresponds exactly to the Bayes factor that was calculated by integrating out θ .

It is clear that the same procedure can be followed when the height of the posterior is not available in closed form, but instead has to be approximated from the histogram of MCMC samples. [Fig. 3](#) shows the logspline estimates ([Stone et al., 1997](#)) for the prior and the posterior densities as obtained from MCMC output. The estimated height of the prior and posterior distributions at $\theta = .5$ equal 1.00 and 0.107, respectively.

In most nested model comparisons, H_0 and H_1 have several free parameters in common. These parameters are usually not of direct interest, and they are not the focus of the hypothesis test. Hence, the common parameters are known as *nuisance parameters*. For instance, one might want to test whether or not the mean of a Normal distribution is zero (i.e., $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$), whereas the variance σ^2 is common to both models and not of immediate interest.

In general then, the framework of nested models features a parameter vector $\theta = (\phi, \psi)$, where ϕ denotes the parameter of substantive interest that is subject to test, and ψ denotes the set of nuisance parameters. The null hypothesis H_0 posits that ϕ is constrained to some special value, i.e. $\phi = \phi_0$. The alternative hypothesis H_1 assumes that ϕ is free to vary. Now consider H_1 , and let $\phi \rightarrow \phi_0$; this effectively means that H_1 reduces to H_0 —it is therefore reasonable to assume that $p(\psi|\phi \rightarrow \phi_0, H_1) = p(\psi|H_0)$ (but see Consonni & Veronese, 2008). In other words, when $\phi \rightarrow \phi_0$ the prior for the nuisance parameters under H_1 should equal the prior for the nuisance parameters under H_0 . When this condition holds, Appendix A shows that the nuisance parameters affect the Bayes factor only through the posterior for ϕ , so that again

$$BF_{01} = \frac{p(D|H_0)}{p(D|H_1)} = \frac{p(\phi = \phi_0|D, H_1)}{p(\phi = \phi_0|H_1)}, \quad (12)$$

which equals the ratio of the heights for the posterior and the prior distribution for ϕ at ϕ_0 . Thus, the Savage–Dickey density ratio holds under relatively general conditions.

Eq. (12) conveys several important messages:

1. *Relevance of the prior for the parameter of interest.* The denominator of Eq. (12) features the height of the prior for ϕ at $\phi = \phi_0$. This means that the choice of prior can greatly influence the Bayes factor, a fact that is also illustrated by Figs. 1 and 3. The choice of prior will also influence the shape of the posterior, of course, but this influence quickly diminishes as the data accumulate. This point underscores the conceptual challenge for the Bayes factors that was noted earlier (e.g., Bartlett, 1957; Liu & Aitkin, 2008). For example, consider again a test for a Normal mean μ , with $H_0 : \mu = 0$ and $H_1 : \mu \neq 0$. Suppose the prior for μ is a uniform distribution that ranges from $-a$ to a , and suppose that the number of observations is reasonably large. In this situation, the data will have overwhelmed the prior, so that the posterior for μ is relatively robust against changes in a . In contrast, the height of the prior at $\mu = 0$ varies directly with a : if a is doubled, the height of the prior at $\mu = 0$ becomes twice as small, and according to Eq. (12) this would about double the Bayes factor in favor of H_0 . In the limit, as a grows very large, the height of the prior at $\mu = 0$ goes to zero, which means that the Bayes factor will go to infinity, indicating decisive support for the null hypothesis.
2. *Irrelevance of the prior for nuisance parameters.* In contrast to the prior for the parameter of interest ϕ , Eq. (12) indicates that the prior for the nuisance parameters ψ is not critical. Hence, priors on the nuisance parameters can be vague or even improper (e.g., Hsiao, 1997, p. 659; Kass & Raftery, 1995, p. 783; Kass & Vaidyanathan, 1992). Intuitively, the prior vagueness of nuisance parameters is present in both models and cancels out in the computation of the Bayes factor (Rouder et al., 2009).
3. *Relative ease of computing the Bayes factor in nested models.* Eq. (12) shows that in nested models, under plausible assumptions on the prior structure for the nuisance parameters, computation of the Bayes factor is relatively straightforward. All that is needed is an estimate of posterior and prior ordinates under the alternative hypothesis H_1 . This computational shortcut is often much less involved than the more generic solution, which involves integrating out nuisance parameters ψ for H_0 , and parameters ψ and ϕ for H_1 , as follows:

$$BF_{01} = \frac{p(D|H_0)}{p(D|H_1)} = \frac{\int p(D|\phi = \phi_0, \psi)p(\phi = \phi_0, \psi)d\psi}{\int \int p(D|\psi, \phi)p(\psi, \phi)d\psi d\phi}. \quad (13)$$

To the best of our knowledge, the Savage–Dickey method has only been used in psychology once before, by Wetzels et al. (2009), who used it to develop a WinBUGS implementation of the t -test proposed by Rouder et al. (2009).

4. Summary and prelude to the examples

So far, we have introduced Bayesian parameter estimation, MCMC sampling, and the advantages and challenges of Bayesian hypothesis testing. In order to address the computational challenge that comes with Bayesian hypothesis testing, we outlined the Savage–Dickey density ratio method. This

straightforward and exact method applies to nested models, and for its computation the user only requires the height of the posterior and the height of the prior distribution—for the parameter that is tested, at the point of interest (see Eq. (12) and Figs. 1 and 3).

Throughout the preceding sections, Bayesian concepts have been discussed by reference to a single, extremely simple binomial example. The next sections discuss three more complicated examples, using real data taken from the psychological literature. This reflects our belief that the advantages of Bayesian hypothesis testing and the practical feasibility of the Savage–Dickey method are best illustrated by concrete examples that are highly relevant to psychological practice.

5. Example 1: equality of proportions

In their article “After the promise: the STD consequences of adolescent virginity pledges”, Brückner and Bearman (2005) analyzed a series of interviews conducted as part of the National Longitudinal Study of Adolescent Health (*Add Health*). The focus of the article was on the sexual behavior of adolescents, aged 18–24, who have made a *virginity pledge*, that is, a public or written pledge to remain a virgin until marriage. Scientific studies suggest that the sexual behavior of pledgers is not very different from that of nonpledgers—except for the fact that pledgers are less likely to use condoms when they first have sex.

The Brückner and Bearman (2005) study presents a wealth of data, but here our focus is on a small subset of the data: 424 out of 777 pledgers ($\approx 54.6\%$) indicated that they had used a condom at first sex, versus 5416 out of 9072 nonpledgers ($\approx 59.7\%$). To what extent does a statistical analysis support the assertion that pledgers are less likely than nonpledgers to use a condom at first sex?

A frequentist test for equality of proportions indicates that $p \approx .006$, which tells us that when H_0 is true (i.e., the proportions of condom users are equal in the two groups), then the probability is about .006 that we would encounter a result at least as extreme as the one that was in fact observed. But this is not the kind of information that researchers really care about; researchers want to know the extent to which the data support the claim that pledgers are less likely than nonpledgers to use a condom at first sex.

Our Bayesian model for these data is simple and general. We assume that the number of condom users ($s_1 = 424$ and $s_2 = 5416$) among the pledgers and the nonpledgers ($n_1 = 777$ and $n_2 = 9072$) is governed by binomial rate parameters θ_1 and θ_2 , respectively. Denote the difference between the two rate parameters by δ , that is, $\delta = \theta_1 - \theta_2$. Fig. 4 shows this model in graphical model notation (for introductions, see Gilks, Thomas, & Spiegelhalter, 1994; Lauritzen, 1996; Lee, 2008; Spiegelhalter, 1998). In this notation, nodes represent variables of interest, and the graph structure is used to

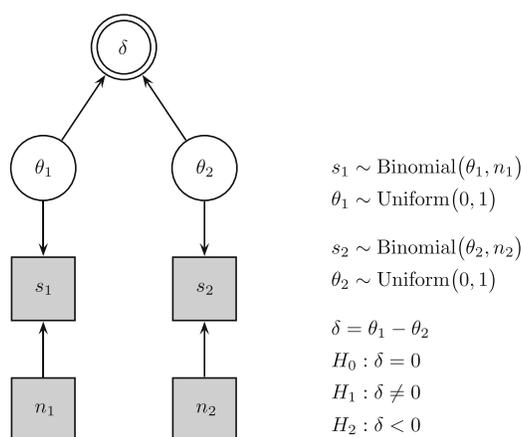


Fig. 4. Bayesian graphical model for the pledger data.

indicate dependencies between the variables, with children depending on their parents. Continuous variables are represented with circular nodes and discrete variables are represented with square nodes; observed variables are shaded and unobserved variables are not shaded. The double borders around the unobserved continuous variable δ indicates that it is deterministic (i.e., calculated without noise from other variables) rather than stochastic. In Fig. 4, for instance, the discrete observed variable s_1 indicates the number of condom users in the group of pledgers. This observed variable depends both on the (discrete, observed) number of pledgers n_1 , and on the continuous, unobserved binomial rate parameter θ_1 .

In our Bayesian model, we assume that the rate parameters θ_1 and θ_2 each have a uniform prior distribution (i.e., $p(\theta_{(i)}) \sim \text{Beta}(1, 1)$). These uniform prior distributions induce a triangular prior distribution for the difference parameter δ :

$$p(\delta) = \begin{cases} 1 + \delta & \text{for } \delta \leq 0, \\ 1 - \delta & \text{for } \delta > 0. \end{cases} \quad (14)$$

The null hypothesis states that the rates θ_1 and θ_2 are equal, and hence $H_0 : \delta = 0$. The unrestricted alternative hypothesis states that the rates are free to vary, $H_1 : \delta \neq 0$, and the restricted alternative hypothesis states that the rate is lower for the pledgers than for the nonpledgers, $H_2 : \delta < 0$. Below we examine these alternative hypothesis in turn.

5.1. Unrestricted analysis

The problem of testing $H_0 : \delta = 0$ versus $H_1 : \delta \neq 0$ is still relatively simple. The Bayes factor in support for the null hypothesis (i.e., $BF_{01} = p(D|H_0)/p(D|H_1)$) is given for instance by de Braganca Pereira and Stern (1999):

$$BF_{01} = \frac{\binom{n_1}{s_1} \binom{n_2}{s_2}}{\binom{n_1 + n_2}{s_1 + s_2}} \frac{(n_1 + 1)(n_2 + 1)}{n_1 + n_2 + 1}. \quad (15)$$

For the pledger data, this yields $BF_{01} \approx 0.45$, which means that the data are about $1/0.45 \approx 2.22$ times more likely under the alternative hypothesis than under the null hypothesis. Note that although the Bayesian hypothesis test supports the alternative hypothesis, the result is much less convincing than a p -value of .006 suggests.

To apply the Savage–Dickey method, we first draw samples from the posterior and the prior distributions for δ (the WinBUGS code can be found in Appendix B). We ran three chains for 100,000 iterations each, and we discarded the first 1000 iterations of each chain as burn-in. After confirming by means of visual inspection and the Gelman and Rubin (1992) \hat{R} statistic that the chains had converged, we collapsed the samples across the three chains.

The left panel of Fig. 5 shows the resulting histograms for the posterior and prior distributions for δ plotted on their entire range. In this panel, the thin solid line for the prior indicates the analytical distribution given in Eq. (14). For the posterior distribution, the thin solid line indicates a logspline non-parametric density estimate (Stone et al., 1997), the procedure that we will use throughout this article to estimate distributions.

The right panel of Fig. 5 zooms in on the relevant region around $\delta = 0$. The almost flat line is the analytical distribution of the prior, and the sharply decreasing line is the logspline estimate for the posterior. The two dots mark the height of both densities at $\delta = 0$. From a visual comparison of the height of the dots, it is clear that the point $\delta = 0$ is supported about twice as much under the prior as it is under the posterior. That is, the data have decreased the support for $\delta = 0$ by a factor of two. Application of the Savage–Dickey method (i.e., Eq. (12)) yields $BF_{01} \approx 0.47$, which leads to the conclusion that the data are about 2.17 times more likely under the alternative hypothesis than under the null. Thus, the result from the MCMC-based Savage–Dickey method (i.e., $BF_{10} = 2.17$) and the analytical solution (i.e., $BF_{10} = 2.22$) are in reasonable agreement.

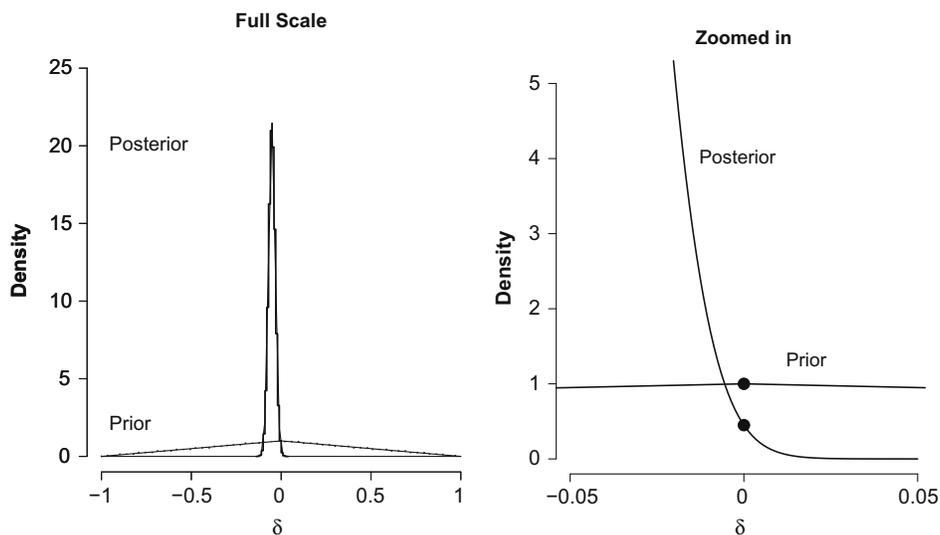


Fig. 5. Prior and posterior distributions of the rate difference δ for the unrestricted analysis of the pledger data. The left panel shows the distributions across their entire range (prior: histogram and analytical result; posterior: histogram and logspline density estimate). The right panel zooms in on the area that is relevant for the test of $H_0 : \delta = 0$ versus $H_1 : \delta \neq 0$ (prior: analytical result; posterior: logspline density estimate). The dots indicate the height of the two distributions at $\delta = 0$.

Finally, note that the conclusions from the Bayesian hypothesis test (i.e., roughly twice as much evidence for H_1 as for H_0) are more conservative than those that follow from Bayesian parameter estimation; the Bayesian 95% confidence interval for the posterior of δ is $(-0.09, -0.01)$ and does not include 0. The reason for the discrepancy is that the Bayesian hypothesis test punishes H_1 for assigning prior mass to values of δ that yield very low likelihoods (i.e., the automatic Ockham's razor discussed previously, see Berger & Delampady, 1987 for a discussion).

5.2. Order-restricted analysis

Many substantive psychological questions can be formulated as order-restrictions (e.g., Hoijtink, Klugkist, & Boelen, 2008; Klugkist et al., 2005a). Here we focus on a test of $H_0 : \delta = 0$ versus $H_2 : \delta < 0$, an order-restricted alternative hypothesis that states that the rate of condom use is lower for the pledgers than for the nonpledgers.

In the Bayesian framework, order-restrictions can be implemented in several ways (e.g., O'Hagan & Forster, 2004, pp. 70–71). For instance, order-restrictions can be enforced before MCMC sampling, by appropriately constraining the prior distributions, or they can be implemented after the MCMC sampling, by retaining only those MCMC samples that obey the order-restriction (e.g., Gelfand, Smith, & Lee, 1992, p. 525).

The left panel of Fig. 6 shows the histograms for the posterior and prior distributions for δ under the restricted model $H_2 : \delta < 0$. These histograms were obtained by selecting from the previous unrestricted analysis only those MCMC samples that obey the order-restriction. For the prior, the thin solid line indicates the analytical distribution, and for the posterior it indicates the order-restricted logspline estimate.

Note that for the prior, the effect of the order-restriction is to double the mass on $\delta = 0$, from a value of 1 to a value of 2. In contrast, the order-restriction does not much affect the posterior, as most of its mass was already smaller than 0. The right panel of Fig. 6 zooms in on the relevant area around $\delta = 0$ and shows the effect of the order-restriction on the Bayesian hypothesis test. Again, the almost flat line is the analytical distribution of the order-restricted prior, and the associated dot indicates its height at $\delta = 0$. The sharply decreasing line is the logspline estimate for the order-restricted posterior,

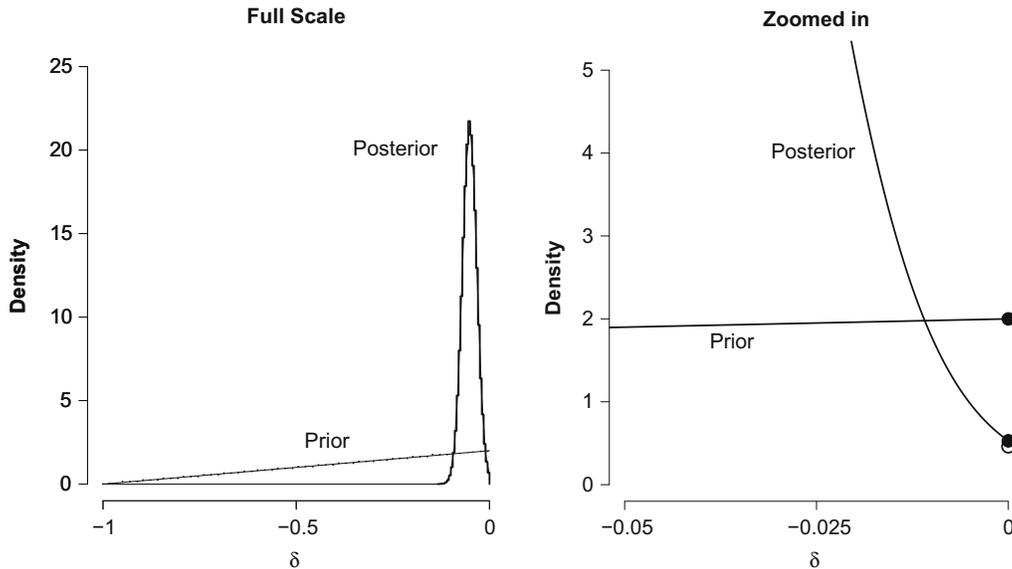


Fig. 6. Prior and posterior distributions of the rate difference δ for the order-restricted analysis of the pledger data. The left panel shows the distributions across their entire range (prior: histogram and analytical result; posterior: histogram and log spline density estimate). The right panel zooms in on the area that is relevant for the test of $H_0 : \delta = 0$ versus $H_2 : \delta < 0$ (prior: analytical result; posterior: log spline density estimate). The dots indicate the height of the two distributions at $\delta = 0$.

and the associated solid dot indicates the log spline estimate of the height of the posterior based on the subset of MCMC samples that obey the order-restriction. The open dot immediately below indicates the height of the posterior estimated from an alternative method, one that is based on renormalizing the order-restricted posterior (i.e., dividing the height of the unrestricted posterior at $\delta = 0$ by the area of the unrestricted posterior that lies to the left of $\delta = 0$).

A visual comparison of the height of the prior and posterior at $\delta = 0$ confirms that the order-restriction has increased the evidence in favor of the alternative hypothesis. Specifically, the log spline estimate yields $BF_{02} \approx 0.26$ (i.e., $BF_{20} \approx 3.78$), and the estimate based on renormalizing the posterior yields $BF_{02} \approx 0.23$ (i.e., $BF_{20} \approx 4.34$). Thus, both methods lead to the conclusion that there is roughly four times as much evidence for H_2 as for H_0 .

The foregoing may lead one to conclude that the effect of order-restrictions are similar in the Bayesian and the frequentist framework; in the Bayesian framework, the order-restriction increased the evidence against H_0 roughly by a factor of two, and in the frequentist framework, a one-sided p -value provides twice as much evidence against H_0 as a two-sided p -value. However, this correspondence only holds because the posterior for δ is largely consistent with the order-restriction. In general, one may distinguish between the following three situations, which form points on a continuum of possibilities:

1. *Posterior largely consistent with the order-restriction.* This situation occurred for the pledger data. The order-restriction increases the height of the prior by two, but it hardly increases the height of the posterior. This means that when the order-restriction is almost fully supported by the data, this can only increase the support in favor of the alternative hypothesis by a factor of two. For example, for the pledger data the unrestricted test of $H_0 : \delta = 0$ versus $H_1 : \delta \neq 0$ yields $BF_{10} \approx 2.22$. This means that the Bayes factor in favor of $H_2 : \delta < 0$ versus $H_0 : \delta = 0$ cannot be larger than $2.22 \times 2 = 4.44$.
2. *Posterior neither consistent nor inconsistent with the order-restriction.* This situation occurs when the data are uninformative with respect to the direction of the effect, so that the posterior is symmetrical around $\delta = 0$ (assuming that δ is the parameter of interest and 0 is the value at which the

alternative model collapses to the null model). In this case, the order-restriction increases the height of both the prior and the posterior by 2, so that the end result is unaffected.

3. *Posterior largely inconsistent with the order-restriction.* This situation occurs when the data suggest that the effect is in the direction opposite to that suggested by the order-restriction. In this case, the order-restriction again increases the height of the prior by 2, but it increases the height of the posterior much more. Consider, for instance, the right panel of Fig. 5, and an order-restricted test of $H_0 : \delta = 0$ versus $H_3 : \delta > 0$. To determine the height of the order-restricted posterior at $\delta = 0$ one may divide the height of the unrestricted posterior (i.e., 0.47 according to the logspline method) by its area to the right of zero, which is approximately .003. The Bayes factor in favor of $H_0 : \delta = 0$ versus $H_3 : \delta > 0$ would then be $(0.47/.003)/2 \approx 78$, which constitutes strong support for the null hypothesis.

6. Example 2: a hierarchical Bayesian one-sample *t*-test

In their article “Priming in implicit memory tasks: Prior study causes enhanced discriminability, not only bias”, Zeelenberg et al. (2002) reported three experiments in two-alternative forced-choice perceptual identification. In the test phase of each experiment, a stimulus (e.g., a picture of a clothes pin) is briefly presented and masked. Immediately after the mask the participant is confronted with two choice options—the target (i.e., the picture of the clothes pin) and a similar foil alternative (e.g., the picture of a stapler; see Fig. 7 for an example); the participant’s goal is to identify the target.

Prior to the test phase, the Zeelenberg et al. experiments featured a study phase, in which participants studied a subset of the choice alternatives that would also be presented in the later test phase. Two conditions were critical: the “study-neither” condition, in which neither choice alternative was studied, and the “study-both” condition, in which both choice alternatives were studied.

In the first two experiments reported by Zeelenberg et al., participants choose the target stimulus more often in the study-both condition than in the study-neither condition. This *both-primed benefit* suggests that prior study leads to enhanced discriminability, not just a bias to prefer the studied alternative (e.g., Ratcliff & McKoon, 1997; for a discussion see also Bowers, 1999; Wagenmakers, Zeelenberg, & Raaijmakers, 2003).

Here we focus on statistical inference for the Experiment 3 from Zeelenberg et al. (2002). In the study phase of this experiment, all 74 participants were presented with 21 pairs of similar pictures (e.g., the clothes pin/stapler example shown in Fig. 7). In the test phase, all participants had to identify briefly presented target pictures among a set of two alternatives. The test phase was composed of 42 pairs of similar pictures, 21 of which had been presented in the study phase.

In order to assess the evidence in favor of the both-primed benefit, the authors carried out a standard analysis and computed a one-sample *t*-test:

“Mean percentage of correctly identified pictures was calculated for each participant. When neither the target nor the foil had been studied, 71.5% of the pictures were correctly identified. When both the target and the foil had been studied, 74.7% of the pictures were correctly identified. The difference between the study-both and study-neither conditions was significant, $t(73) = 2.19, p < .05$.”

This analysis has two main disadvantages. First, the *t*-test assumes that the data are Normally distributed. For the Zeelenberg experiment, this assumption is certainly incorrect, as the difference between two proportions is constrained to lie between -1 and 1 (see Example 1). Second, the analysis from Zeelenberg et al. ignores the fact that the experimental design is nested (i.e., trials are

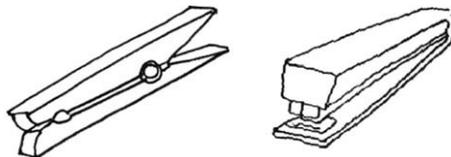


Fig. 7. Example pair of similar pictures used in Experiment 3 from Zeelenberg et al. (2002).

nested within participants), a situation that calls for a hierarchical or multi-level analysis (e.g., Gelman & Hill, 2007; Rouder, Lu, Morey, Sun, & Speckman, 2008). In other words, it is unlikely that the both-primed benefit is a fixed effect, in the sense that it is the same for each and every participant—it is more reasonable to assume that the both-primed benefit is a random effect (cf. Rouder et al., 2007).

Our Bayesian test of the both-primed benefit proceeds as follows. We start by assuming that for participant i the number of correct choices is binomially distributed with parameter θ_i . Unfortunately, θ_i is defined on the rate scale, which ranges from 0 to 1. This is an awkward scale for modeling additive effects, as a change from .55 to .65 is not the same as a change from .85 to .95. Hence, we do not model θ_i , but instead choose to model ϕ_i , the deterministic *probit transformation* of θ_i .

The probit transform is the inverse cumulative distribution function of the standard Normal distribution, so that, for instance, a rate of $\theta_i = 0.5$ maps onto a probit rate of $\phi_i = 0$, and a rate of $\theta_i = 0.975$ maps onto a probit rate of $\phi_i = 1.96$. The probit transform is shown in Fig. 8. In contrast to the rate scale, the probit scale covers the entire real line, and lends itself easily to hierarchical modeling (Rouder & Lu, 2005).

For each participant i , the both-primed benefit α_i is given by the difference between performance in the study-both and study-neither condition, $\alpha_i = \phi_{SB,i} - \phi_{SN,i}$. Our model incorporates two random effects; first, each participant's baseline level of performance $\phi_{SN,i}$ is assumed to be drawn from a group-level Normal distribution with mean μ_ϕ and standard deviation σ_ϕ . Second, each participant's both-primed benefit is assumed to be drawn from a group-level Normal distribution with mean μ_α and standard deviation σ_α . Note that such normal distributions are easily defined on the probit scale, but not on the rate scale. Fig. 9 shows the model in graphical form. In order to accommodate the hierarchical structure of the model, we use plate notation, enclosing with square boundaries subsets of the graph that have independent replications. Because each participant contributes to both the study-neither and the study-both conditions, the design is “within-subjects” and the square boundaries therefore enclose both conditions.

For the parameters that are not subject to statistical test (i.e., μ_ϕ , σ_ϕ , and σ_α) we specified uninformative priors. The prior for the group mean of the study-neither condition, μ_ϕ , is a truncated standard Normal (i.e., greater than zero only on the positive real line), which on the rate scale translates to a uniform distribution from 0.5 to 1 (cf. Rouder & Lu, 2005, p. 588). For σ_ϕ and σ_α , we chose priors that are uniform from 0 to 10.

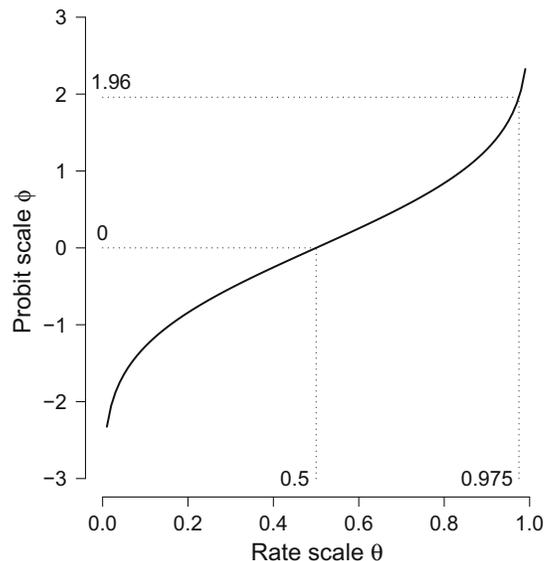


Fig. 8. The probit transformation. $\theta = \Phi(\phi)$ and $\phi = \Phi^{-1}(\theta)$, where Φ denotes the cumulative distribution function of the standard normal distribution.

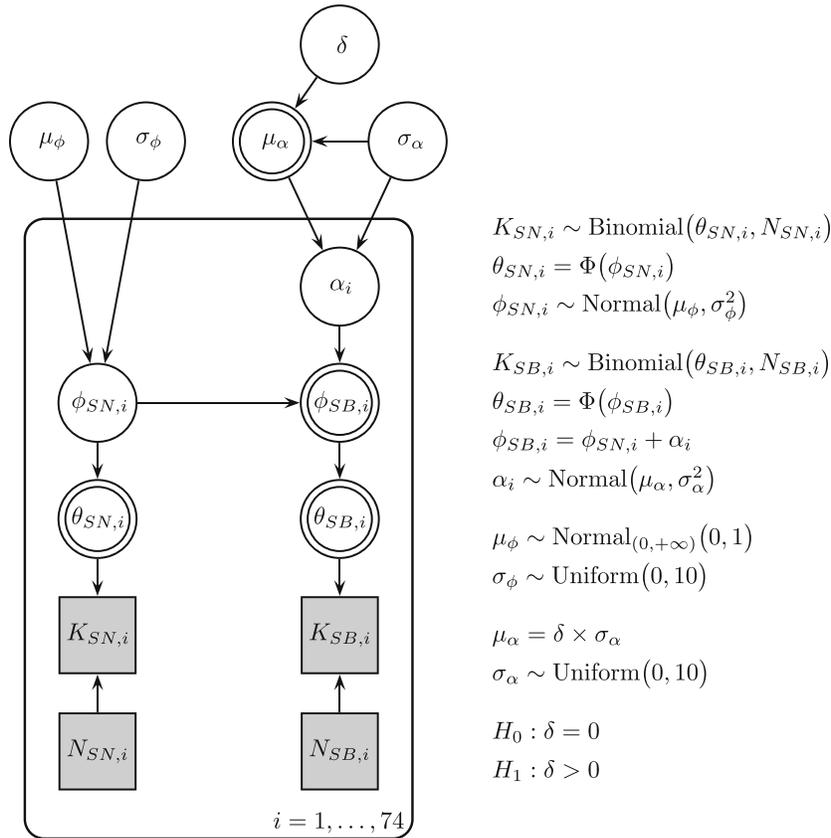


Fig. 9. Bayesian graphical model for the Zeelenberg data. In a within-subjects design, 74 participants performed a two-alternative forced-choice perceptual identification task, in both “study-neither” (SN) and “study-both” (SB) conditions.

Finally, and critically, our model incorporates a parameter δ that quantifies *effect size*, $\delta = \mu_\alpha / \sigma_\alpha$. Effect size is a dimensionless quantity, and this makes it relatively easy to define a principled prior. Reasonable default choices for priors on effect size include the Cauchy distribution (i.e., a t distribution with one degree of freedom) and the standard Normal distribution (e.g., Gönen, Johnson, Lu, & Westfall, 2005; Rouder et al., 2009). The latter prior is known as the “unit information prior”, as it carries as much information as a single observation (Kass & Wasserman, 1995). The standard Normal distribution is the prior for effect size that we will use in this example and the next.

With the statistical model in place, we can now turn to hypothesis testing. The null hypothesis states that there is no both-primed benefit, and hence the effect size is zero: $H_0 : \delta = 0$. The alternative, order-restricted hypothesis states that there is a both-primed benefit, and hence $H_1 : \delta > 0$. This test is, in fact, a hierarchical extension of the Bayesian one-sample t -test proposed by Gönen et al. (2005), Rouder et al. (2009); the difference is that our hierarchical t -test is defined on the level of individual parameters instead of raw data. Our model can therefore be thought of as a Bayesian hierarchical one-sample t -test.

We implemented our Bayesian hierarchical t -test by means of the Savage–Dickey method. First we drew MCMC samples from the posterior distribution for δ (the WinBUGS code can be found in Appendix B). As in Example 1, we ran three chains for 100,000 iterations each, and we discarded the first 1000 iterations of each chain as burn-in. After confirming by means of visual inspection and the Gelman and Rubin (1992) \hat{R} statistic that the chains had converged, we collapsed the samples across the three chains.

Fig. 10 visualizes the results—for the prior on effect size δ , the thin solid line indicates the Normal distribution that has been truncated and renormalized to take into account the order restriction that $\delta > 0$. For the posterior order-restricted distribution on effect size δ , the thin solid line indicates the logspline nonparametric density estimate, and the thick solid line indicates the histogram of MCMC samples. As in Example 1, the two dots mark the height of prior and posterior densities at $\delta = 0$. From a visual comparison of the height of the dots, it is clear that the point $\delta = 0$ is supported about four times as much under the prior as it is under the posterior. That is, the data have decreased the support for $\delta = 0$ by a factor of four. Application of the Savage–Dickey method (i.e., Eq. (12)) yields $BF_{01} \approx 0.22$, which leads to the conclusion that the data are about 4.49 times more likely under the alternative hypothesis than under the null hypothesis.

Thus, the data support the assertion that there is a both-primed benefit, but the extent of this support is somewhat weaker than is suggested by the p -value.

7. Example 3: a hierarchical Bayesian two-sample t -test

In their article “How specific are executive functioning deficits in Attention Deficit Hyperactivity Disorder and autism?”, Geurts et al. (2004) studied the performance of children with ADHD and autism on a range of cognitive tasks. Here we focus on a small subset of the data and consider the question whether children that develop typically (i.e., “normal controls”) outperform children with ADHD on the Wisconsin Card Sorting Test (WCST; Grant & Berg, 1948; Heaton, Chelune, Talley, Kay, & Curtiss, 1993). The WCST requires that participants learn, by trial and error, to sort cards according to an implicit rule. The complication is that, over the course of the experiment, the sorting rule sometimes changes. This means that in order to avoid too many mistakes, participants have to suppress the tendency to perseverate and quickly discover and adopt the new rule. Because of these task demands, performance on the WCST is thought to quantify cognitive flexibility or set shifting ability.

The experiment of interest contains data from 26 normal controls and 52 children with ADHD. Each child performed the WCST, and the measure of interest is the number of correctly sorted cards relative

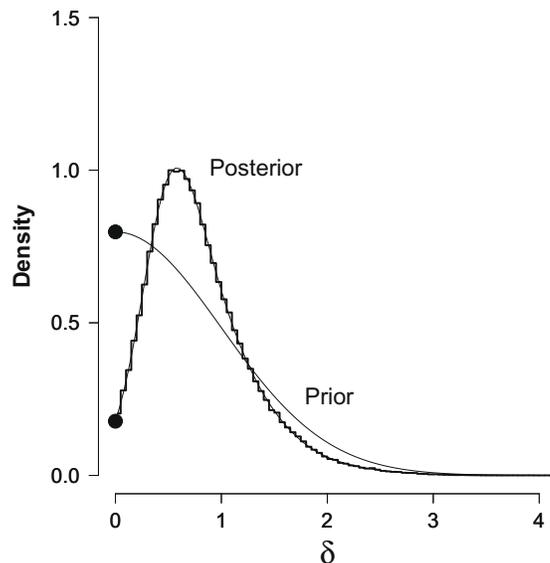


Fig. 10. Prior and posterior distribution of the effect size δ for the hierarchical, order-restricted analysis of the Zeelenberg data. For the prior, the thin line gives the analytical result; for the posterior, the thick line gives the histogram and the thin line gives the logspline density estimate. The dots indicate the height of the two distributions at $\delta = 0$.

to the total number of sorting opportunities. The WCST provides a maximum of 128 cards to sort, but, depending on a child's performance, this number could also be lower. Overall, the group of normal controls sorted the cards correctly on 65.4% of the cases, and the group of ADHD children sorted the cards correctly on 66.9% of the cases. A between-subjects (i.e., two-sample) frequentist t -test on the proportion of correctly sorted cards does not allow one to reject the null hypothesis, $t(40.2) = 0.37, p = .72$. But this statistic does not quantify the evidence in favor of the null hypothesis. Another problem with this frequentist t -test is that it ignores the fact that trials are nested in participants—a design that, as in Example 2, calls for a hierarchical/multi-level/random effects analysis.

Our hierarchical model is specified as follows. We assume that for child i in the group of normal controls, the number of correctly sorted cards $K_{NC,i}$ (out of $N_{NC,i}$ opportunities) is binomially distributed with rate parameter $\theta_{NC,i}$. As in the previous example, this rate parameter is then transformed to the probit scale (cf. Fig. 8), which yields the corresponding parameter $\phi_{NC,i}$. The comparable assumptions are made for child j in the group of ADHD children, resulting in the associated parameter $\phi_{AD,j}$.

Next, our model incorporates random effects; for both the normal controls and the group of ADHD children, the probitized rates of correct responding (i.e., ϕ_{NC} and ϕ_{AD}) are assumed to be drawn from group-level Normal distributions. Denoting the grand mean by μ , and the group difference in means by α , the group-level Normal distribution for the normal controls is defined as $N(\mu + \alpha/2, \sigma^2)$ and that for the ADHD children as $N(\mu - \alpha/2, \sigma^2)$, where σ denotes the standard deviation for the group-level distribution.

Fig. 11 shows the model in graphical form. As in Fig. 9, the hierarchical structure of the model is accommodated by plate notation, enclosing with square boundaries subsets of the graph that have independent replications. Because every child participants in only one of the two conditions, the design is “between-subjects” and the square boundaries enclose each condition separately.

For the parameters that are not subject to statistical test (i.e., μ and σ) we specified uninformative priors. The prior for the grand mean μ is a standard Normal, which on the rate scale translates to a uniform distribution from 0 to 1 (cf. Rouder & Lu, 2005, p. 588). For σ , we chose a prior that is uniform from 0 to 10. As in the previous example, the key aspect of our model is a parameter δ that quantifies effect size, $\delta = \alpha/\sigma$. We again use the “unit information” standard normal prior on δ , completing the specification of the model.

Hypothesis testing now proceeds as before. The null hypothesis states that normal controls and ADHD children perform the same on the WCST, and hence the effect size is zero: $H_0 : \delta = 0$. The unrestricted alternative hypothesis states that there is a difference in performance, and hence $H_1 : \delta \neq 0$. Lastly, the order-restricted hypothesis states that normal controls perform better than ADHD children, such that $H_2 : \delta > 0$. These tests are hierarchical extensions of the Bayesian one-sample t -test (Gönen et al., 2005; Rouder et al., 2009); as in Example 2, the difference is that our hierarchical t -tests are defined on the level of individual parameters instead of raw data. Below we examine the unrestricted analysis (i.e., H_0 versus H_1) and the restricted analysis (i.e., H_0 versus H_2) in turn.

7.1. Unrestricted analysis

We implemented our Bayesian hierarchical two-sample t -test by means of the Savage–Dickey method. As in the previous two examples, we drew MCMC samples from the posterior distribution for δ (the WinBUGS code can be found in Appendix B), we ran three chains for 100,000 iterations each, and we discarded the first 1000 iterations of each chain as burn-in. We also confirmed by means of visual inspection and the Gelman and Rubin (1992) \hat{R} statistic that the chains had converged, and we then collapsed the samples across the three chains.

The left panel of Fig. 12 visualizes the result. The ADHD children performed slightly better than the normal controls, and this is reflected in a posterior distribution for δ which is slightly asymmetrical around zero, assigning more mass to negative than to positive values of δ . The Bayesian 95% confidence interval for δ is $(-0.54, 0.42)$.

The left panel of Fig. 12 also shows that the data have made the value $\delta = 0$ more likely than it was before (i.e., at $\delta = 0$, the posterior is higher than the prior). Specifically, the ratio of the heights yields $BF_{01} = 3.96$, which indicates that the data are about four times more likely under H_0 than they are

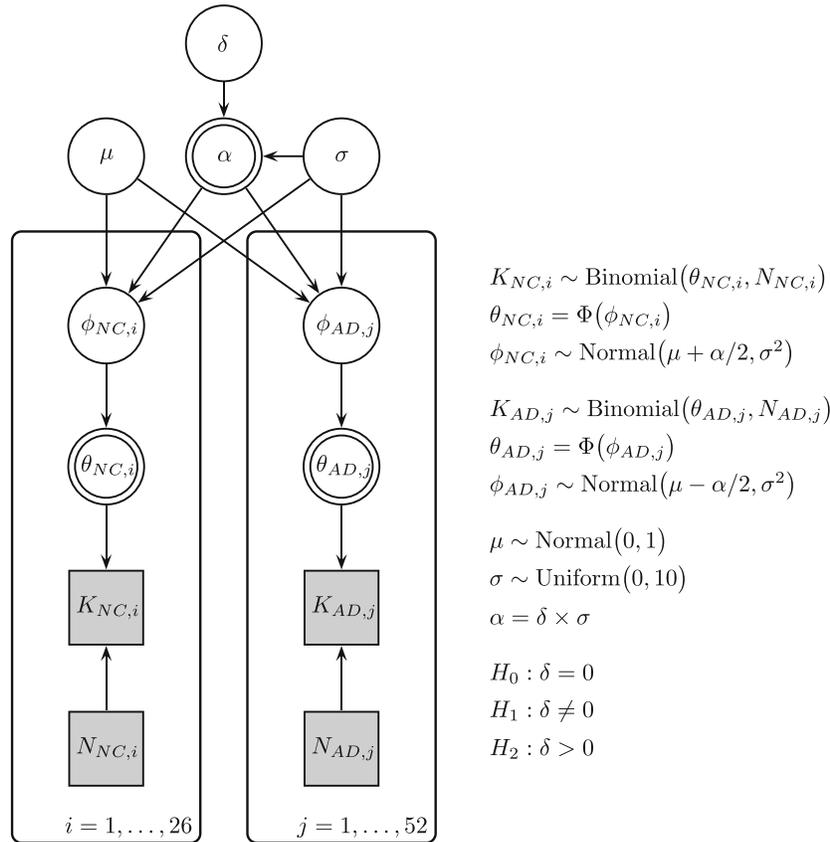


Fig. 11. Bayesian graphical model for the Geurts data. In a between-subjects design, 26 typically developing children (i.e. “normal controls”, NC) and 52 children with ADHD (AD) performed the Wisconsin Card Sorting Test.

under H_1 . Thus, the data support the claim that normal controls and ADHD children perform equally well on the WCST over the claim that these groups perform differently.

7.2. Order-restricted analysis

The order-restricted hypothesis states that normal controls outperform children with ADHD on the WCST (i.e., $H_2 : \delta > 0$). This hypothesis may be entertained because it is plausible *a priori*; However, the data show that, if anything, the reverse is true: the mean percentage of correct card selections was 1.5% higher for the group of ADHD children than for the normal controls. What can we expect when we test $H_0 : \delta = 0$ versus $H_2 : \delta > 0$?

First, note that the posterior for δ is not far from being symmetrical around zero. If it were completely symmetrical, we would have “case 2” discussed above: “Posterior neither consistent nor inconsistent with the order-restriction”. In this case the height of both the prior and the posterior is multiplied by 2, so that their ratio stays the same. Second, the the posterior for δ is not quite symmetrical around zero, and assigns slightly more mass to values that are inconsistent with H_2 . This will slightly increase the support for H_0 over H_2 . These two considerations lead us to expect that the evidence in favor of H_0 over H_2 (i.e., BF_{02}) will be slightly larger than that of H_0 over H_1 (i.e., $BF_{01} = 3.96$).

The right panel of Fig. 12 shows the result of the order-restricted analysis. As before, the relatively flat line is the analytical distribution of the order-restricted prior, and the associated dot indicates its

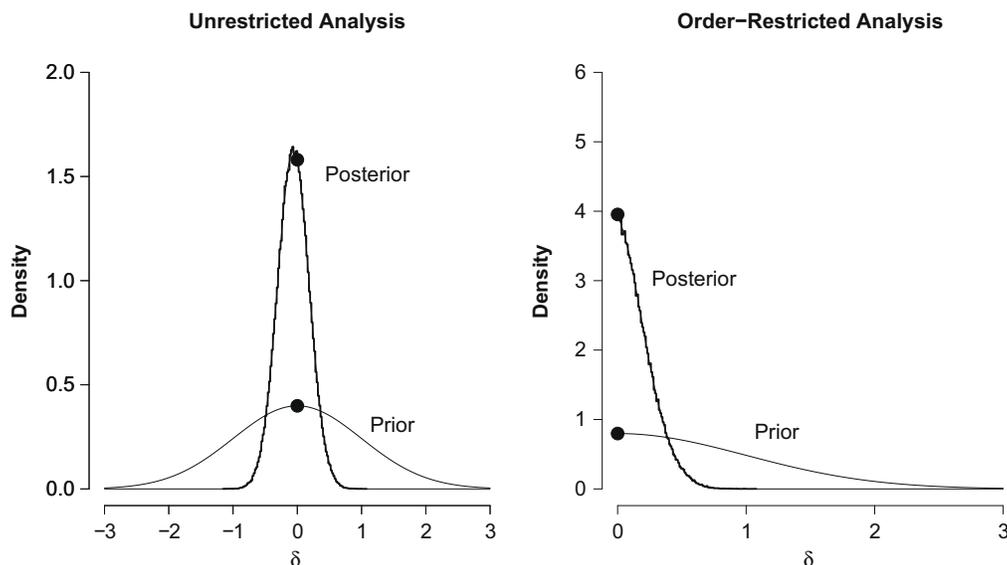


Fig. 12. Prior and posterior distribution of the effect size δ for the hierarchical analysis of the Geurts data (left panel: unrestricted analysis; right panel: order-restricted analysis) For the prior, the thin line gives the analytical result; for the posterior, the thick line gives the histogram and the thin line gives the logspline density estimate. The dots indicate the height of the two distributions at $\delta = 0$.

height at $\delta = 0$. The sharply decreasing line is the logspline estimate for the order-restricted posterior, and the associated solid dot indicates the logspline estimate of the height of the posterior based on the subset of MCMC samples that obey the order-restriction. The alternative method based on renormalizing the order-restricted posterior yielded a virtually identical result.

A quantitative comparison of the height of the prior and posterior at $\delta = 0$ confirms our expectation that the order-restriction slightly increases the evidence in favor of H_0 . Specifically, the logspline estimate yields $BF_{02} = 4.94$. Thus, under H_0 the data are about five times more likely than they are under the order-restricted alternative, a result that is slightly more convincing than the one obtained when H_0 is pitted against the unrestricted alternative. In sum, the data support the assertion that normal controls and children with ADHD perform similarly on the WCST, even though the evidence is not overwhelming.

8. Limitations of the Savage–Dickey density ratio

So far we have focused on the advantages of the Savage–Dickey density ratio method. However, the method also has its limitations, and these include the following:

1. *Markov chain Monte Carlo.* The Savage–Dickey method depends directly on the posterior distribution for the parameter that is subject to test. For most interesting models, this posterior is not available in closed-form, but instead has to be approximated by MCMC techniques. Fortunately, these MCMC techniques are implemented in the popular WinBUGS program (Lunn et al., 2000; Lunn et al., 2009; Ntzoufras, 2009); when using WinBUGS, all researchers have to do is to describe their model using an intuitive scripting language, and the details of the sampling process are automatically taken care of by WinBUGS (see Appendix B for examples).
2. *Convergence.* As explained in the section on Bayesian parameter estimation, with MCMC comes an obligation to monitor convergence; if the MCMC chains have not converged, the samples do not come from the posterior distribution, and the Savage–Dickey test will produce the wrong results. For the simple statistical models that are popular in psychology, convergence is generally very fast.

3. *Density estimation.* For its computation, the Savage–Dickey method requires an estimate of the height of a uni-dimensional posterior distribution at a single point. In this article, we have used the logspline nonparametric density estimator proposed by Stone et al. (1997). This estimator is implemented in the R package `polspline`, and concrete examples of its use are provided in the online R code associated with this article.

We chose the logspline density algorithm because it generally performs well when the posterior is restricted (e.g., only positive values are allowed), and we chose a nonparametric estimator because we wanted to avoid any assumptions about the form of the posterior distribution.

Nevertheless, the nonparametric density estimators will not be reliable when the results are extreme, that is, when the point of interest lies in the extreme tail of the posterior distribution—in the tail, the information about height is based on relatively few samples. This problem can be diagnosed by using the Savage–Dickey method multiple times to see whether the result is stable. Also, one might argue that when the point of interest is far out in the tails of the posterior distribution, the qualitative conclusion is evident and reliable (i.e., the data support H_1 over H_0), even if the quantitative result is not.

4. *Nested models.* The Savage–Dickey method can only be applied to nested models. This means that one model—the null hypothesis—needs to be a special case of a more general model. Although this is a clear limitation, scientific practice has shown that in the field of psychology, the overwhelming majority of model comparisons involve nested models (i.e., those models that also allow the computation of a p -value).
5. *Borel–Kolmogorov paradox.* A final limitation of the Savage–Dickey method originates from the way in which the priors need to be specified: $p(\psi|\phi \rightarrow \phi_0, H_1) = p(\psi|H_0)$, where ψ are the nuisance parameters and ϕ is the parameter that is subject to test (i.e., H_0 holds that ϕ takes on the specific value ϕ_0). This prior specification is intuitively plausible, but it has two disadvantages. The first disadvantage is that it is implicitly assumed that the common nuisance parameters fulfill exactly the same roles, whether they are part of H_0 or H_1 ; some people believe this assumption is too strict (for a discussion see Consonni & Veronese, 2008). The second disadvantage is that the priors are constructed by conditioning on an event that has probability zero, namely $\phi \rightarrow \phi_0$. This way of conditioning invokes the Borel–Kolmogorov paradox, a paradox that makes the results of the Bayesian hypothesis test depend on the parametrization used (Consonni & Veronese, 2008; for a summary see Wetzels, Grasman, & Wagenmakers, submitted for publication). Concretely, this means that a test of $\mu = 0$ can yield a result that differs from a conceptually equivalent test of $\mu/\sigma = 0$.

Several alternative procedures have been proposed to circumvent the Borel–Kolmogorov paradox. These alternatives define priors for nested models so that one does not condition on an event of probability zero. Unfortunately, all methods appear to come with drawbacks of their own (reviewed in Consonni & Veronese, 2008), and presently their does not appear to be a single method that is universally accepted. For most models used in psychology (e.g., regression) the choice of parametrization is clear, but this only slightly alleviates the general concern.

Despite these limitations, we hope that our examples have shown that the Savage–Dickey method can be a useful as a relatively straightforward implementation of the Bayesian hypothesis test.

9. Concluding comments

The goal of this article was to familiarize psychologists with Bayesian hypothesis testing as an alternative to calculating p -values. We have outlined a simple yet general Bayesian hypothesis test, implemented via the Savage–Dickey density ratio, that can be used to quantify the statistical evidence for and against members from a set of nested models. We have illustrated the use of this hypothesis test with concrete examples that are relevant to the analysis of routine psychological experiments. In particular, we have shown how the Bayesian hypothesis can be applied to hierarchical designs that involve order-restrictions, and how the results can quantify statistical support both for and against the null hypothesis.

Throughout this article we have illustrated the Savage–Dickey method with applications that required relatively simple statistical models; for instance, the applications had only two conditions, did not contain any covariates, and did not assume any variability across items. It is clearly desirable

to extend the Bayes factor hypothesis test to more general scenarios such as those that involve generalized linear models (Dey, Ghosh, & Mallick, 2000) and variable selection in regression (Liang, Paulo, Molina, Clyde, & Berger, 2008). The extension of the Bayesian hypothesis test to more general statistical models is ongoing, and it is likely that MCMC-based methods will be crucial for their flexible application (e.g., Ntzoufras, 2009, chap. 11).

Outside of the context of basic statistical models, the Savage–Dickey method could also be used for Bayesian hypothesis testing in a range of relatively complex mathematical process models such as the Expectancy–Valence model for the Iowa Gambling Task (Busemeyer & Stout, 2002; Wetzels, Vandekerckhove, Tuerlinckx, & Wagenmakers, *in press*), the Ratcliff diffusion model for response times and accuracy (Vandekerckhove, Tuerlinckx, & Lee, 2008; Wagenmakers, 2009), models of categorization such as ALCOVE (Kruschke, 1992), multinomial processing trees (Batchelder & Riefer, 1999), and the ACT-R model (Weaver, 2008).

For instance, a team of researchers might study the effect of alcohol on the parameters of the Ratcliff diffusion model; at some point, they might wish to test the hypothesis that alcohol has an effect on response caution. The Savage–Dickey method allows them to calculate easily the statistical support for and against this hypothesis without having to integrate out all other parameters in the model, a requirement that necessitates the use of relatively complicated numerical techniques.

For decades, cognitive psychologists carried out their statistical analysis within a single paradigm, the paradigm of p -values. This is unfortunate, not only because such a narrow focus restricts one's statistical horizon, but also because p -values only indirectly answer the kinds of questions that researchers would like to see answered. This article does not just provide a theoretical analysis of the problem (see also Nickerson, 2000; Wagenmakers, 2007), but it also offers a practical and flexible alternative. Cast in Gigerenzer's Freudian analogy, it is our hope that the Bayesian hypothesis test will help to resolve the unconscious conflict that plagues cognitive psychologists, and resolve it so that the Id can finally see its wish granted: probabilities assigned to parameters and hypotheses!

Acknowledgments

This research was supported by a Vidi grant from the Dutch Organization for Scientific Research (NWO). We thank Rene Zeelenberg for sending us the perceptual identification data (Zeelenberg et al., 2002, Experiment 3), and we thank Hilde Geurts for sending us the Wisconsin Card Sorting Test data (Geurts et al., 2004). Correspondence concerning this article may be addressed to Eric-Jan Wagenmakers, University of Amsterdam, Department of Psychology, Roetersstraat 15, 1018 WB Amsterdam, the Netherlands.

Appendix A. Derivation of the Savage–Dickey density ratio

This appendix provides the derivation of the Savage–Dickey density ratio (e.g., Dickey & Lientz, 1970, Lindley, 1972, pp. 30–32, O'Hagan & Forster, 2004, pp. 174–177).

Consider a simple model or null hypothesis, H_0 , that consists of parameter vector $\theta = (\phi, \psi)$, with ϕ set equal to some special value of substantive interest, i.e., $\phi = \phi_0$. The complex model or alternative hypothesis, H_1 , augments H_0 and states that $\phi \neq \phi_0$. Thus, parameter ϕ is the focus of interest, whereas common parameters ψ are so-called nuisance parameters. A well-known example is that of the Normal model, in which one might test whether or not the mean is zero (i.e., $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$), whereas the variance σ^2 is not of interest. To simplify notation, let subscripts 0 and 1 denote the densities under hypothesis H_0 and H_1 , respectively.

Now assume that the conditional density for ϕ is continuous at ϕ_0 , such that $\lim_{\phi \rightarrow \phi_0} p_1(\psi|\phi) = p_0(\psi)$. Then, the prior for the nuisance parameters in the complex model, conditional on $\phi \rightarrow \phi_0$ equals the prior for the nuisance parameters in the simple model, where $\phi = \phi_0$ by definition, so that $p_1(\psi|\phi = \phi_0) = p_0(\psi)$.

As explained in the main text, the Bayes factor is the ratio of marginal likelihoods, $BF_{01} = p(D|H_0)/p(D|H_1) = p_0(D)/p_1(D)$. The marginal likelihood under H_0 is given by

$$p_0(D) = \int p_0(D|\psi)p_0(\psi)d\psi. \quad (16)$$

Using the continuity condition, this can be rewritten as

$$p_0(D) = \int p_1(D|\psi, \phi = \phi_0)p_1(\psi|\phi = \phi_0)d\psi = p_1(D|\phi = \phi_0). \quad (17)$$

We then apply Bayes' rule to the right-hand side of Eq. (17) to obtain

$$p_0(D) = \frac{p_1(\phi = \phi_0|D)p_1(D)}{p_1(\phi = \phi_0)}. \quad (18)$$

We can now obtain the Bayes factor by dividing $p_0(D)$ —written as in Eq. (18)—by $p_1(D)$. The latter factor cancels, and we are left with

$$BF_{01} = \frac{p_0(D)}{p_1(D)} = \frac{p_1(\phi = \phi_0|D)}{p_1(\phi = \phi_0)}, \quad (19)$$

which is the ratio of the posterior and prior ordinate, a.k.a. the Savage–Dickey density ratio. As shown by [Wetzels, Grasman, and Wagenmakers \(submitted for publication\)](#), the Savage–Dickey method is a special, “exact equality” case of the more general encompassing prior approach advocated by [Hojtink, Klugkist, and colleagues \(Hojtink et al., 2008; Klugkist et al., 2005a; Klugkist, Laudy, & Hojtink, 2005b; Mulder et al., in press\)](#). Another generalization of the Savage–Dickey method was presented by [Verdinelli and Wasserman \(1995\)](#).

Appendix B. WinBUGS code

This appendix provides the WinBUGS computer code that implements the models discussed in this article. The WinBUGS program (e.g., [Lunn et al., 2000](#), <http://www.mrc-bsu.cam.ac.uk/bugs/>) requires that the user constructs a file containing the model specification, a file containing initial values for the model parameters, and a file containing the data. Below, we provide only the model specification files. The additional computer code is available on the first author's website, <http://www.users.fmg.uva.nl/ewagenmakers/papers.html>.

B.1. Example 1: pledger data

The WinBUGS code below implements the graphical model shown in [Fig. 4](#).

```

model
{
  # Uniform Prior on Rates:
  theta1 ~ dbeta(1,1)
  theta2 ~ dbeta(1,1)

  # Binomial Distribution for Observed Counts:
  s1 ~ dbin(theta1,n1)
  s2 ~ dbin(theta2,n2)

  # Difference between Rates:
  delta <- theta1-theta2

  # Priors
  # Make "Dummy" Variables That Copy The Prior,
  # But Are Never Updated By Data
  theta1prior ~ dbeta(1,1)
  theta2prior ~ dbeta(1,1)
  deltaprior <- theta1prior-theta2prior
}

```

B.2. Example 2: Zeelenberg data

The WinBUGS code below implements the graphical model shown in Fig. 9.

```

model
{
  for(i in 1:74) # 74 Participants
  {
    # Binomial Distributions for Observed Counts:
    K.SN[i] ~ dbin(theta.SN[i], N.SN[i])
    K.SB[i] ~ dbin(theta.SB[i], N.SB[i])

    # Transformation to Parameters on the Probit Scale:
    theta.SN[i] <- phi(phi.SN[i])
    theta.SB[i] <- phi(phi.SB[i])

    # Individual Parameters that Quantify Performance in
    # the Study-Neither Condition Come From a Group-Level Distribution:
    phi.SN[i] ~ dnorm(mu.phi,tau.phi)
    # NB. tau.phi is the precision, defined as 1/variance

    # On the Probit Scale, Priming Effects Are Additive:
    phi.SB[i] <- phi.SN[i] + alpha[i]
    # alpha[i] is the priming effect for participant i

    # Individual Priming Effects Come From a Group-Level Distribution:
    alpha[i] ~ dnorm(mu.alpha,tau.alpha)
    # NB. tau.alpha is the precision, defined as 1/variance
  }

  # Group-Level Priors for the Study-Neither Condition:
  mu.phi ~ dnorm(0,1)I(0,)
  # NB1. The I(0,) command ensures that all samples for mu.phi are > 0
  # NB2. This prior for mu.phi corresponds to a uniform prior the rate scale,
  # ranging from 0.5 to 1.

  # Uninformative Prior on the Group-Level Standard Deviation:
  sigma.phi ~ dunif(0,10)
  # Transformation from Standard Deviation to Precision:
  tau.phi <- pow(sigma.phi,-2)

  # Priors for the Group-Level Priming Effect (cf. Rouder et al., PBR):
  mu.alpha <- delta * sigma.alpha
  # Uninformative Prior for sigma.alpha:
  sigma.alpha ~ dunif(0,10)
  # Transformation from Standard Deviation to Precision:
  tau.alpha <- pow(sigma.alpha,-2)

  # The "Unit Information Prior" on Effect Size delta (cf. Rouder et al., PBR):
  delta ~ dnorm(0,1)I(0,)
  # NB. The I(0,) incorporates the order-restriction that allows only
  # positive values for delta
}

```

B.3. Example 3: Geurts data

The WinBUGS code below implements the graphical model shown in Fig. 11.

```

model
{
  for(i in 1:26) # 26 Normal Control Participants
  {
    # Binomial Distributions for Observed Counts:
    K.NC[i] ~ dbin(theta.NC[i],N.NC[i])
    # Transformation to Parameters on the Probit Scale:
    theta.NC[i] <- phi(phi.NC[i])
    # Individual Parameters Come From a Group-Level Distribution:
    phi.NC[i] ~ dnorm(mu.NC,tau)
    # NB. tau is the precision, defined as 1/variance
  }
  for(j in 1:52) # 52 ADHD Participants
  {
    # Binomial Distributions for Observed Counts:
    K.AD[j] ~ dbin(theta.AD[j],N.AD[j])
    # Transformation to Parameters on the Probit Scale:
    theta.AD[j] <- phi(phi.AD[j])
    # Individual Parameters Come From a Group-Level Distribution:
    phi.AD[j] ~ dnorm(mu.AD,tau)
    # NB. tau is the precision, defined as 1/variance
  }
  mu.NC <- mu + (.5*alpha)
  mu.AD <- mu - (.5*alpha)
  # NB. mu is the grand mean, alpha is the effect (i.e., the group difference)

  # Group-Level Priors:
  mu ~ dnorm(0,1)
  # NB. This prior for mu corresponds to a uniform prior the rate scale,
  # ranging from 0 to 1.

  # Uninformative Prior on the Group-Level Standard Deviation:
  sigma ~ dunif(0,10)
  # Transformation from Standard Deviation to Precision:
  tau <- pow(sigma, -2)
  alpha <- delta * sigma
  # NB. This allows one to put a prior on effect size delta (cf. Rouder et al., PBR)

  # The "Unit Information Prior" on Effect Size delta (cf. Rouder et al., PBR):
  delta ~ dnorm(0,1)
}

```

References

- Aitchison, J., & Dunsmore, I. R. (1975). *Statistical prediction analysis*. Cambridge: Cambridge University Press.
- Anscombe, F. J. (1963). Sequential medical trials. *Journal of the American Statistical Association*, 58, 365–383.
- Bartlett, M. S. (1957). A comment on D.V. Lindley's statistical paradox. *Biometrika*, 44, 533–534.
- Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review*, 6, 57–86.
- Berger, J. O., & Berry, D. A. (1988a). The relevance of stopping rules in statistical inference. In S. S. Gupta & J. O. Berger (Eds.), *Statistical decision theory and related topics* (Vol. 1, pp. 29–72). New York: Springer Verlag.

- Berger, J. O., & Berry, D. A. (1988b). Statistical analysis and the illusion of objectivity. *American Scientist*, 76, 159–165.
- Berger, J. O., & Delampady, M. (1987). Testing precise hypotheses. *Statistical Science*, 2, 317–352.
- Berger, J. O., & Mortera, J. (1999). Default Bayes factors for nonnested hypothesis testing. *Journal of the American Statistical Association*, 94, 542–554.
- Berger, J. O., & Pericchi, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, 91, 109–122.
- Berger, J. O., & Pericchi, L. R. (2001). Objective Bayesian methods for model selection: Introduction and comparison (with discussion). In P. Lahiri (Ed.), *Model selection* (Vol. 38, pp. 135–207). Beachwood, OH: Institute of Mathematical Statistics Lecture Notes—Monograph Series.
- Bernardo, J. M., & Smith, A. F. M. (1994). *Bayesian theory*. New York: Wiley.
- Besag, J. (1989). A candidate's formula: A curious result in Bayesian prediction. *Biometrika*, 76, 183.
- Bowers, J. S. (1999). Priming is not all bias: Commentary on Ratcliff and McKoon (1997). *Psychological Review*, 106, 582–596.
- Bowers, J. S., Vigliocco, G., & Haan, R. (1998). Orthographic, phonological, and articulatory contributions to masked letter and word priming. *Journal of Experimental Psychology: Human Perception and Performance*, 24, 1705–1719.
- Brückner, H., & Bearman, P. (2005). After the promise: The STD consequences of adolescent virginity pledges. *Journal of Adolescent Health*, 36, 271–278.
- Busemeyer, J. R., & Stout, J. C. (2002). A contribution of cognitive decision models to clinical assessment: Decomposing performance on the Bechara gambling task. *Psychological Assessment*, 14, 253–262.
- Carlin, B. P., & Chib, S. (1995). Bayesian model choice via Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B*, 57, 473–484.
- Chen, M.-H. (2005). Computing marginal likelihoods from a single MCMC output. *Statistica Neerlandica*, 59, 16–29.
- Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90, 1313–1321.
- Chib, S., & Jeliazkov, I. (2001). Marginal likelihood from the Metropolis–Hastings output. *Journal of the American Statistical Association*, 96, 270–281.
- Christensen, R. (2005). Testing fisher, Neyman, Pearson, and Bayes. *The American Statistician*, 59, 121–126.
- Congdon, P. (2003). *Applied Bayesian modelling*. Chichester, UK: Wiley.
- Consonni, G., & Veronese, P. (2008). Compatibility of prior specifications across linear models. *Statistical Science*, 23, 332–353.
- de Braganca Pereira, C. A., & Stern, J. M. (1999). Evidence and credibility: Full Bayesian significance test for precise hypotheses. *Entropy*, 1, 99–110.
- Dennis, S., & Humphreys, M. S. (2001). A context noise model of episodic word recognition. *Psychological Review*, 108, 452–477.
- Dennis, S. J., Lee, M. D., & Kinnell, A. (2008). Bayesian analysis of recognition memory: The case of the list – length effect. *Journal of Memory and Language*, 59, 361–376.
- Dey, D. K., Ghosh, S. K., & Mallick, B. K. (Eds.). (2000). *Generalized linear models: A Bayesian perspective*. Boca Raton, FL: Taylor & Francis.
- Dickey, J. M. (1971). The weighted likelihood ratio, linear hypotheses on normal location parameters. *The Annals of Mathematical Statistics*, 42, 204–223.
- 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.
- Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society, Series B*, 57, 45–97.
- Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, 70, 193–242.
- Forster, K. I., Mohan, K., & Hector, J. (2003). The mechanics of masked priming. In S. Kinoshita & S. J. Lupker (Eds.), *Masked priming: The state of the art* (pp. 3–38). New York, NY: Psychology Press.
- Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review*, 116, 439–453.
- Gamerman, D., & Lopes, H. F. (2006). *Markov chain Monte Carlo: Stochastic simulation for Bayesian inference*. Boca Raton, FL: Chapman & Hall.
- Gelfand, A. E., Smith, A. F. M., & Lee, T.-M. (1992). Bayesian analysis of constrained parameter and truncated data problems using Gibbs sampling. *Journal of the American Statistical Association*, 87, 523–532.
- Gelman, A. (1996). Inference and monitoring convergence. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice* (pp. 131–143). Boca Raton, FL: Chapman & Hall/CRC.
- Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.
- Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, 7, 457–472.
- Geurts, H. M., Verté, S., Oosterlaan, J., Roeyers, H., & Sergeant, J. A. (2004). How specific are executive functioning deficits in attention deficit hyperactivity disorder and autism? *Journal of Child Psychology and Psychiatry*, 45, 836–854.
- Gigerenzer, G. (1993). The Superego, the Ego, and the Id in statistical reasoning. In C. Lewis & G. Keren (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 311–339). Hillsdale, NJ: Erlbaum.
- Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, 33, 587–606.
- Gigerenzer, G., Krauss, S., & Vitouch, O. (2004). The null ritual: What you always wanted to know about significance testing but were afraid to ask. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 391–408). Thousand Oaks, CA: Sage.
- Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (Eds.). (1996). *Markov chain Monte Carlo in practice*. Boca Raton, FL: Chapman & Hall/CRC.
- Gilks, W. R., Thomas, A., & Spiegelhalter, D. J. (1994). A language and program for complex Bayesian modelling. *The Statistician*, 43, 169–177.
- Gill, J. (2002). *Bayesian methods: A social and behavioral sciences approach*. Boca Raton, FL: CRC Press.

- Gönen, M., Johnson, W. O., Lu, Y., & Westfall, P. H. (2005). The Bayesian two-sample t test. *The American Statistician*, 59, 252–257.
- Good, I. J. (1985). Weight of evidence: A brief survey. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, & A. F. M. Smith (Eds.), *Bayesian statistics 2* (pp. 249–269). New York: Elsevier.
- Grant, D. A., & Berg, E. A. (1948). A behavioral analysis of degree of reinforcement and ease of shifting to new responses in a Weigl-type card-sorting problem. *Journal of Experimental Psychology*, 38, 404–411.
- Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82, 711–732.
- Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research*, 7, 1–20.
- Heaton, R. K., Chelune, G. J., Talley, J. L., Kay, G. G., & Curtiss, G. (1993). *Wisconsin card sorting test manual: Revised and expanded*. Odessa, FL: Psychological Assessment Resources, Inc..
- Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, 14, 382–417.
- Hojtink, H., Klugkist, I., & Boelen, P. (2008). *Bayesian evaluation of informative hypotheses that are of practical value for social scientists*. New York: Springer.
- Hsiao, C. K. (1997). Approximate Bayes factors when a mode occurs on the boundary. *Journal of the American Statistical Association*, 92, 656–663.
- Hubbard, R., & Bayarri, M. J. (2003). Confusion over measures of evidence (p 's) versus errors (α 's) in classical statistical testing. *The American Statistician*, 57, 171–182.
- Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge, UK: Cambridge University Press.
- Jeffreys, H. (1961). *Theory of probability*. Oxford, UK: Oxford University Press.
- Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90, 773–795.
- Kass, R. E., & Vaidyanathan, S. K. (1992). Approximate Bayes factors and orthogonal parameters, with application to testing equality of two binomial proportions. *Journal of the Royal Statistical Society, Series B*, 54, 129–144.
- Kass, R. E., & Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association*, 90, 928–934.
- Klugkist, I., Laudy, O., & Hoijtink, H. (2005a). Inequality constrained analysis of variance: A Bayesian approach. *Psychological Methods*, 10, 477–493.
- Klugkist, I., Laudy, O., & Hoijtink, H. (2005b). Bayesian eggs and Bayesian omelettes: Reply to Stern (2005). *Psychological Methods*, 10, 500–503.
- Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, 99, 22–44.
- Lauritzen, S. L. (1996). *Graphical models*. Oxford: Clarendon.
- Lee, M. D. (2008). Three case studies in the Bayesian analysis of cognitive models. *Psychonomic Bulletin & Review*, 15, 1–15.
- 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.
- Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, 103, 410–423.
- Lindley, D. V. (1972). *Bayesian statistics, a review*. Philadelphia, PA: SIAM.
- Lindley, D. V. (2000). The philosophy of statistics. *The Statistician*, 49, 293–337.
- Liu, C. C., & Aitkin, M. (2008). Bayes factors: Prior sensitivity and model generalizability. *Journal of Mathematical Psychology*, 52, 362–375.
- Lodewyckx, T., Lee, M.D., & Wagenmakers, E.-J. (submitted for publication). A general computational method for estimating Bayes factors.
- Lunn, D., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The BUGS project: Evolution, critique and future directions. *Statistics in Medicine*, 28, 3049–3067.
- Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS – a Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, 10, 325–337.
- MacKay, D. J. C. (2003). *Information theory, inference, and learning algorithms*. Cambridge: Cambridge University Press.
- Madigan, D., & Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, 89, 1535–1546.
- Mulder, J., Klugkist, I., van de Schoot, R., Meeus, W.H.J., Selfhout, M., & Hoijtink, H. (in press). Bayesian model selection of informative hypotheses for repeated measurements. *Journal of Mathematical Psychology*.
- Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, 47, 90–100.
- Myung, I. J., Forster, M. R., & Browne, M. W. (2000). Model selection [Special issue]. *Journal of Mathematical Psychology*, 44(1–2).
- Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, 4, 79–95.
- Nickerson, R. S. (2000). Null hypothesis statistical testing: A review of an old and continuing controversy. *Psychological Methods*, 5, 241–301.
- Ntzoufras, I. (2009). *Bayesian modeling using WinBUGS*. Hoboken, NJ: Wiley.
- O'Hagan, A. (1995). Fractional Bayes factors for model comparison. *Journal of the Royal Statistical Society B*, 57, 99–138.
- O'Hagan, A., & Forster, J. (2004). *Kendall's advanced theory of statistics* (2nd ed.). *Bayesian inference* (Vol. 2B). London: Arnold.
- Ratcliff, R., & McKoon, G. (1997). A counter model for implicit priming in perceptual word identification. *Psychological Review*, 104, 319–343.
- 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., Sun, D., Speckman, P., Morey, R., & Naveh-Benjamin, M. (2007). Signal detection models with random participant and item effects. *Psychometrika*, 72, 621–642.

- 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.
- Royall, R. M. (1997). *Statistical evidence: A likelihood paradigm*. London: Chapman & Hall.
- Sellke, T., Bayarri, M. J., & Berger, J. O. (2001). Calibration of p values for testing precise null hypotheses. *The American Statistician*, 55, 62–71.
- Sheu, C.-F., & O'Curry, S. L. (1998). Simulation-based Bayesian inference using BUGS. *Behavioral Research Methods, Instruments, & Computers*, 30, 232–237.
- Sisson, S. A. (2005). Transdimensional Markov chains: A decade of progress and future perspectives. *Journal of the American Statistical Association*, 100, 1077–1089.
- Smith, A. F. M., & Spiegelhalter, D. J. (1980). Bayes factors and choice criteria for linear models. *Journal of the Royal Statistical Society B*, 42, 213–220.
- Spiegelhalter, D. J. (1998). Bayesian graphical modelling: A case-study in monitoring health outcomes. *Applied Statistics*, 47, 115–133.
- Stone, C. J., Hansen, M. H., Kooperberg, C., & Truong, Y. K. (1997). Polynomial splines and their tensor products in extended linear modeling (with discussion). *The Annals of Statistics*, 25, 1371–1470.
- 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). Cognitive Science Society.
- Verdinelli, I., & Wasserman, L. (1995). Computing Bayes factors using a generalization of the Savage–Dickey density ratio. *Journal of the American Statistical Association*, 90, 614–618.
- Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, 14, 779–804.
- Wagenmakers, E.-J. (2009). Methodological and empirical developments for the Ratcliff diffusion model of response times and accuracy. *European Journal of Cognitive Psychology*, 21, 641–671.
- Wagenmakers, E.-J., Lee, M. D., Lodewyckx, T., & Iverson, G. (2008). Bayesian versus frequentist inference. In H. Hoijtink, I. Klugkist, & P. A. Boelen (Eds.), *Bayesian evaluation of informative hypotheses* (pp. 181–207). New York: Springer Verlag.
- Wagenmakers, E.-J., Ratcliff, R., Gomez, P., & Iverson, G. J. (2004). Assessing model mimicry using the parametric bootstrap. *Journal of Mathematical Psychology*, 48, 28–50.
- Wagenmakers, E.-J., & Waldorp, L. (2006). Model selection: Theoretical developments and applications [Special issue]. *Journal of Mathematical Psychology*, 50(2).
- Wagenmakers, E.-J., Zeelenberg, R., & Raaijmakers, J. G. W. (2003). Testing the counter model for perceptual identification: Effects of repetition priming and word frequency. *Psychonomic Bulletin & Review*, 7, 662–667.
- Weaver, R. (2008). Parameters, predictions, and evidence in computational modeling: A statistical view informed by ACT-R. *Cognitive Science*, 32, 1349–1375.
- Wetzels, R., Grasman, R.P.P., & Wagenmakers, E.-J. (submitted for publication). An encompassing prior generalization of the Savage–Dickey density ratio test.
- Wetzels, R., Vandekerckhove, J., Tuerlinckx, F., & Wagenmakers, E.-J. (in press). Bayesian parameter estimation in the Expectancy Valence model of the Iowa gambling task. *Journal of Mathematical Psychology*.
- 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. *Psychonomic Bulletin & Review*, 16, 752–760.
- Wilkinson, L., & The Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54, 594–604.
- Zeelenberg, R., Wagenmakers, E.-J., & Raaijmakers, J. G. W. (2002). Priming in implicit memory tasks: Prior study causes enhanced discriminability, not only bias. *Journal of Experimental Psychology: General*, 131, 38–47.