Psychonomic Bulletin & Review
2009, 16 (4), 752-760
doi:10.3758/PBR.16.4.752

Notes and Comment

How to quantify support for and against the null hypothesis: A flexible WinBUGS implementation of a default Bayesian *t* test

Ruud Wetzels, Jeroen G. W. Raaijmakers, Emöke Jakab, and Eric-Jan Wagenmakers
University of Amsterdam, Amsterdam, The Netherlands

We propose a sampling-based Bayesian t test that allows researchers to quantify the statistical evidence in favor of the null hypothesis. This Savage–Dickey (SD) t test is inspired by the Jeffreys–Zellner–Siow (JZS) t test recently proposed by Rouder, Speckman, Sun, Morey, and Iverson (2009). The SD test retains the key concepts of the JZS test but is applicable to a wider range of statistical problems. The SD test allows researchers to test order restrictions and applies to two-sample situations in which the different groups do not share the same variance.

> Never use the unfortunate expression "accept the null-hypothesis." (Wilkinson and the Task Force on Statistical Inference, 1999, p. 599)

Popular theories are difficult to overthrow. Consider, for instance, the following hypothetical sequence of events. First, Dr. John proposes a seasonal memory model (SMM). The model is intuitively attractive and quickly gains in popularity. Dr. Smith, however, remains unconvinced and decides to put one of SMMs predictions to the test. Specifically, SMM predicts that the increase in recall performance due to the intake of glucose is more pronounced in summer than in winter. Dr. Smith conducts the relevant experiment using a within-subjects design and finds the exact opposite, although the result is not significant. More specifically, Dr. Smith finds that with $N = 41$, the t value equals 0.79, which corresponds to a two-sided p value of .44 (see Table 1).

Clearly, Dr. Smith's data do not support SMMs prediction that the glucose-driven increase in performance is larger in summer than in winter. Instead, the data seem to suggest that the null hypothesis is plausible and that no difference between summer and winter is evident. Dr. Smith submits his findings to the *Journal of Experimental Psychology: Learning, Memory, and the Seasons*. Three months later, Dr. Smith receives the reviews, and one of them is from Dr. John. This review includes the following comment:

> From a null result, we cannot conclude that no difference exists, merely that we cannot reject the null hypothesis. Although some have argued that with enough data we can argue for the null hypothesis, most agree that this is only a reasonable thing to do in the face of a sizeable amount [sic] of data [which] has been collected over many experiments that control for all concerns. These conditions are not met here. Thus, the empirical contribution here does not enable readers to conclude very much, and so is quite weak . . . .[1]

In this article, we outline a statistical method that allows Dr. Smith to quantify the evidence for the null hypothesis versus the SMM hypothesis. More generally, this method is appropriate for a test between two hypotheses, where one is nested in the other. Our work is inspired by the automatic Jeffreys–Zellner–Siow (JZS) Bayesian t test that was recently proposed by Rouder, Speckman, Sun, Morey, and Iverson (2009). Although the JZS test is able to quantify support in favor of the null hypothesis, it does not help Dr. Smith, because the prediction of SMM (i.e., the alternative hypothesis) is directional, one-sided, or order restricted (e.g., Hoijtink, Klugkist, & Boelen, 2008; Klugkist, Laudy, & Hoijtink, 2005). In other words, SMM does not merely predict that the increase in recall performance differs from summer to winter, but it makes the more specific prediction that the increase in recall performance is *larger* in summer than it is in winter. The JZS test does not directly apply to this scenario. In addition, the JZS two-sample test assumes that both groups share the same variance. When this assumption is violated, the test may no longer be reliable, a phenomenon that statisticians have studied extensively (i.e., the Behrens–Fisher problem; Kim & Cohen, 1998). To address these limitations, we have developed a flexible sampling-based alternative to the JZS test. This alternative procedure, which we name the Savage–Dickey (SD) test, retains the key concepts of the JZS test but applies to a wider range of statistical problems. The computer code for the SD test and step-by-step procedures for implementing the program can be found on the first author's Web site, www.ruudwetzels.com.

The outline of this article is as follows. First, we will provide the necessary Bayesian background, and then we will discuss the statistical details of Rouder et al.'s (2009) JZS test. Next, we will explain our own procedure, the SD test, and will demonstrate by simulation that it mimics the JZS test—for both the one-sample and two-sample cases. Subsequently, we will outline two ways in which the SD test extends the JZS test. First, the SD test enables researchers such as Dr. Smith to test order-restricted hypoth-

R. Wetzels, wetzels.ruud@gmail.com

Table 1
Increase in Recall Performance Due to Intake of Glucose in Summer and Winter: A Hypothetical Example

Season	*N*	*M*	*SD*
Winter	41	0.11	0.15
Summer	41	0.07	0.23

Note—$t = 0.79, p = .44$.

eses (i.e., one-sided *t* test). Second, the SD test can deal with two-sample situations in which the different groups do not share the same variance.

Bayesian Hypothesis Testing

In order to keep this article self-contained, we will briefly recapitulate the basic principles of Bayesian hypothesis testing (for details, see Kass & Raftery, 1995; Myung & Pitt, 1997; O'Hagan & Forster, 2004; Wasserman, 2000). First, we will explain the concept of *Bayes factors*, and then we will discuss Rouder et al.'s (2009) JZS test, on which our method is based.

Bayes factors. In Bayesian inference, competing hypotheses (i.e., statistical models) are assigned probabilities. For instance, assume that you entertain two hypotheses, a null hypothesis H_0 and an alternative hypothesis H_1. Before the data D are seen, these hypotheses have *prior* probabilities $p(H_0)$ and $p(H_1)$. The ratio of these two probabilities defines the *prior odds*. When the data D come in, the prior odds are updated to *posterior odds*, which is defined as the ratio of posterior probabilities $p(H_0|D)$ and $p(H_1|D)$:

$$\frac{p(H_0 \mid D)}{p(H_1 \mid D)} = \frac{p(D \mid H_0)}{p(D \mid H_1)} \times \frac{p(H_0)}{p(H_1)}. \tag{1}$$

Equation 1 shows that the change from prior odds to posterior odds is quantified by $p(D|H_0)/p(D|H_1)$, the so-called *Bayes factor*. Thus, Equation 1 reads,

$$\text{posterior odds} = \text{Bayes factor} \times \text{prior odds}. \tag{2}$$

When the Bayes factor is, say, 14, this indicates that the data are 14 times more likely to have occurred under H_0 than under H_1, irrespective of the prior probabilities that you may assign to H_0 and H_1. When H_0 and H_1 are equally likely a priori, however, a Bayes factor of 14 translates directly to posterior probability; here, this means that after the data are seen, H_0 is 14 times more likely than is H_1. Alternatively, one may state that the posterior probability in favor of H_0 equals $14/15 \approx .93$ and the posterior probability in favor of H_1 is its complement—that is, $p(H_1|D) = 1 - p(H_0|D) \approx .07$.[2]

One of the attractions of the Bayes factor is that it follows the principle of parsimony: When two models fit the data equally well, the Bayes factor prefers the simple model over the more complex one (Berger & Jefferys, 1992; Myung & Pitt, 1997). This fact can be appreciated by considering how the components of the Bayes factor are calculated. Specifically, both $p(D|H_0)$ and $p(D|H_1)$ are derived by averaging the likelihood over the prior:

$$p(D|H) = \int_{\theta \in \Theta_H} f_H(D|\theta) p_H(\theta) d\theta, \tag{3}$$

where Θ_H denotes the parameter space under the hypothesis of interest H, f_H is the likelihood, and p_H denotes the prior distribution on the model parameters θ. Note that a complex model has a relatively large parameter space; a complex model tends to have many parameters, some of which may furthermore have a complicated functional form. Because of its large parameter space, a complex model has to spread out its prior probability quite thinly over the parameter space. As a result, the occurrence of any particular event will not greatly add to that model's credibility. A prior that is very spread out will occupy a relatively large part of the parameter space in which the likelihood for the observed data is almost zero, and this decreases the average likelihood $p(D|H)$ (Myung & Pitt, 1997).

Rouder et al.'s default Bayesian JZS *t* test. Consider the one-sample *t* test. We assume that the data are normally distributed with unknown mean μ and unknown variance σ^2. The null hypothesis states that the mean is equal to zero—that is, $H_0 : \mu = 0$. The alternative hypothesis states that the mean is not equal to zero—that is, $H_1 : \mu \neq 0$. Denote by BF_{01} the Bayes factor in favor of H_0 over H_1. From Equation 3, the separate components of BF_{01} are given by

$$p(D|H_0) = \int_0^\infty f_0(D \mid \mu = 0, \sigma^2) p_0(\mu = 0, \sigma^2) d\sigma^2 \tag{4A}$$

and

$$p(D|H_1) = \int_{-\infty}^\infty \int_0^\infty f_1(D \mid \mu, \sigma^2) p_1(\mu, \sigma^2) d\sigma^2 d\mu. \tag{4B}$$

These equations feature priors on the model parameters (i.e., p_0 and p_1). Rouder et al. (2009) followed Jeffreys (1961) and proposed a prior on effect size $\delta = \mu/\sigma$, instead of on the mean μ. Specifically, Rouder et al. (2009) defined a Cauchy prior on δ with location parameter 0 and scale parameter 1 (i.e., a *t* distribution with 1 *df*) and a Jeffreys' prior (Jeffreys, 1961) on the variance:

$$\delta \sim \text{Cauchy}(0,1), \tag{5}$$

and

$$p(\sigma^2) \propto 1/\sigma^2, \tag{6}$$

where $\propto$ denotes *is proportional to*. This completes the specification of H_0 and H_1. Rouder et al. (2009) then derived Equation 7, below, for the JZS Bayes factor. In this equation, *t* is the *t* statistic for the one-sided *t* test, *N* is the number of observations, $\nu = N - 1$ equals the degrees of

$$BF_{01} = \frac{\left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2}}{\int_0^\infty (1 + Ng)^{-1/2} \left(1 + \frac{t^2}{(1+Ng)\nu}\right)^{-(\nu+1)/2} (2\pi)^{-1/2} g^{-3/2} \exp\left[-1/(2g)\right] dg} \tag{7}$$

freedom, and g represents Zellner's g-prior (for a detailed explanation, see Liang, Paulo, Molina, Clyde, & Berger, 2008; Zellner, 1986; Zellner & Siow, 1980).

In order to apply this Bayesian t test to two-sample designs, Equation 7 needs to be adjusted in three ways: (1) Replace the one-sample t value with the two-sample t value; (2) calculate N as $N_X N_Y/(N_X + N_Y)$, where X and Y denote the separate groups; and (3) calculate ν as $N_X + N_Y - 2$.

Now recall the data collected by Dr. Smith (see Table 1). Dr. Smith used a within-subjects design, and hence, a one-sample t test on the difference scores is appropriate. From the Bayes factor calculator provided on Rouder's Web site,[3] we obtain a Bayes factor of 6.08; this means that the data are about six times more likely under the null hypothesis than under the alternative hypothesis. When we assume that both hypotheses are equally likely a priori, we can compute $p(H_0/D)$, the posterior probability for the null hypothesis, as $6.08/7.08 \approx .86$.

Unfortunately, the test developed by Rouder et al. (2009) does not apply to the problem that confronts Dr. Smith. As was mentioned earlier, the SMM predicts that the effect will go in a specific direction—a direction other than the one that is observed in Dr. Smith's experiment. In order to calculate the Bayes factors that are appropriate for a one-sided test, we have developed a sampling-based alternative test.[4]

SD: An MCMC Sampling-Based *t* Test

Calculation of the SD t test involves four steps. The associated computer programs can be found on the first author's Web site.

Step 1: Rescaling the data. Prior to the analyses, we rescale the data such that one group has a mean of 0 and a standard deviation of 1. This scaling does not affect the test statistic. For the data from Dr. Smith, for instance, the *summer mean* of 0.07 is subtracted from all observations, both in the winter condition and in the summer condition. Next, all observations are divided by the *summer standard deviation*. The main advantage of this rescaling procedure is that the prior distributions for the parameters hold regardless of the scale of measurement: For our Bayesian SD test, it does not matter whether, say, response times are measured in seconds or in milliseconds.

Step 2: Defining prior distributions. We follow Rouder et al. (2009) and use a Cauchy(0,1) prior for effect size δ. For the standard deviation σ, we use a half-Cauchy(0,1) (Gelman & Hill, 2007)—that is, a Cauchy(0,1) distribution that is defined only for positive numbers. This choice for σ is reasonably uninformative, but—in contrast to Jeffrey's prior in Equation 6—the distribution is still proper (i.e., the area under the distribution is finite).[5] For the two-sample t test, we specify a Cauchy(0,1) prior for the grand mean μ.

Step 3: Obtaining posteriors using WinBUGS. The WinBUGS program[6] (Lunn, Thomas, Best, & Spiegelhalter, 2000) uses built-in Markov chain Monte Carlo techniques (MCMC; Gamerman & Lopes, 2006) to obtain samples from posterior distributions. After specifying the SD model in WinBUGS, the posterior distribution for effect size δ can be approximated to any desired degree of accuracy by increasing the number of samples. Because the SD model is relatively simple, we can draw as many as one million samples in a matter of minutes.

Step 4: Calculating Bayes factors using the SD density ratio. To obtain the Bayes factor, we use a method that is simple, intuitive, and flexible: the SD density ratio method (e.g., Dickey & Lientz, 1970; O'Hagan & Forster, 2004, pp. 174–177; Verdinelli & Wasserman, 1995). This method applies only to nested model comparisons, but it greatly simplifies the computation of the Bayes factor: The only information that is required is the height of the prior and the posterior distributions for the parameter of interest (i.e., δ) under the alternative hypothesis H_1 at the point that is subject to test. The reader who is not interested in the mathematical derivation may safely skip to Equation 10.

Let δ be the parameter of interest and σ the nuisance parameter. We assume, as is reasonable in many cases, that the conditional density for δ is continuous at $\delta = 0$, such that

$$\lim_{\delta\to 0} p(\sigma^2 | H_1, \delta) = p(\sigma^2 | H_0).$$

This means that the prior for the nuisance parameter in the complex model, conditional on $\delta \to 0$, equals the prior for the nuisance parameters in the simple model for which $\delta = 0$ by definition. We can then write $p(\sigma^2 | H_1, \delta = 0) = p(\sigma^2 | H_0)$, an equality that holds automatically when the prior distributions are specified to be independent.

The foregoing allows us to simplify the marginal likelihood for H_0 as follows:

$$\begin{aligned} p(D|H_0) &= \int_0^\infty f(D|H_0, \sigma^2) p(\sigma^2 | H_0) d\sigma^2 \\ &= \int_0^\infty f(D|H_1, \sigma^2, \delta = 0) p(\sigma^2 | H_1, \delta = 0) d\sigma^2 \\ &= p(D|H_1, \delta = 0). \end{aligned} \tag{8}$$

We now apply Bayes's rule to the results of Equation 8 and obtain

$$\begin{aligned} p(D|H_0) &= p(D|H_1, \delta = 0) \\ &= \frac{p(\delta = 0|H_1, D) p(D|H_1)}{p(\delta = 0|H_1)}. \end{aligned} \tag{9}$$

Dividing both sides of Equation 9 by $p(D | H_1)$ results in

$$BF_{01} = \frac{p(D|H_0)}{p(D|H_1)} = \frac{p(\delta = 0|H_1, D)}{p(\delta = 0|H_1)}. \tag{10}$$

This result is generally known as the SD density ratio (Dickey & Lientz, 1970; O'Hagan & Forster, 2004), and it shows that the Bayes factor equals the ratio of the posterior and prior ordinate under H_1 at the point of interest (i.e., $\delta = 0$). Note that there is no need to integrate out any model parameters, that the only distribution that matters is the one for the parameter of interest δ, and that the only hypothesis that needs to be considered is H_1. These are considerable simplifications, as compared with the standard procedure (cf. Equation 4).

Thus, Equation 10 shows that all that is required to compute the Bayes factor is the height of the prior and posterior distributions for δ at $\delta = 0$. The height of the prior distribution at $\delta = 0$ can be immediately computed from the Cauchy(0,1) distribution. The height of the posterior distribution at $\delta = 0$ can be easily estimated from the MCMC samples—for instance, by applying a nonparametric density estimator (e.g., Stone, Hansen, Kooperberg, & Truong, 1997) or a normal approximation to the posterior (i.e., parametric density estimation). The normal approximation is motivated by the Bayesian central limit theorem (Carlin & Louis, 2000, pp. 122–124), which states that under general regularity conditions, all posterior distributions tend to a normal distribution as the number of observations grows large.

Our experience with the SD test suggests that the difference between nonparametric and parametric estimation is negligible. In the work reported here, we choose to use the normal approximation because it is computationally more efficient. However, it is prudent to always plot the posterior distributions and check whether the posterior ordinate at $\delta = 0$ is estimated correctly. For practical applications, we also advise the user to use both the nonparametric and the parametric estimators and confirm that they yield approximately the same result.

The One-Sample SD *t* Test: Comparison With Rouder et al. (2009)

The one-sample *t* test is used to test whether the population mean of one particular sample of observations is equal to zero or not. In experimental psychology, the one-sample *t* test is often used for within-subjects designs, in which the scores for two conditions can be reduced to a single difference score.

In order to clarify the structure of the one-sample *t* test, we use graphical model notation (e.g., 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 indicate dependencies between the variables, with children depending on their parents. Double borders indicate that the variable under consideration is deterministic (i.e., they are calculated without noise from other variables), rather than stochastic. Finally, observed variables are shaded, and unobserved variables are not shaded. The graphical model for the one-sample *t* test is shown in Figure 1.

In the graphical model, X represents the observed data, distributed according to a normal distribution with a mean of μ_X and a variance of σ_X^2. Because $\delta = \mu_X/\sigma_X$, μ_X is given by $\mu_X = \delta \times \sigma_X$. The null hypothesis puts all prior mass for δ on a single point—that is, $H_0 : \delta = 0$—whereas the alternative hypothesis assumes that δ is Cauchy(0,1) distributed: $H_1 : \delta \sim$ Cauchy(0,1). It is relatively straightforward to implement this graphical model in WinBUGS, obtain samples from the posterior distribution for δ, and carry out the SD test.

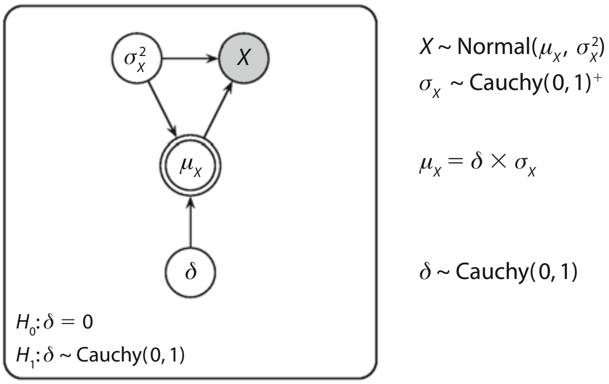
σ_X^2
X
μ_X
δ
$H_0: \delta = 0$
$H_1: \delta \sim$ Cauchy(0, 1)
$X \sim$ Normal(μ_X, σ_X^2)
$\sigma_X \sim$ Cauchy(0, 1)$^+$
$\mu_X = \delta \times \sigma_X$
$\delta \sim$ Cauchy(0, 1)

Figure 1. Graphical model for the Savage–Dickey one-sample *t* test. Cauchy(0,1)$^+$ denotes the half-Cauchy(0,1) defined for positive numbers only.

Because our SD *t* test is based on a sampling-based procedure that relies on the convergence of a stochastic process, it is desirable to verify whether the results of the SD test coincide with those from the JZS test, which is based on an analytical solution. This verification was carried out by means of a simulation study, the results of which are shown in Figure 2. We simulated 100 data sets by systematically increasing the difference between the group means to yield a set of 100 different *t* values. For each of the 100 data sets, we then compared the Bayes factor calculated by the JZS test with the SD Bayes factor. For all panels, the *x*-axis gives the *t* statistic, and the *y*-axis gives the associated posterior probability for the null hypothesis, $p(H_0 \mid D)$, derived from the Bayes factor under the assumption that H_0 and H_1 are equally likely a priori. Each panel shows the overlap between the JZS test and the SD test for a specific sample size (i.e., $N \in \{20,40,80,160\}$), on the basis of 100 simulated data sets. The results demonstrate that for the one-sample scenario, the SD test closely mimics the JZS test.

The Two-Sample SD *t* Test: Comparison With Rouder et al. (2009)

The two-sample *t* test is used to test whether the population means of two independent samples of observations are equal to each other or not. In experimental psychology, the two-sample *t* test is often used for between-subjects designs.

The graphical model for the two-sample *t* test is shown in Figure 3. The graphical model shows that X and Y represent the two groups of observed data. Both X and Y are distributed according to a normal distribution with shared variance σ^2. The mean of X is given by $\mu + \alpha/2$, and the mean of Y is given by $\mu - \alpha/2$.

Because $\delta = \alpha/\sigma$, α is given by $\alpha = \delta \times \sigma$. As for the one-sample scenario, the null hypothesis puts all prior mass for δ on a single point—that is, $H_0 : \delta = 0$—whereas the alternative hypothesis assumes that δ is Cauchy(0,1) distributed—$H_1 : \delta \sim$ Cauchy(0,1).

To compare this SD test with Rouder et al.'s (2009) JZS test, we conducted a simulation study, identical to the one-sample scenario in all respects except for the number of

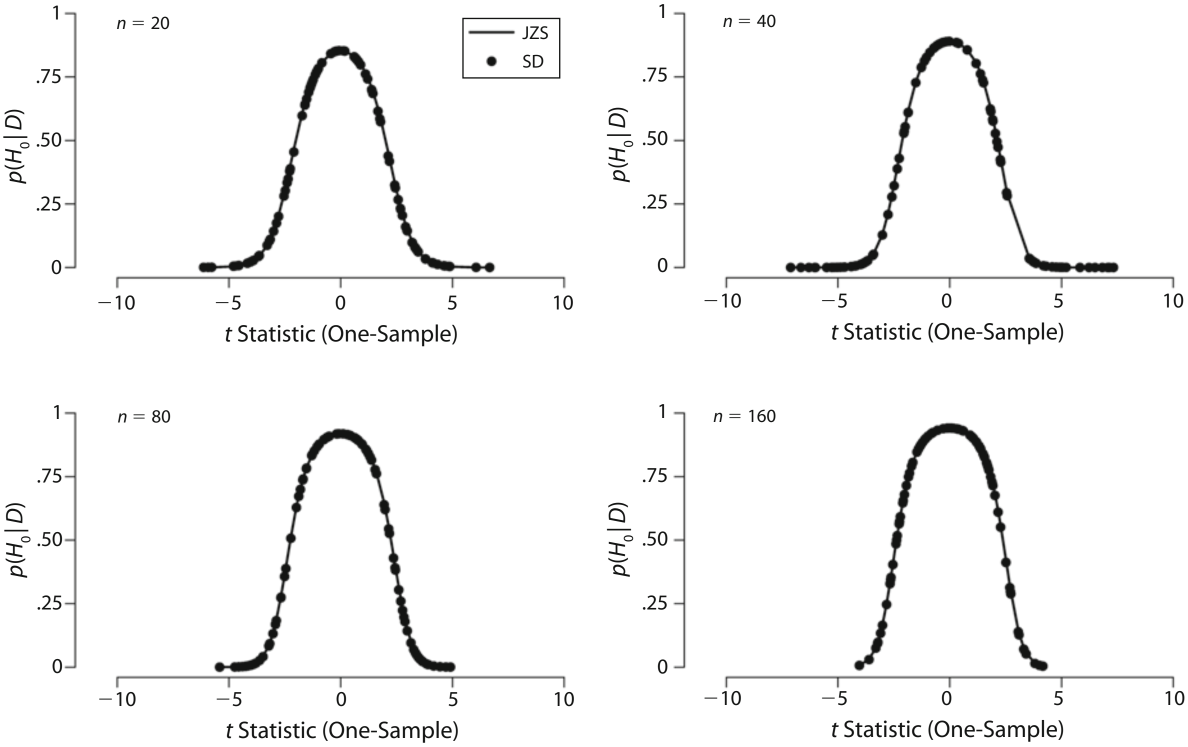

Figure 2. Comparison between the one-sample Savage–Dickey (SD) values and Jeffreys–Zellner–Siow (JZS) values for various sample sizes. The black dots represent the SD values, and the solid line represents the JZS values.

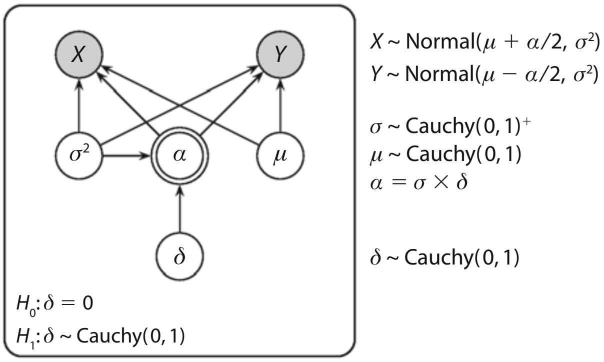
X
Y
σ^2
α
μ
δ
H_0: $\delta = 0$
H_1: $\delta \sim$ Cauchy(0, 1)
$X \sim$ Normal($\mu + \alpha/2$, σ^2)
$Y \sim$ Normal($\mu - \alpha/2$, σ^2)
$\sigma \sim$ Cauchy(0, 1)$^+$
$\mu \sim$ Cauchy(0, 1)
$\alpha = \sigma \times \delta$
$\delta \sim$ Cauchy(0, 1)

Figure 3. Graphical model for the Savage–Dickey two-sample *t* test. Cauchy(0,1)$^+$ denotes the half-Cauchy(0,1) defined for positive numbers only.

groups. The results of this simulation study are shown in Figure 4. The results demonstrate that for the two-sample scenario, the SD test closely mimics the JZS test.

Extension 1: Order Restrictions

Recall once again the experiment by Dr. Smith (see Table 1). The SMM predicted that the effect of glucose would be larger in summer than in winter. We now show how the SD test can be used to test such order-restricted hypotheses, allowing Dr. Smith to quantify exactly the extent to which the data support the null hypothesis versus the alternative SMM hypothesis.

The top panel of Figure 5 shows the unrestricted prior and posterior distributions for δ for the data from Dr. Smith. Negative values of δ indicate that the effect of glucose is larger in summer than in winter. From the SD method, we can compute the Bayes factor in favor of $H_0 : \delta = 0$ versus the unrestricted alternative $H_1 : \delta \neq 0$, instantiated as $\delta \sim$ Cauchy(0,1). Note that the result—$BF_{01} = 6.08$—is identical to the Bayes factor that is obtained from the JZS test: The data are about six times more likely under H_0 than under H_1.

The middle panel of Figure 5 shows the *SD* test that applies to the prediction that Dr. Smith seeks to test—that is,

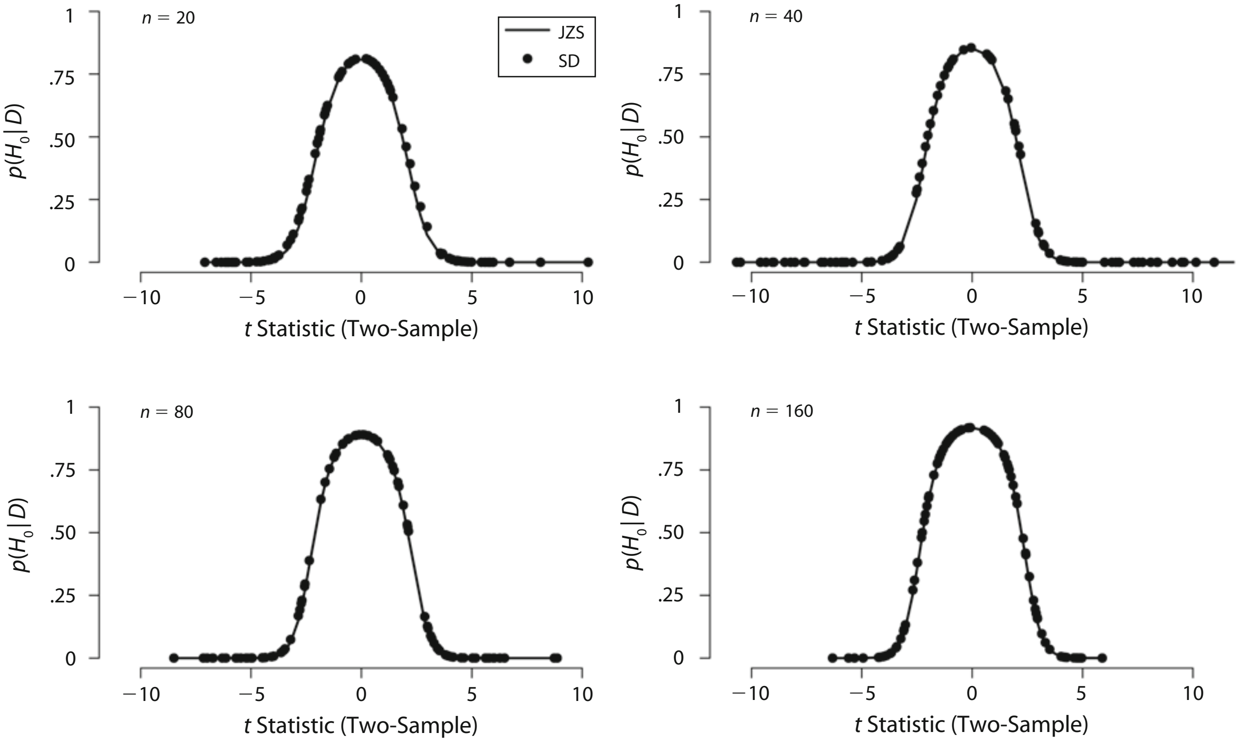

Figure 4. Comparison between the two-sample Savage–Dickey (SD) values and Jeffreys–Zellner–Siow (JZS) values for various sample sizes. The black dots represent the SD values, and the solid line represents the JZS values.

$H_0 : \delta = 0$ versus the order-restricted hypothesis $H_1 : \delta < 0$, instantiated as $\delta \sim$ Cauchy(0,1)$^-$, a half-Cauchy(0,1) distribution that is defined only for negative numbers. In order to calculate the height of the order-restricted posterior distribution at $\delta = 0$, we focus solely on that part of the unrestricted posterior for which $\delta < 0$. After renormalizing, we obtain a truncated but proper posterior distribution that ranges from $\delta = -\infty$ to $\delta = 0$. Figure 5 shows both the half-Cauchy(0,1) prior (solid line) and the truncated posterior (dashed line). The SD ratio at $\delta = 0$ yields a Bayes factor of $BF_{01} = 13.75$. This means that the data are almost 14 times more likely under H_0 than under the order-restricted H_1 that is associated with the SMM. When H_0 and H_1 are equally likely a priori, the posterior probability in favor of the null hypothesis is about $13.75/14.75 \approx .93$, which is considered *positive evidence* for the null hypothesis (Raftery, 1995; Wagenmakers, 2007).

For completeness, the bottom panel of Figure 5 shows the SD test for the alternative order restriction. In this case, we seek to test $H_0 : \delta = 0$ versus $H_1 : \delta > 0$, instantiated as $\delta \sim$ Cauchy(0,1)$^+$, a half-Cauchy(0,1) distribution that is defined only for positive numbers. The SD density ratio yields a Bayes factor of $BF_{01} = 3.91$, which indicates that the data are almost four times more likely under H_0 than under H_1.

Extension 2: Variances Free to Vary in the Two-Sample *t* Test

For the two-sample scenario, the JZS test assumes that the separate samples share a common unknown variance. When this assumption is false and both groups have unequal numbers of observations, results of the JZS *t* test should be interpreted with care.

This complication (i.e., testing for the difference of two normal means with unequal variances) is known as the Behrens–Fisher problem, and it is one of the oldest problems in statistics. Within the paradigm of *p* value hypothesis testing, several solutions to the Behrens–Fisher problem have been proposed (Kim & Cohen, 1998). These solutions (i.e., corrections for unequal variances) have been implemented in popular statistical software packages such as SPSS and R.

In order to address the Behrens–Fisher problem, we adjusted the SD test in two ways. First, as is illustrated in Figure 6, each of the two groups now has its own variance. Second, the previous relation $\alpha = \delta \times \sigma$ no longer holds, since we now have two σ parameters. We use a standard solution and calculate the pooled standard deviation (Hedges, 1981):

$$\alpha = \delta \times \sqrt{\frac{\left[\sigma_1^2 \times (n_1 - 1)\right] + \left[\sigma_2^2 \times (n_2 - 1)\right]}{n_1 + n_2 - 2}}. \quad (11)$$

After implementing these changes, calculation of the Bayes factor proceeds in the same fashion as before.

To illustrate the behavior of the separate variance SD Bayes factors, we follow Moreno, Bertolino, and Racugno (1999) and apply the tests to hypothetical data from Box and Tiao (1973, p. 107). These data have the following properties: $n_1 = 20$, $\text{var}_1 = 12$, $n_2 = 12$, and $\text{var}_2 = 40$. As can be seen from Table 2, the support for the null hypothesis decreases as the difference in group means increases. The separate variance SD test tends to favor the null hypothesis more than does the shared variance SD test, although the difference is small. The intrinsic Bayes factor (i.e., a default Bayes factor that uses minimal training samples and uninformative priors; Berger & Pericchi, 1996; Moreno et al., 1999) supports the null hypothesis the most. A more detailed treatment of the Behrens–Fisher problem is beyond the scope of the present article; we include it here only to highlight the flexibility of the SD test.

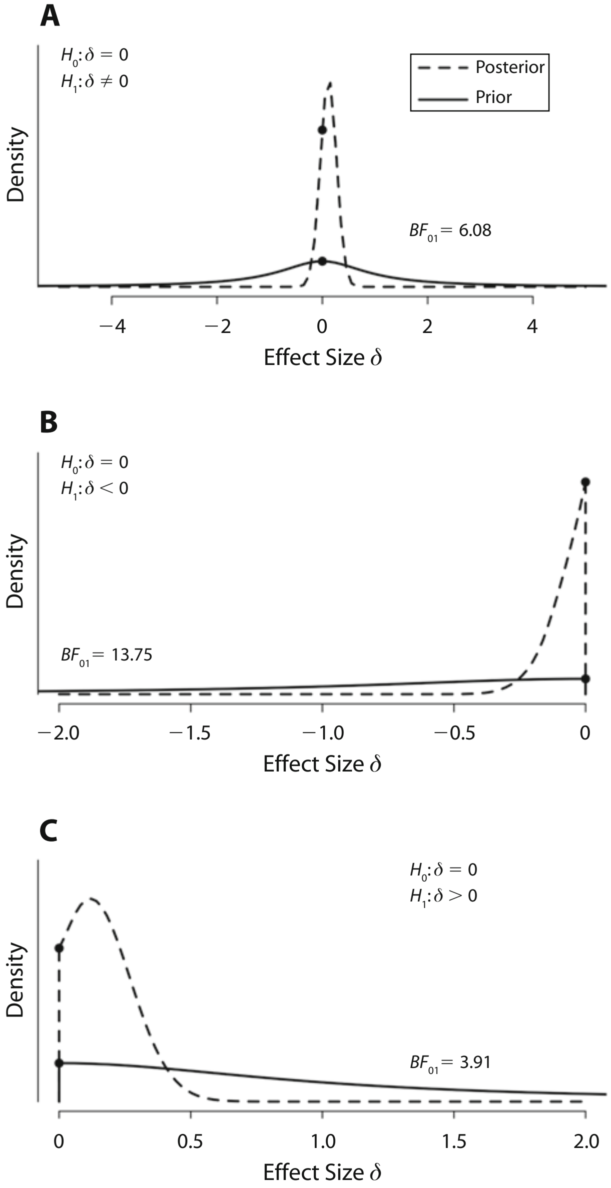

Figure 5. The prior and posterior distributions of effect size δ, based on the data from Dr. Smith (see Table 1). The top panel illustrates the unrestricted Savage–Dickey (SD) test, the middle panel illustrates the order-restricted test associated with the seasonal memory model, and the bottom panel illustrates the SD test for the alternative order restriction. The dots mark the height of the prior and posterior distributions at $\delta = 0$.

Summary and Conclusion

In this article, we have developed an SD Bayesian *t* test that extends the Bayesian JZS *t* test recently proposed by Rouder et al. (2009). Our sampling-based SD test can handle order restrictions and addresses the situation in which two groups have unequal variance.

One of the advantages of the SD test is its flexibility; for instance, it would be trivial to replace the default priors with priors that are informed by previous experiments or detailed expert knowledge about the problem at hand. We chose to use the Cauchy(0,1) prior for effect size δ, as proposed by Rouder et al. (2009), but many more prior distributions are possible. For example, Killeen (2007) argued that, on the basis of extensive research in social psychology (Richard, Bond, & Stokes-Zoota, 2003), the distribution of effect sizes is normally distributed with a variance of 0.3.

Another advantage of the SD test, and Bayesian methods in general, is that they allow for *sequential inference*. As has been stated by Edwards, Lindman, and Savage (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" (p. 193). More concretely, this means that one can apply the SD *t* test and monitor the resulting Bayes factor after every new subject, stopping data collection whenever the evidence is sufficiently compelling. Note that within

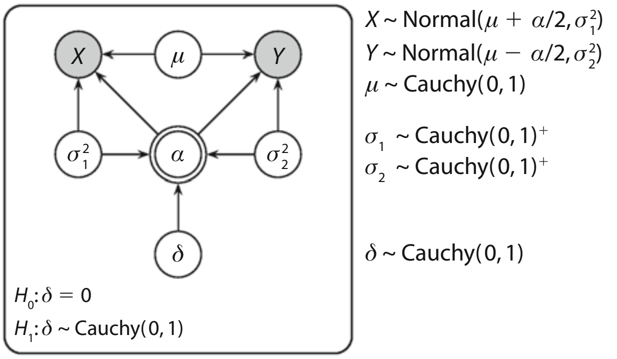
X
μ
Y
σ_1^2
α
σ_2^2
δ
H_0: $\delta = 0$
H_1: $\delta \sim$ Cauchy(0, 1)
$X \sim$ Normal($\mu + \alpha/2, \sigma_1^2$)
$Y \sim$ Normal($\mu - \alpha/2, \sigma_2^2$)
$\mu \sim$ Cauchy(0, 1)
$\sigma_1 \sim$ Cauchy(0, 1)$^+$
$\sigma_2 \sim$ Cauchy(0, 1)$^+$
$\delta \sim$ Cauchy(0, 1)

Figure 6. Graphical model for Rouder's default Bayesian two-sided *t* test with unequal variances.

Table 2
Comparison of Savage–Dickey (SD) Bayes Factors With the Intrinsic Bayes Factor for Hypothetical Data Reported in Box and Tiao (1973, p. 107) and Analyzed in Moreno, Bertolino, and Racugno (1999)

$\bar{X} - \bar{Y}$	$BF_{01}^{\mathrm{SD}1\sigma}$	$BF_{01}^{\mathrm{SD}2\sigma}$	BF_{01}^{I}
0.00	3.93	3.36	5.00
2.20	2.08	2.16	2.86
4.22	0.45	0.81	0.76
5.00	0.21	0.51	0.40
10.0	<0.02	<0.02	<0.02

Note—$BF_{01}^{\mathrm{SD}1\sigma}$ denotes the SD Bayes factor using a shared variance, $BF_{01}^{\mathrm{SD}2\sigma}$ denotes the SD Bayes factor using two separate variances, and BF_{01}^{I} denotes the intrinsic Bayes factor reported by Moreno et al. (1999).

the paradigm of p value hypothesis testing, such practice amounts to cheating; with enough time, money, and patience, *optional stopping* is guaranteed to yield a significant result (for a discussion, see Wagenmakers, 2007).

Here, we have limited ourselves to the t test. Nevertheless, the SD idea is quite general, and it can facilitate Bayesian hypothesis testing for a wide 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) or GCM (Nosofsky, 1986), multinomial processing trees (Batchelder & Riefer, 1999), the ACT–R model (Weaver, 2008), and many more. Another exciting possibility is to apply the SD method to facilitate Bayesian hypothesis testing in hierarchical models (i.e., models with random effects for subjects or items) such as those advocated by Rouder and others (Rouder & Lu, 2005; Rouder, Lu, Morey, Sun, & Speckman, 2008; Rouder et al., 2007; Shiffrin, Lee, Kim, & Wagenmakers, 2008).

For example, one might wish to study the effect of an antidepressant on the parameters of the Ratcliff diffusion model. Specifically, the hypothesis of interest may hold that the antidepressant decreases response caution a. This means that $H_0 : \delta = 0$ and $H_1 : \delta > 0$, where δ indicates the difference in response caution ($\delta = \alpha_{\text{off}} - a_{\text{on}}$) between people who are either on or off medication. Standard approaches for computing the Bayes factor require that one integrates out all the other parameters of the diffusion model (i.e., drift rate, nondecision time, starting point, the probability of a response contaminant, and the across-trial variabilities), separately for H_0 and H_1. In contrast, the SD approach requires one only to estimate the height of the posterior distribution at $\delta = 0$—a considerable simplification.

In closing, we agree with Rouder et al. (2009) that many scientific hypotheses are formulated in terms of invariances and that invariances can be formulated in terms of statistical null hypotheses (Wagenmakers, Lee, Lodewyckx, & Iverson, 2008). To quantify the statistical evidence in favor of such substantive null hypotheses, we need to move away from p value hypothesis testing (with which one can only *fail to reject* a null hypothesis) and move toward Bayesian hypothesis testing. In this article, we have discussed a related problem of considerable scientific importance: A substantive hypothesis (i.e., the SMM) makes a specific prediction, and falsification of the theory requires that one is able to quantify the support in favor of the null hypothesis.

We believe not only that Bayesian hypothesis testing provides a coherent framework to quantify knowledge and uncertainty, but also that it addresses the kinds of questions that experimental psychologists would like to see answered. Bayesian t tests such as Rouder et al.'s (2009) JZS test and our SD test are the first steps toward a more rational and informative method for testing statistical hypotheses in psychology.

AUTHOR NOTE

This research was supported by a Vidi grant from the Dutch Organization for Scientific Research (NWO). The SD computer code and the data of Dr. Smith's hypothetical experiment can be found on the first author's Web site, www.ruudwetzels.com. Correspondence concerning this article may be addressed to R. Wetzels or E.-J. Wagenmakers, Department of Psychology, University of Amsterdam, Roetersstraat 15, 1018 WB Amsterdam, The Netherlands (e-mail: wetzels.ruud@gmail.com).

REFERENCES

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., & Jefferys, W. H. (1992). The application of robust Bayesian analysis to hypothesis testing and Occam's razor. *Statistical Methods & Applications*, **1**, 17-32.

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.

Box, G. E. P., & Tiao, G. C. (1973). *Bayesian inference in statistical analysis*. Reading, MA: Addison-Wesley.

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., & Louis, T. A. (2000). *Bayes and empirical Bayes methods for data analysis* (2nd ed.). London: Chapman & Hall.

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

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, **70**, 193-242.

Gamerman, D., & Lopes, H. F. (2006). *Markov chain Monte Carlo: Stochastic simulation for Bayesian inference*. Boca Raton, FL: Chapman & Hall/CRC.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.

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

Hedges, L. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational & Behavioral Statistics*, **6**, 107.

Hoijtink, H., Klugkist, I., & Boelen, P. (2008). *Bayesian evaluation of informative hypotheses that are of practical value for social scientists*. New York: Springer.

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

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

Killeen, P. R. (2007). Replication statistics as a replacement for significance testing: Best practices in scientific decision-making. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 103-124). Thousand Oaks, CA: Sage.

Kim, S., & Cohen, A. (1998). On the Behrens-Fisher problem: A review. *Journal of Educational & Behavioral Statistics*, **23**, 356-377.

Klugkist, I., Laudy, O., & Hoijtink, H. (2005). Inequality constrained analysis of variance: A Bayesian approach. *Psychological Methods*, **10**, 477-493.

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: Oxford University Press, Clarendon Press.

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

Liang, F., Paulo, R., Molina, G., Clyde, M., & Berger, J. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association*, **103**, 410.

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000).

WinBUGS—a Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics & Computing*, **10**, 325-337.

Moreno, E., Bertolino, F., & Racugno, W. (1999). Default Bayesian analysis of the Behrens–Fisher problem. *Journal of Statistical Planning & Inference*, **81**, 323-333.

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

Nosofsky, R. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, **115**, 39-57.

O'Hagan, A., & Forster, J. (2004). *Kendall's advanced theory of statistics: Vol. 2B. Bayesian inference* (2nd ed.). London: Arnold.

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

Richard, F. D., Bond, C. F. J., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology*, **7**, 331-363.

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.

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

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). *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). Austin, TX: 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.

Wasserman, L. (2000). Bayesian model selection and model averaging. *Journal of Mathematical Psychology*, **44**, 92-107.

Weaver, R. (2008). Parameters, predictions, and evidence in computational modeling: A statistical view informed by ACT–R. *Cognitive Science*, **32**, 1349-1375.

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*.

Wilkinson, L., & the Task Force on Statistical Inference (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, **54**, 594-604.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with *g*-prior distributions. In P. K. Goel & A. Zellner (Eds.), *Bayesian inference and decision techniques: Essays in honor of Bruno de Finetti* (pp. 233-243). Amsterdam: North-Holland.

Zellner, A., & Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, & A. F. M. Smith (Eds.), *Bayesian statistics* (pp. 585-603). Valencia: Valencia University Press.

NOTES

1. This quote is taken from an actual review.

2. The absolute posterior model probabilities hold only when H_0 and H_1 are the sole two models under consideration.

3. Available at http://pcl.missouri.edu/bayesfactor.

4. There may or may not be an analytical solution to the order-restricted problem, and here we do not attempt to derive such a solution. Instead, the goal is to illustrate the flexibility of the SD test using the order-restricted hypothesis test as an example.

5. This is helpful since WinBUGS does not allow the specification of improper priors. In any case, because sigma is a nuisance parameter in this model, the prior for sigma has a negligible effect on the calculation of the Bayes factor.

6. WinBUGS is easy to learn and is supported by a large community of active researchers (see www.mrc-bsu.cam.ac.uk/bugs).

(Manuscript received November 7, 2008;
revision accepted for publication March 10, 2009.)