Behav Res (2011) 43:679–690
DOI 10.3758/s13428-010-0049-5

A tutorial on a practical Bayesian alternative to null-hypothesis significance testing

Michael E. J. Masson

Published online: 8 February 2011

Abstract Null-hypothesis significance testing remains the standard inferential tool in cognitive science despite its serious disadvantages. Primary among these is the fact that the resulting probability value does not tell the researcher what he or she usually wants to know: How probable is a hypothesis, given the obtained data? Inspired by developments presented by Wagenmakers (*Psychonomic Bulletin & Review, 14*, 779–804, 2007), I provide a tutorial on a Bayesian model selection approach that requires only a simple transformation of sum-of-squares values generated by the standard analysis of variance. This approach generates a graded level of evidence regarding which model (e.g., effect absent [null hypothesis] vs. effect present [alternative hypothesis]) is more strongly supported by the data. This method also obviates admonitions never to speak of accepting the null hypothesis. An Excel worksheet for computing the Bayesian analysis is provided as supplemental material.

Keywords Bayesian analysis · Null-hypothesis significance testing

The widespread use of null-hypothesis significance testing (NHST) in psychological research has withstood numerous rounds of debate (e.g., Chow, 1998; Cohen, 1994; Hagen, 1997; Krueger, 2001; Nickerson, 2000) and continues to be the field's most widely applied method for evaluating evidence. There are, however, clear shortcomings attached to standard applications of NHST. In this article, I briefly review the most important of these problems and then present a summary of a Bayesian alternative developed by Wagenmakers (2007). My goal is to provide readers with a highly accessible, practical tutorial on the Bayesian approach that Wagenmakers presented. Motivation for such a tutorial comes from the fact that in the time since publication of Wagenmakers's article, very little appears to have changed with respect to the reliance on NHST methods in the behavioral sciences. I suspect that one reason that Bayesian methods have not been more readily adopted is that researchers have failed to recognize the availability of a sufficiently simple means of computing and interpreting a Bayesian analysis. This tutorial is intended as a remedy for that obstacle.

Electronic supplementary material The online version of this article (doi:10.3758/s13428-010-0049-5) contains supplementary material, which is available to authorized users.

M. E. J. Masson (✉)
Department of Psychology, University of Victoria,
Room A234, Cornett Building, P.O. Box 3050 STN CSC,
Victoria, BC V8W 3P5, Canada
e-mail: mmasson@uvic.ca

Serious problems with NHST

The p value generated by NHST is misleading, particularly in the sense that it fails to provide the information that a researcher actually wants to have. That is, the NHST p value is a conditional probability that indicates the likelihood of an observed result (or any more extreme result), given that the null hypothesis is correct: $p(D|H_0)$. Researchers draw inferences from this conditional probability regarding the status of the null hypothesis, reasoning that if the p value is very low, the null hypothesis can be rejected. But what are the grounds for this rejection? One possibility is that a low p value signals that the null hypothesis is unlikely to be true. This conclusion, however, does not follow directly from a low p value (see Cohen, 1994). The probability of the null hypothesis being true, given the obtained data, would be expressed as $p(H_0|D)$, which is not equivalent to $p(D|H_0)$. The only way to translate $p(D|H_0)$, which is readily found by applying

NHST methods, into $p(H_0|D)$ is by applying the Bayes theorem. Trafimow and Rice (2009) did just that in a Monte Carlo simulation to investigate the extent to which $p(D|H_0)$ and $p(H_0|D)$ are correlated, as they must be if one is to justify decisions about the null hypothesis under NHST. Their results indicated a correlation of only .396 between these two measures. When they dichotomized the $p(D|H_0)$ values as significant or not significant, on the basis of the typical cutoff of .05, the correlation between the dichotomized category and the $p(H_0|D)$ values dropped to .289. Surprisingly, this correlation fell even more when more stringent cutoff values (.01 and .001) were applied. Thus, it is far from safe to assume that the magnitude of $p(H_0|D)$ can be directly inferred from $p(D|H_0)$. The Bayesian approach proposed by Wagenmakers (2007) eliminates this problem by directly computing $p(H_0|D)$.

A second concern with NHST stems from the fact that researchers are permitted to make one of two decisions regarding the null hypothesis: reject or fail to reject. This approach to hypothesis testing does not provide a means of quantifying evidence in favor of the null hypothesis and even prohibits the concept of accepting the null hypothesis (Wilkinson & the Task Force on Statistical Inference, 1999). Thus, researchers are typically left with little to say when their statistical test fails to reject the null hypothesis. The Bayesian approach, because it can directly evaluate the relative strength of evidence for the null and alternative hypotheses, provides clear quantification of the degree to which the data support either hypothesis.

Fundamentals of a Bayesian alternative to NHST

The first point to note about the Bayesian approach to hypothesis testing is that, unlike NHST, this approach does not rest on a binary decision about a single (null) hypothesis. Rather, the Bayesian approach is essentially a model selection procedure that computes a comparison between competing hypotheses or models. For most applications in which NHST is currently the method of choice, the Bayesian alternative consists of a comparison between two models, which we can continue to characterize as the null and alternative hypotheses, or which can be two competing models that assume different configurations of effects. Non-Bayesian methods involving a similar model comparison approach have also been developed and have their own merits. For example, Dixon (2003) and Glover and Dixon (2004) proposed using likelihood ratios as a means of determining which of two models is more likely given a set of data. The likelihood ratio is equal to the likelihood of the observed data given one model, relative to the likelihood of the data given a competing model. As with the Bayesian method, likelihood ratios provide graded evidence concerning which model is more strongly supported by the data, rather than a thresholded binary decision. In the method advocated by Glover and Dixon (2004), the Akaike information criterion (Akaike, 1974) may be used to adjust the likelihood ratio to take into account the different number of parameters on which competing models are based. Glover and Dixon (2004) also pointed out that one could use the Bayesian information criterion (BIC; described more fully below), instead of the AIC, in generating a likelihood ratio that takes into account differences in the number of free parameters. In this sense, the Glover and Dixon (2004) formulation anticipates the approach developed by Wagenmakers (2007), which can be seen as a special case of the more general likelihood-ratio method proposed by Glover and Dixon (2004). Unlike the Bayesian method, evaluating models on the basis of a likelihood ratio does not involve computing a posterior probability such as $p(H_0|D)$. Rather, the ratio itself is taken as a measure of the strength of evidence favoring one model over another. The likelihood ratio can be computed for various experimental designs as readily as the Bayesian analysis described here (Bortolussi & Dixon, 2003; Dixon, 2003; Dixon & O'Reilly, 1999; Glover & Dixon, 2004), and therefore, it constitutes another practical alternative to NHST.

In the Bayesian model selection procedure, the ultimate objective is to compute a probability reflecting which model is more likely to be correct, on the basis of the obtained data. The core concept in this method is the Bayes theorem, which can be expressed as

$$p(H|D) = \frac{p(D|H) \cdot p(H)}{p(D)}, \tag{1}$$

where $p(H)$ is the a priori probability that the hypothesis is correct and $p(D)$ is the probability of obtaining the observed data independently of any particular hypothesis. It is readily apparent that this expression yields the type of information that researchers seek—namely, the posterior probability that a hypothesis is correct given a set of observed data: $p(H|D)$. In the Bayesian approach developed by Wagenmakers (2007), the relative posterior probability of the null and alternative hypotheses is computed, yielding the odds in favor on one hypothesis over the other:

$$\frac{p(H_0|D)}{p(H_1|D)} = \frac{\frac{p(D|H_0)\cdot p(H_0)}{p(D)}}{\frac{p(D|H_1)\cdot p(H_1)}{p(D)}}. \tag{2}$$

Equation 2 can be simplified by canceling the $p(D)$ terms common to the upper and lower halves of the right side of the equation to generate

$$\frac{p(H_0|D)}{p(H_1|D)} = \frac{p(D|H_0)}{p(D|H_1)} \cdot \frac{p(H_0)}{p(H_1)}. \tag{3}$$

The posterior odds (left side of Eq. 3), then, are determined by the product of what is called the *Bayes factor* (first term on the right side of the equation) and the prior odds. The posterior odds give the relative strength of evidence in favor of H_0 relative to H_1.

Furthermore, if it is assumed that the prior odds equal 1 (i.e., the two hypotheses are deemed equally likely before the data are collected), the posterior odds are equal to the Bayes factor. It is clear, then, that the Bayes factor plays a crucial role in establishing the relative evidential support for the null and alternative hypotheses. One might argue that it is not reasonable to hold that H_0 and H_1 are equally plausible a priori when one of the models is the null hypothesis, which specifies a precise effect size of 0, whereas the competing model allows a range of possible effect sizes. Moreover, it seems unlikely, for example, that two populations would have identical means, so the exact value specified by H_0 is unlikely ever to be literally correct. There are a number of possible responses to this concern. First, the Bayesian analysis holds either for the case where H_0 specifies an effect size of exactly 0 or for a distribution of very small effect sizes, centered at 0, that would require a very large sample size to detect (Berger & Delampady, 1987; Iverson, Wagenmakers, & Lee, 2010), which usually is what one has in mind when arguing that H_0 is never precisely correct (e.g., Cohen, 1994). Second, theoretical models tested in experiments often predict precisely no effect, setting up the classic H_0 as a viable model with a precise effect size of 0 (Iverson et al., 2010). Third, one has the option to specify prior odds other than 1 and to apply simple algebra to adjust the posterior odds accordingly before converting those odds to posterior probabilities. Finally, one could forego computation of the posterior probabilities and simply take the Bayes factor as a measure of the relative adequacy of the two competing models, which would be analogous to the method of using likelihood ratios discussed earlier (e.g., Glover & Dixon, 2004).

The difficulty we face in computing the Bayes factor is that although the probability of obtaining the observed data, given the null hypothesis, can be computed rather easily (akin to what is done in NHST), the corresponding conditional probability based on the alternative hypothesis is another matter. Unlike the null hypothesis, the alternative hypothesis does not specify one particular a priori value for the effect in question. Rather, the alternative hypothesis is associated with a distribution of possible effect sizes, and the value of the Bayes factor depends on the nature of that distribution. Therefore, exact computation of the Bayes factor quickly becomes complex, involving integration over the space of possible effect sizes using procedures such as Markov chain Monte Carlo methods (e.g., Kass & Raftery, 1995). These are not methods that most experimental psychologists are readily equipped to apply.

The much more practical alternative described by Wagenmakers (2007) is to generate an estimate of the Bayes factor using the BIC. This measure is often used to quantify a formal model's goodness of fit to data, taking into account the number of free parameters in the model. Of crucial importance in the present context, the difference in BIC values for two competing models (e.g., null vs. alternative hypotheses) can be transformed into an estimate of the Bayes factor. The BIC value for a model or hypothesis is defined as

$$\mathrm{BIC} = -2\ln(L) + k\ \ln(n), \tag{4}$$

where ln is the natural logarithm function, L is the maximum likelihood of the data assuming that the model is correct (see below), k is the number of free parameters in the model, and n is the number of independent observations. The Bayes factor (BF) can be estimated using the following transformation of the difference in BIC values for two competing models:

$$\mathrm{BF} \approx \frac{p_{\mathrm{BIC}}(\mathrm{D}|\mathrm{H}_0)}{p_{\mathrm{BIC}}(\mathrm{D}|\mathrm{H}_1)} = \mathrm{e}^{(\Delta\mathrm{BIC})/2}, \tag{5}$$

where $\Delta\mathrm{BIC} = \mathrm{BIC}(\mathrm{H}_1) - \mathrm{BIC}(\mathrm{H}_0)$. The resulting estimate of the Bayes factor yields the odds favoring the null hypothesis, relative to the alternative hypothesis. BF can then be converted to the posterior probability that the data favor the null hypothesis as follows (assuming equal priors, as discussed above):

$$p_{\mathrm{BIC}}(\mathrm{H}_0|\mathrm{D}) = \frac{\mathrm{BF}}{\mathrm{BF}+1}. \tag{6}$$

Note that I have used the symbol p_{BIC} to denote a posterior probability generated by the BIC estimate of the Bayes factor. There being only two candidate models, the posterior probability that the data favor the alternative hypothesis is just the complement of Eq. 6:

$$p_{\mathrm{BIC}}(\mathrm{H}_1|\mathrm{D}) = 1 - p_{\mathrm{BIC}}(\mathrm{H}_0|\mathrm{D}). \tag{7}$$

Before describing a practical method for computing BIC values, or at least ΔBIC (the critical element in Eq. 5), and the resulting posterior probabilities, an important caveat is in order. Computation of the Bayes factor depends on the specification of a prior distribution for the effect size parameter that distinguishes the alternative hypothesis from the null hypothesis. That is, assuming (as per the alternative hypothesis) that there is some nonzero effect size, what are the probable values of this effect size? These values cover some distribution whose characteristics influence the eventual posterior odds. The method of estimating the Bayes factor implemented in Eq. 5 is consistent with a prior distribution of possible effect size parameter values known as the *unit information prior* (Kass & Wasserman, 1995).

This distribution is the standard normal distribution centered at the value of the effect size observed in the data and extending over the distribution of observed data (Raftery, 1999). Because the unit information prior is normal in shape, its coverage emphasizes the plausible values of the effect size parameter without being excessively spread out (i.e., very little likelihood is associated with the more extreme values of effect size). As Raftery (1999) pointed out, this is a reasonable prior distribution in the sense that a researcher likely has some idea in advance of the general range in which the observed effect size is likely to fall and so will not put much prior probability outside that range. The application of Eq. 5 to estimate the Bayes factor, then, implicitly assumes the unit information prior as the distribution of the effect size parameter under the alternative hypothesis. It should be noted, however, that the unit information prior is more spread out than other potential prior distributions that could be informed by more specific a priori knowledge of the likely effect size. Prior distributions with greater spread tend to favor the null hypothesis more than do prior distributions with less spread. In this sense, the BIC estimate of the Bayesian posterior probabilities should be considered somewhat conservative with respect to providing evidence for the alternative hypothesis (Raftery, 1999).

Computation of posterior probabilities

I now turn to the question of how to compute BIC values associated with a particular hypothesis. The difficult term in Eq. 4 is L, the maximum likelihood of the data. Wagenmakers (2007) explains that if we assume normally distributed errors of measurement—a fundamental assumption in the standard use of analysis of variance (ANOVA) and regression—we have the following variant of Eq. 4:

$$\mathrm{BIC} = n \ln(1 - R^2) + k \ln(n), \tag{8}$$

where $1 - R^2$ is the proportion of variability that the model fails to explain. Recall that in the earlier definition of BIC (Eq. 4), n was defined as the number of independent observations in the data. Wagenmakers instead defines n, when used in Eq. 8, as the number of subjects. This difference in definition causes no difficulty for independent-samples designs, which are the only type of design considered in the Wagenmakers article. The issue of how to define n becomes more complicated, however, when one considers repeated measures factors. In a repeated measures design where each subject has c scores, with c indicating the number of conditions in which each subject was tested and s indicating the number of subjects, the number of independent observations is actually $n = s(c - 1)$. Indeed, this is how Dixon and colleagues defined n when repeated measures factors were involved in their computation of likelihood ratios (e.g., Bortolussi & Dixon, 2003). The issue of whether, in repeated measures designs, the value of n should be the number of subjects or the number of independent observations appears at present to be unsettled (Wagenmakers, personal communication, October 19, 2010). But given that the interpretation of n for the log likelihood (the first term in Eq. 8) is the number of independent observations, and given that there is no indication that the n in the second term of Eq. 8 represents a different quantity than does the n in the first term, I have opted to use the number of independent observations interpretation for n throughout.

Now, the link between Eq. 8 and ANOVA lies in the following fact:

$$(1 - R^2) = \frac{SSE}{SS_{\mathrm{total}}}, \tag{9}$$

where SSE is the sum of squares for the error term. Suppose that we obtain the BIC values for the alternative and the null hypotheses, using the relevant SS terms. When computing $\Delta\mathrm{BIC} = \mathrm{BIC}(\mathrm{H}_1) - \mathrm{BIC}(\mathrm{H}_0)$, note that both the null and the alternative hypothesis models share the same SS_{total} term (both models are attempts to explain the same collection of scores), although they differ with respect to SSE. In particular, SSE will be larger for the null hypothesis model because it uses one less free parameter to fit the data (the effect size parameter is fixed at 0 for the null model). The SS_{total} term common to both BIC values cancels out in computing ΔBIC, producing

$$\Delta\mathrm{BIC} = n \ln\left(\frac{SSE_1}{SSE_0}\right) + (k_1 - k_0)\ln(n), \tag{10}$$

where SSE_1 and SSE_0 are the sum of squares for the error terms in the alternative and the null hypothesis models, respectively. Note that the term SSE_1/SSE_0 is just the complement of partial eta-squared (η_p^2), a measure of effect size corresponding to the proportion of variability accounted for by the independent variable (i.e., $SSE_1/SSE_0 = 1 - \eta_p^2$). The $k_1 - k_0$ term represents the difference in the number of free parameters between the two models. This difference will be equal to the degrees of freedom associated with an effect when null and alternative hypotheses are contrasted. For example, when the difference between two condition means are tested in an ANOVA, the alternative hypothesis has one additional free parameter (the size of the difference between the two condition means), relative to the null hypothesis, so that, in that case, $k_1 - k_0 = 1$. Additional cases are described in the examples provided below. Once ΔBIC has been computed, Equations 5 through 7 may be applied to yield the posterior probabilities for the two competing hypotheses.

Example applications of the Bayesian method

Example 1 It may now be apparent that to implement the Bayesian test described here for experimental designs that fit the ANOVA paradigm requires a relatively simple transformation of sum-of-squares values generated by the standard ANOVA. I will illustrate this application using three examples with actual published data. For the first case, consider an experiment on long-term priming in an object identification task (Breuer, Masson, Cohen, & Lindsay, 2009, Experiment 2A). In a study phase, subjects searched rapidly presented lists of pictures for a target object. Nontarget items from these trials then served as primed targets on a masked identification task in which a target was briefly presented and masked. Primed targets appeared either in their original study phase form or in a mirror image version. Successful identification of original and mirror image targets and unprimed targets were compared in a repeated measures ANOVA. Table 1 presents the condition means and the ANOVA summary table for these data.

For the Breuer et al. (2009) results, the NHST method clearly leads to rejection of the null hypothesis. The Bayesian analysis of these same data, based on the BIC estimate of posterior probabilities, could be conducted as follows if the goal were to evaluate the standard null hypothesis against a model that claimed that priming should be present for one or both types of primed items. First, note that there were 40 subjects in this study (as indicated by 39 degrees of freedom for the *subjects* source of variability), each tested in three conditions. In applying Eq. 10, then, $n = 40(3 - 1) = 80$ for hypotheses involving all three conditions. On the alternative hypothesis, there are true differences among the three condition means, so the variability among those means is treated as explained, rather than as error variability. Consequently, SSE_1, the sum of squares error for the alternative model, is simply equal to the sum of squares for the error term in the standard ANOVA. In this case, $SSE_1 = 1.078$, as is shown in Table 1. For the null hypothesis, variability among condition means is unexplained, so the full amount of unexplained variability on this model is the additive combination of sum of squares for the *item* source and the *item* × *subjects* source, yielding $SSE_0 = 0.357 + 1.078 = 1.435$. Finally, the alternative hypothesis model has two more free parameters than does the null hypothesis model (e.g., the difference between the original and mirror image condition could be different from the difference between the mirror image condition and the unprimed condition), so $k_1 - k_0 = 2$. This difference in number of free parameters corresponds to the number of degrees of freedom for the effect of study condition. Substituting these numerical values into Eq. 10 produces

$$\Delta\text{BIC} = n \ln\left(\frac{SSE_1}{SSE_0}\right) + (k_1 - k_0)\ln(n)$$
$$= (80)\ln\left(\frac{1.078}{1.435}\right) + (2)\ln(80) = -14.121.$$

The ΔBIC value of –14.121 can now be used in Eq. 5 to generate an estimate of the Bayes factor as follows:

$$\text{BF} \approx \frac{p_{\text{BIC}}(\text{D}|\text{H}_0)}{p_{\text{BIC}}(\text{D}|\text{H}_1)} = \text{e}^{(\Delta\text{BIC})/2} = \text{e}^{-14.121/2} = 0.000859.$$

The final step is to convert the Bayes factor into the posterior probabilities for the two competing hypotheses, using Eqs. 6 and 7:

$$p_{\text{BIC}}(\text{H}_0|\text{D}) = \frac{\text{BF}}{\text{BF} + 1} = \frac{0.000859}{1 + 0.000859}$$
$$= .00086 \text{ and } p_{\text{BIC}}(\text{H}_1|\text{D}) = 1 - p_{\text{BIC}}(\text{H}_0|\text{D})$$
$$= .9991\,.$$

These posterior probability values indicate that the data very clearly favor the alternative hypothesis over the null hypothesis. I have implemented in Excel a routine for computing SSE_0, ΔBIC, the Bayes factor, and the posterior probabilities for the null and alternative hypotheses from input consisting of n (number of independent observations), $k_1 - k_0$ (the difference between the two models with respect to number of free parameters), sum of squares for the effect of interest, and sum of squares for the error term associated with the effect of interest (SSE_1). The Excel worksheet is available as supplementary material for this article.

Given that the difference between posterior probabilities for the null and alternative hypotheses may vary continuously and will not always convincingly favor one hypothesis over the other, a convention for describing or labeling the strength of evidence in favor of one or the other hypothesis would be very helpful. Raftery (1995) has

Table 1 Mean proportions correct and ANOVA summary table from Breuer et al. (2009, Experiment 2A)

Study Condition	Mean	Source	*SS*	*df*	*MS*	*F*	*p*
Original	.700	Subjects	0.668	39			
Mirror image	.621	Item	0.357	2	0.178	12.90	< .001
Unprimed	.567	Item × Subjects	1.078	78	0.014		
		Total	2.103	119			

95% within-subjects confidence interval for the means is ±.037

provided a suggested categorization of degrees of evidence, as shown in Table 2, that meets this need. According to this system, the obtained posterior probability for the alternative hypothesis constitutes "positive" evidence for the conclusion that a real effect is present.

In this example experiment, there were three levels of the items factor. Additional or more specific models could be tested in this case. For example, one could conduct analyses to determine which conditions are different from which, as might be done with planned or post hoc comparisons following an ANOVA. One approach using the Bayesian method would be to analyze the data for just one pair of conditions at a time. I will present one example of that here. Consider the two conditions involving primed items, differing in the orientation of items when presented at test, relative to their appearance in the study phase (original vs. mirror image). A separate ANOVA involving just those two conditions yielded $SS_{\text{Item}} = 0.124$ and $SS_{\text{Item} \times \text{Subjects}} = 0.635$. For these data,

$$\Delta\text{BIC} = n\ln\left(\frac{SSE_1}{SSE_0}\right) + (k_1 - k_0)\ln(n) = (40)\ln\left(\frac{0.635}{0.759}\right) + (1)\ln(40) = -3.45$$

Note that, in this test, only two scores per subject are considered, so the number of independent observations is equal to the number of subjects. The resulting Bayes factor is $e^{-3.45/2} = 0.1782$, and the posterior probabilities are $p_{\text{BIC}}(\text{H}_0|\text{D}) = .151$ and $p_{\text{BIC}}(\text{H}_1|\text{D}) = .849$. This outcome provides positive evidence, using Raftery's (1995) classification, that changing an item's orientation between study and test reduces priming.

In addition to following the usual path of testing null and alternative hypotheses associated with an ANOVA design, one can also use the Bayesian method to compare two models that differ in the pattern of means that is expected. In the Breuer et al. (2009) example, suppose that one theoretical perspective (H_0) anticipated that presenting studied items in mirror image form during the test should eliminate any priming effect, whereas another model (H_1) predicts that there should be a roughly linear relationship between identification performance and study condition, moving from original, to mirror image, to unprimed. Both of these patterns can be captured as a single degree of freedom contrast. In the first case, contrast coefficients of 1, –0.5, and –0.5 could be applied to the three conditions, respectively, whereas in the second case, coefficients of 1, 0, and –1 would constitute a linear trend. Each set of contrast coefficients can be applied, in turn, to the condition means to yield a sum-of-squares value representing variability explained by the model, using this equation

$$SS_{\text{contrast}} = \frac{n_i(\sum c_i M_i)^2}{\sum c_i^2}, \tag{11}$$

where n_i is the number of subjects in a particular condition (in a fully repeated measures design, such as the present example, this would be all the subjects), c_i is the contrast coefficient for a particular condition, and M_i is the mean for that condition. The resulting sum-of-squares values for the two models are .300 for H_0, and .354 for H_1. The total amount of variability to be explained is the total sum of squares shown in Table 1, less the between-subjects sum of squares: $2.103 - 0.668 = 1.435$. Thus, the unexplained variability for the two models is $SSE_0 = 1.135$, and $SSE_1 = 1.081$, for the first and second models, respectively. The number of independent observations for single-*df* contrasts such as these is just the number of subjects, because only two scores per subject are considered when fitting the contrast model to the data. In the first model, two conditions are averaged together to get a single score, and in the second model, the 0 contrast weight effectively eliminates one condition from consideration. These two models do not differ in the number of free parameters, so we have

$$\Delta\text{BIC} = n\ln\left(\frac{SSE_1}{SSE_0}\right) + (k_1 - k_0)\ln(n) = (40)\ln\left(\frac{1.081}{1.135}\right) + (0)\ln(40) = -1.94\ .$$

The Bayes factor in this case is .377, which produces a posterior probability for the linear trend model of $p_{\text{BIC}}(\text{H}_1|\text{D}) = .726$, which qualifies only as weak evidence in favor of that model over the model that assumes priming only in the original condition.

Example 2 For the second numerical example, I present a multifactor design that again can be handled with a standard factorial ANOVA. The data are from Bub and Masson (2010, Experiment 1), in which subjects made a speeded reach-and-grasp action with either the left or the right hand, cued by one of two possible colors. The color cue was carried by a picture of a handled object, with the handle oriented either to the left or to the right. Handle orientation and cued response hand were manipulated independently, making alignment of response hand and object handle a factor in the design. A second factor was the timing of the presentation of the color. Either the onset of the color was simultaneous with the pictured object (delay = 0 ms), or the color was presented after the object had been in view in

Table 2 Descriptive terms for strength of evidence corresponding to ranges of p_{bic} values as suggested by Raftery (1995)

$p_{\text{BIC}}(\text{H}_i\|\text{D})$	Evidence
.50—.75	weak
.75–.95	positive
.95–.99	strong
> .99	very strong

achromatic form for 195 ms. There was also a between-subjects factor in this experiment—namely, a manipulation of whether the performed action was compatible or incompatible with the pictured object. For the sake of simplicity, I present an analysis of data from only the compatible condition. The mean response latency (time taken to initiate the reach-and-grasp response after color onset) in each of the four conditions and the associated ANOVA summary table are shown in Table 3.

Just as each of the three effects (two main effects and an interaction) can be evaluated with an F test, each can also be evaluated by the Bayesian method. For the main effect of alignment, the *SSE* for the alternative hypothesis is 12,103, and the *SSE* for the null hypothesis is 12,103 + 1,906 = 14,009. The resulting ΔBIC is

$$\Delta\text{BIC} = n\ln\left(\frac{SSE_1}{SSE_0}\right) + (k_1 - k_0)\ln(n)$$
$$= (48)\ln\left(\frac{12,103}{14,009}\right) + (1)\ln(48) = -3.15,$$

and the Bayes factor is $e^{-3.15/2} = 0.2070$. The corresponding posterior probabilities are $p_{\text{BIC}}(\text{H}_0|\text{D}) = .171$ and $p_{\text{BIC}}(\text{H}_1|\text{D}) = .829$. This test yields positive evidence for an alignment effect, according to Raftery's (1995) classification system. For the main effect of delay, we have

$$\Delta\text{BIC} = n\ln\left(\frac{SSE_1}{SSE_0}\right) + (k_1 - k_0)\ln(n)$$
$$= (48)\ln\left(\frac{17,964}{37,344}\right) + (1)\ln(48) = -31.26.$$

For this effect, the Bayes factor is $e^{-31.26/2} = 0.0000002$, yielding $p_{\text{BIC}}(\text{H}_0|\text{D}) < .0001$ and $p_{\text{BIC}}(\text{H}_1|\text{D}) > .9999$, which is very strong evidence for an effect of delay. Finally, the interaction effect also produced a clear outcome, with

$$\Delta\text{BIC} = n\ln\left(\frac{SSE_1}{SSE_0}\right) + (k_1 - k_0)\ln(n)$$
$$= (48)\ln\left(\frac{4,421}{5,957}\right) + (1)\ln(48) = -10.44,$$

and the Bayes factor is $e^{-10.44/2} = 0.0054$. The posterior probabilities are $p_{\text{BIC}}(\text{H}_0|\text{D}) = .005$ and $p_{\text{BIC}}(\text{H}_1|\text{D}) = .995$ (very strong evidence for an interaction effect). The significant interaction could be examined by testing simple effects, as might be done with the usual ANOVA. In this experiment, interest was in the alignment effect at each level of the delay factor. The Bayesian analysis was carried out by computing separate ANOVAs for each delay condition so that the correct *SSE* values could be obtained in each case. The resulting posterior probabilities for the 0-ms delay condition were $p_{\text{BIC}}(\text{H}_0|\text{D}) = .870$ and $p_{\text{BIC}}(\text{H}_1|\text{D}) = .130$, which is positive evidence in support of the null hypothesis (no effect). In the 195-ms delay condition, there was very strong evidence for an alignment effect, $p_{\text{BIC}}(\text{H}_1|\text{D}) = .995$.

An alternative approach to the Bayesian analysis of these data could have involved a comparison between two theoretically motivated models. Following the example presented by Dixon (2003), consider a model that assumes a main effect of alignment that is independent of delay and a competing model that assumes an interaction between alignment and delay such that an alignment effect is expected only after a delay. Once again, these models may be specified by sets of contrast coefficients, one set specifying the main effect model (H_0), and one set for the interaction model (H_1). For the main effect model, the coefficients are +1, +1, –1, –1 (with +1 designating not-aligned conditions and –1 for the aligned conditions), and for the interaction model, the coefficients are –3, +1, +1, +1 (with –3 for the delayed/aligned condition and +1 for the three remaining conditions). Variability to be explained by either model is the total sum of squares shown in Table 3, less the sum of squares associated with between-subjects variability: $291,716 - 234,406 = 57,310$. The sum-of-squares value for the main effect model using Eq. 11 is 2,028, and the value for the interaction model is 16,900. The unexplained variability associated with the models is 55,282 for the main effect model (H_0) and 40,410 for the interaction model (H_1). Again, these are single-*df* contrast models, so the number of independent observations is equal to the number of subjects. The comparison

Table 3 Mean response times (in milliseconds) and ANOVA summary table from Bub and Masson (2010, Experiment 1)

	Delay (ms)		Source	*SS*	*df*	*MS*	*F*	*p*
Alignment	0	195	Subjects	234,406	47			
Aligned	523	497	Alignment	1,906	1	1,906	7.40	.009
Not aligned	524	509	Align. × Subj.	12,103	47	258		
			Delay	19,380	1	19,380	50.71	< .001
			Delay × Subj.	17,964	47	382		
			Align. × Delay	1,536	1	1,536	16.33	< .001
			Align. × Del. × Subj.	4,421	47	94		
			Total	291,716	191			

95% within-subjects confidence interval for the means is ±4.5

between these two models, which have the same number of free parameters, yields

$$\Delta BIC = n\ln\left(\frac{SSE_1}{SSE_0}\right) + (k_1 - k_0)\ln(n)$$
$$= (48)\ln\left(\frac{40,410}{55,282}\right) + (0)\ln(48) = -15.04.$$

The Bayes factor for this model comparison is 0.0005, and the posterior probability for the interaction model is $p_{BIC}(H_1|D)$ = .999. Thus, there is very strong evidence favoring the interaction model over the main effect model.

Example 3 The outcome of the simple effects analysis of the interaction in Example 2 provides a hint regarding another very powerful advantage of the Bayesian approach. Namely, the Bayesian analysis can provide a quantitative assessment of the degree of evidence supporting the null hypothesis. The final example that I will present highlights this aspect of the Bayesian approach and also illustrates how evidence may be aggregated across multiple experiments to provide particularly strong support for a hypothesis. Kantner and Lindsay (2010) tested the conjecture that providing feedback during a recognition memory test can improve accuracy. In a series of six experiments, they gave one group of subjects valid trial-by-trial feedback during a yes/no recognition memory test, whereas the control group in each experiment received no feedback. Relevant data from each experiment are shown in Table 4. The means in the table are corrected recognition scores (hits minus false alarms) for the feedback and the control conditions in each experiment, collapsed across any other factors that were manipulated in the original experiments, such as trial block within the test phase. All of the experiments produced a null result for the test of the difference between the feedback and control group means. The Bayesian analysis shows that, considered individually, the experiments produce weak or, at best, positive evidence in support of a null effect of feedback. A more convincing case can be made for the null hypothesis by simply combining the data from the entire set of experiments. Although Wagenmakers (2007) proposed that evidence can be aggregated across experiments by multiplying the Bayes factor from each experiment to yield an overall Bayes factor, this was an incorrect claim. In a corrigendum to the original article (available at http://www.ejwagenmakers.com/2007/CorrigendumPvalues.pdf), Wagenmakers points out that after an initial experiment has been conducted, the outcome of that experiment would modify the prior distribution used in the analysis of data from the second experiment. There is no straightforward way to take into account the updating of the prior distribution, so a simpler alternative is illustrated here. Namely, the data from the experiments were aggregated, and a new analysis treated the data as though they had come from a single experiment. The bottom row of Table 4 shows the aggregated means and the corresponding analyses. The resulting posterior probability favoring the null hypothesis was $p_{BIC}(H_0|D)$ = .938. Thus, the aggregated data provide stronger evidence in support of a null effect than do any of the experiments taken alone. Unlike the NHST stricture against accepting the null hypothesis, the Bayesian approach offers researchers a mechanism for quantifying evidence in support of the null hypothesis. More extensive coverage of the use of Bayesian analyses to evaluate the degree to which observed data favor the null hypothesis has been given in recent articles by Gallistel (2009) and Rouder, Speckman, Sun, Morey, and Iverson (2009).

In the context of NHST, one can compute power estimates to help interpret data when the null hypothesis is not rejected. To compare this approach to the Bayesian method for the aggregated data shown in Table 4, I computed a power analysis using G*Power 3 (Faul, Erdfelder, Lang, & Buchner, 2007), assuming three different effect sizes, equal to Cohen's (1988) benchmark values. Although the aggregated data had substantial power to detect medium (d = .5, power = .97) or large (d = .8; power > .99) effect sizes, power to detect a small effect size (d = .2) was very low at .33. This collapse of power as assumed effect size shrinks reflects the stricture within the NHST framework against accepting the null hypothesis. When using NHST, the best one can do is to claim that there is good evidence that the true effect size lies below some upper bound.

Table 4 Mean corrected recognition scores (95% confidence interval), ANOVA results, and Bayesian analysis for Kantner and Lindsay's (2010) experiments

Experiment	Feedback	Control	n	SS_{effect}	SS_{error}	F	BF	$p_{BIC}(H_0\|D)$
1	.510 (±.05)	.524 (±.05)	46	0.002	0.570	< 1	6.26	.862
2: 75% old	.594 (±.07)	.643 (±.07)	36	0.022	0.645	1.15	3.28	.766
2: 25% old	.600 (±.08)	.579 (±.07)	35	0.004	0.750	< 1	5.39	.844
3	.667 (±.07)	.688 (±.06)	43	0.005	0.922	< 1	5.84	.854
4: CRR	.480 (±.07)	.410 (±.06)	45	0.054	0.929	2.52	1.88	.653
4: SR	.332 (±.09)	.351 (±.05)	29	0.002	0.452	< 1	5.05	.835
All exps.	.537 (±.03)	.538 (±.03)	234	0.00013	7.060	< 1	15.27	.938

SS_{effect} = sum of squares for the effect of feedback; BF = Bayes factor.

Implications of using the Bayesian method

For any researcher new to the Bayesian method proposed by Wagenmakers (2007), questions regarding the relationship between the Bayesian posterior probabilities and the more familiar NHST *p* values are bound to arise. I address this general issue in three ways. First, I present a plot showing how Bayesian posterior probabilities vary as a function of sample size and effect size (i.e., how well the data are described by the alternative vs. the null hypothesis). Second, a plot is provided that compares the posterior probabilities for the null hypothesis with the *p* values generated by NHST for the sample size and effect size combinations shown in the first plot. Finally, I present a plot (which is a variant of one used by Wagenmakers, 2007) to show how Bayesian posterior probabilities diverge from the NHST *p* value when effect size varies but sample size is adjusted to keep the NHST *p* value constant. All of these results are based on a repeated measures design with one factor having two levels, but the relationships revealed by these plots should generally hold for other designs as well.

As indicated by Eqs. 5 and 10, the Bayes factor is crucially dependent on the relative goodness of fit of the null and alternative models to the data (effect size) and on sample size. The relative fits of the two competing models is expressed in Eq. 10 as the ratio between the error variability (*SSE*) under the alternative model to the error variability associated with the null model. As was noted above, this ratio is simply the complement of a standard measure of effect size, η_p^2, so larger values of this ratio indicate smaller effect sizes and, therefore, favor the null hypothesis. Smaller values of the ratio indicate larger effect sizes and so favor the alternative hypothesis. Sample size modulates the relationship between this ratio and the degree of support afforded one or the other hypothesis. This modulation is apparent in Fig. 1, which shows the posterior probability favoring the alternative hypothesis as a function of sample size and the ratio between the two relevant *SSE* values. At one extreme, there is little or no error variance when the alternative hypothesis model is fit to the data, indicating that a large effect is present and generating an *SSE* ratio very near zero. In such cases, the posterior probability for the alternative hypothesis is at or near its upper asymptotic value of 1.0, regardless of sample size.

As the goodness fit of the alternative model (effect size) decreases, the posterior probability for the alternative hypothesis decreases. Then, as the null model begins to be favored over the alternative model, the posterior probability for the alternative hypothesis approaches its lower asymptotic value. Note that the maximum value for the ratio of the two *SSE* terms is 1.0 (denoting an effect size of zero), so Fig. 1 indicates that the lower asymptotic value

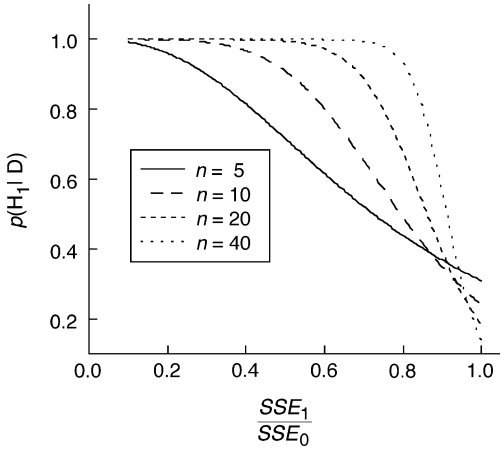

Fig. 1 The posterior probability favoring the alternative hypothesis as a function of sample size and model fit error for the alternative relative to the null hypothesis. Note that the ratio SSE_1/SSE_0 is equivalent to $1 - \eta_p^2$ (the complement of effect size). This plot is based on a repeated measures design with one factor having two levels.

for $p_{BIC}(H_1|D)$ varies as a function of sample size. In particular, with smaller samples, the lower asymptote for $p_{BIC}(H_1|D)$ is greater than when sample size is larger. The implication of this fact is that sample size limits the strength of evidence in favor of the null hypothesis (evidence is stronger with larger sample sizes). Figure 1 also shows that with larger sample sizes, $p_{BIC}(H_1|D)$ retains a relatively large value as the *SSE* ratio increases, until a critical value is reached (about .75 for $n = 40$), at which point $p_{BIC}(H_1|D)$ drops sharply and the Bayesian analysis begins to favor the null hypothesis.

Figure 2 presents a direct comparison of the estimated Bayesian posterior probability for the null hypothesis, $p_{BIC}(H_0|D)$, and the *p* value produced by NHST. The plots in Fig. 2 represent the same experimental design, sample sizes, and effect size range as in Fig. 1. In general, the Bayesian method yields posterior probabilities for the null hypothesis that begin to rise sooner than does the NHST *p* value as effect size dwindles (recall that effect size is the complement of the *SSE* ratio). To appreciate the implication of this difference, consider the relatively typical case of $n = 20$ for a repeated measures design. At the point where the NHST *p* value is approximately .05, the effect size measured by η_p^2 is slightly less than .20 (i.e., SSE_1/SSE_0 is just over .80). For this same effect size, $p_{BIC}(H_0|D)$ is already greater than .32, indicating only weak evidence in favor of an effect [$p_{BIC}(H_1|D) < .68$]. This comparison suggests that from the Bayesian perspective, researchers should be very cautious when interpreting evidence for an effect based on an NHST *p* value that hovers near .05 (see also Hubbard & Lindsay, 2008).

A further cautionary note is provided in Fig. 3 regarding NHST results that lie near the standard significance value of .05 or that are only "marginally significant." This figure

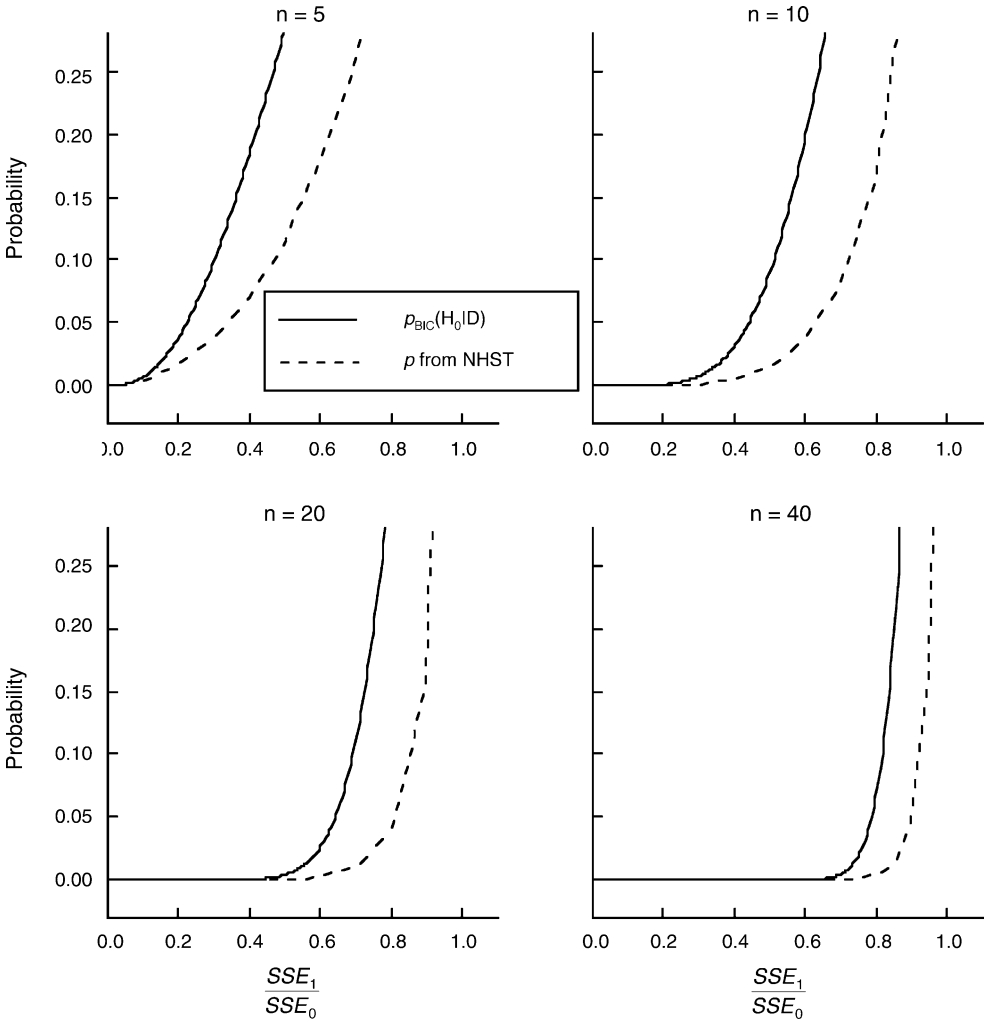

Fig. 2 Probability values associated with the null hypothesis from the Bayesian approach and from NHST as a function of sample size and model fit error for the alternative relative to the null hypothesis. Note that the ratio SSE_1/SSE_0 is equivalent to $1 - \eta_p^2$ (the complement of effect size). This plot is based on a repeated measures design with one factor having two levels.

displays the Bayesian posterior probability associated with the null hypothesis as a function of sample size when effect size is varied to maintain a constant value of .05 for the NHST p. Again, the design on which these data are based is a repeated measures design with a single factor having two levels. According to the standard hypothesis-testing approach, the evidence supporting rejection of the null hypothesis is equivalent across the range of sample sizes shown (p = .05 in all cases). In the Bayesian analysis, however, a very different outcome is apparent. As sample size increases, the posterior probability favoring the null hypothesis grows; indeed, with infinitely large sample size, it reaches an asymptote of 1.0 (Wagenmakers, 2007). This illustration is a troubling example of Lindley's (1957) paradox, which states that for any NHST p value, a sample size can be found such that the Bayesian posterior probability of the null hypothesis is 1 – p. Thus, with the appropriate sample size, a null hypothesis that is rejected in the NHST system can have strong support under a Bayesian analysis. This paradox and its illustration in Fig. 3 powerfully demonstrate the truism that p(H|D) and p(D|H) cannot be used interchangeably in statistical reasoning.

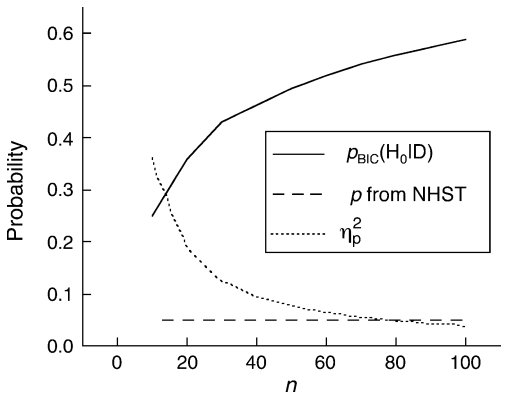

Fig. 3 Bayesian posterior probability for the null hypothesis as a function of sample size when effect size (η_p^2) is varied to maintain a constant NHST p value of .05. This plot is based on a repeated measures design with one factor having two levels

Why do the Bayesian and NHST approaches to evaluating the null hypothesis diverge to the striking degree shown in Fig. 3? The crucial element here is the fact that to

maintain a constant p value under NHST, the effect size (represented by η_p^2) must shrink as sample size increases. The difficulty this situation poses for NHST is that that method is based on a decision-making process that considers only the plausibility of the null hypothesis. But the drop in effect size also has implications for how strongly the alternative model is supported by the data. As the effect size shrinks, so too does the value added by the extra parameter (nonzero effect size) carried by the alternative hypothesis. In the Bayesian analysis, the reduced contribution of that additional parameter in accounting for variability in the data shows up as a liability when the penalty for this parameter is taken into account (the last term in Eq. 10). The critical advantage of the Bayesian approach is that it consists of a comparative evaluation of two models or hypotheses, rather than driving toward a binary decision about a single (null) hypothesis, as in NHST.

Conclusion

The BIC approximation of Bayesian posterior probabilities introduced by Wagenmakers (2007) offers significant advantages over the NHST method. Foremost among these is the fact that the Bayesian approach provides exactly the information that researchers often seek—namely, the probability that a hypothesis or model should be preferred, given the obtained data. In addition, the Bayesian approach generates graded evidence regarding both the alternative and the null hypotheses, including the degree of support favoring the null (in contrast with NHST, which proscribes acceptance of the null hypothesis), and provides a simple means of aggregating evidence across replications. Researchers who wish to consider using this Bayesian method may be reluctant to break new ground. Why give any additional cause for reviewers or editors to react negatively to a manuscript during peer review? Any such reticence is understandable. Nevertheless, my goal is to encourage researchers to overcome this concern. Studies are beginning to appear that include Bayesian analyses of data, sometimes presented alongside the results of NHST p values to give readers some sense of how the Bayesian analysis compares with NHST results (e.g., Winkel, Wijnen, Ridderinkof, Groen, Derrfuss, Danielmeier & Forstmann, 2009). Moreover, the computation of estimated posterior probabilities and the Bayes factor is relatively straightforward, as has been shown here, and BIC values associated with specific models or hypotheses can be computed in the open-source statistical package R (R development core team, 2010).

In my roles as action editor, reviewer, and author, I have found that participants in the peer review process generally are surprisingly receptive to alternative methods of data analysis. Various studies have appeared over the years that report analyses of data using techniques in place of NHST, such as likelihood ratios (e.g., Glover & Dixon, 2001; Glover, Dixon, Castiello, & Rushworth, 2005) and confidence intervals (e.g., Bernstein, Loftus, & Meltzoff, 2005; Loftus & Harley, 2005; Loftus & Irwin, 1998). The BIC approximation to Bayesian posterior probabilities is an excellent and highly practical candidate for inclusion among these alternatives.

Author Note Preparation of this article was supported by a discovery grant to the author from the Natural Sciences and Engineering Research Council of Canada. I am grateful to Jo-Anne Lefevre for suggesting that an article of this nature would be useful, to Peter Dixon and Geoffrey Loftus for helpful comments on an earlier version of the manuscript, and especially to E.-J. Wagenmakers for valuable suggestions and guidance. I also thank Justin Kantner and Stephen Lindsay for making their raw data available. Correspondence regarding this article should be sent to Michael Masson, Department of Psychology, University of Victoria, P.O. Box 3050 STN CSC, Victoria, British Columbia V8W 3P5, Canada (e-mail: mmasson@uvic.ca).

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716–723. doi:10.1109/TAC.1974.1100705

Berger, J. O., & Delampady, M. (1987). Testing precise hypotheses. *Statistical Science, 2*, 317–352. doi:10.1109/TAC.1974.1100705

Bernstein, D. M., Loftus, G. R., & Meltzoff, A. N. (2005). Object identification in preschool children and adults. *Developmental Science, 8*, 151–161. doi:10.1111/j.1467-7687.2005.00402.x

Bortolussi, M., & Dixon, P. (2003). *Psychonarratology: Foundations for the empirical study of literary response*. Cambridge: Cambridge University Press.

Breuer, A. T., Masson, M. E. J., Cohen, A.-L., & Lindsay, D. S. (2009). Long-term repetition priming of briefly identified objects. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 35*, 487–498. doi:10.1037/a0014734

Bub, D. N., & Masson, M. E. J. (2010). Grasping beer mugs: On the dynamics of alignment effects induced by handled objects. *Journal of Experimental Psychology: Human Perception and Performance, 36*, 341–358. doi:10.1037/a0017606

Chow, S. L. (1998). Précis of statistical significance: Rationale, validity, and utility. *The Behavioral and Brain Sciences, 21*, 169–239. doi:10.1017/S0140525X98001162

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale: Erlbaum.

Cohen, J. (1994). The earth is round ($p < .05$). *The American Psychologist, 49*, 997–1003. doi:10.1037/0003-066X.49.12.997

Dixon, P. (2003). The p-value fallacy and how to avoid it. *Canadian Journal of Experimental Psychology, 57*(189-202), 133–149. doi:10.1037/h0087425

Dixon, P., & O'Reilly, T. (1999). Scientific versus statistical inference. *Canadian Journal of Experimental Psychology, 53*, 189–202. doi:10.1037/h0087305

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175–191.

Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review, 116*, 439–453. doi:10.1037/a0015251

Glover, S., & Dixon, P. (2001). Dynamic illusion effects in a reaching task: Evidence for separate visual representations in the planning and control of reaching. *Journal of Experimental Psychology: Human Perception and Performance, 27*, 560–572. doi:10.1037/0096-1523.27.3.560

Glover, S., & Dixon, P. (2004). Likelihood ratios: A simple and flexible statistic for empirical psychologists. *Psychonomic Bulletin & Review, 11*, 791–806.

Glover, S., Dixon, P., Castiello, U., & Rushworth, M. F. S. (2005). Effects of an orientation illusion on motor performance and motor imagery. *Experimental Brain Research, 166*, 17–22.

Hagen, R. L. (1997). In praise of the null hypothesis statistical test. *The American Psychologist, 52*, 15–24. doi:10.1037/0003-066X.52.1.15

Hubbard, R., & Lindsay, R. M. (2008). Why *p* values are not a useful measure of evidence in statistical significance testing. *Theory & Psychology, 18*, 69–88. doi:10.1177/0959354307086923

Iverson, G. J., Wagenmakers, E.-J., & Lee, M. D. (2010). A model-averaging approach to replication: The case of p_{rep}. *Psychological Methods, 15*, 172–181.

Kantner, J., & Lindsay, D. S. (2010). Can corrective feedback improve recognition memory? *Memory & Cognition, 38*, 389–406. doi:10.3758/MC.38.4.389

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*, 773–795.

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.

Krueger, J. (2001). Null hypothesis significance testing: On the survival of a flawed method. *The American Psychologist, 56*, 16–26. doi:10.1037/0003-066X.56.1.16

Lindley, D. V. (1957). A statistical paradox. *Biometrika, 44*, 187–192. doi:10.1093/biomet/44.1-2.187

Loftus, G. R., & Harley, E. M. (2005). Why is it easier to identify someone close than far away? *Psychonomic Bulletin & Review, 12*, 43–65.

Loftus, G. R., & Irwin, D. E. (1998). On the relations among different measures of visible and informational persistence. *Cognitive Psychology, 35*, 135–199. doi:10.1006/cogp.1998.0678

Nickerson, R. S. (2000). Null hypothesis statistical testing: A review of an old and continuing controversy. *Psychological Methods, 5*, 241–301. doi:10.1037/1082-989X.5.2.241

R development core team. (2010). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

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

Raftery, A. E. (1999). Bayes factors and BIC: Comment on ”A critique of the Bayesian information criterion for model selection.”. *Sociological Methods & Research, 27*, 411–427. doi:10.1177/0049124199027003005

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. doi:10.3758/PBR.16.2.225

Trafimow, D., & Rice, S. (2009). A test of the null hypothesis significance testing procedure correlation argument. *The Journal of General Psychology, 136*, 261–269. doi:10.3200/GENP.136.3.261-270

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

Wilkinson, L., & the Task Force on Statistical Inference, (1999). Statistical methods in psychology journals: Guidelines and explanations. *The American Psychologist, 54*, 594–604. doi:10.1037/0003-066X.54.8.594

Winkel, J., Wijnen, J. G., Ridderinkof, K. R., Groen, I. I. A., Derrfuss, J., Danielmeier, C., et al. (2009). Your conflict matters to me! Behavioral and neural manifestations of control adjustment after self-experienced and observed decision-conflict. *Frontiers in Human Neuroscience, 3*, 57. doi:10.3389/neuro.09.057.2009

Errata for:
Masson, M. E. J. (2011). A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. *Behavior Research Methods*, *43*, 679-690.

On p. 684, the first paragraph indicates that the posterior probability favoring the alternative hypothesis constitutes "positive" evidence, based on Raftery's (1995) descriptive terms. In fact, the posterior probability qualifies as "very strong" evidence.

Near the bottom of p. 685, and on to p. 686, a specification for the value of n is given for the comparison between two models that are fit to all four conditions of the experiment. Even though the models can be characterized as single-*df* contrasts, the fact that they are being fit to all four within-subjects conditions means that the correct number of independent observations (n) is the number of subjects (48) multiplied by the number of conditions minus one (4 – 1), yielding 48(3) = 144. Thus, the correct formulation for the difference between BIC values for the main effect model and the interaction model is

$$\Delta\text{BIC} = n \ln\left(\frac{SSE_1}{SSE_0}\right) + (k_1 - k_0)\ln(n)$$

$$= (144)\ln\left(\frac{40,410}{55,282}\right) + (0)\ln(144) = -45.13 \ .$$

The resulting Bayes factor is 1.589E-10, and the posterior probability favoring the interaction model is $p_{\text{BIC}}(\text{H}_1|\text{D}) > .999$.