SPECIAL ISSUE

Journal of Problem Solving

What Are the Odds? A Practical Guide to Computing and Reporting Bayes Factors

Andrew F. Jarosz[1] and Jennifer Wiley[1]

[1] University of Illinois at Chicago

Correspondence:
Correspondence concerning this article should be addressed to Andrew F. Jarosz, 1007 W. Harrison St. MC 285, University of Illinois at Chicago, Chicago, IL, 60647, or via email to ajaros5@uic.edu.

Keywords:
Statistics, Bayes Factor

The purpose of this paper is to provide an easy template for the inclusion of the Bayes factor in reporting experimental results, particularly as a recommendation for articles in the Journal of Problem Solving. The Bayes factor provides information with a similar purpose to the *p*-value—to allow the researcher to make statistical inferences from data provided by experiments. While the *p*-value is widely used, the Bayes factor provides several advantages, particularly in that it allows the researcher to make a statement about the alternative hypothesis, rather than just the null hypothesis. In addition, it provides a clearer estimate of the amount of evidence present in the data. Building on previous work by authors such as Wagenmakers (2007), Rouder et al. (2009), and Masson (2011), this article provides a short introduction to Bayes factors, before providing a practical guide to their computation using examples from published work on problem solving.

The world of psychology and many related disciplines is dominated by the *p*-value. The publication of studies hinges upon a magical number—either .01 or .05—that plays a deciding role in whether the data are thought to reflect an actual difference, or random happenstance. The *p*-value is, in a word, pervasive. However, just because a measure is ubiquitous does not necessarily mean that it is the best measure. There have long been arguments for alternative statistical approaches (e.g., Edwards, Lindman, & Savage, 1963), and recently there has been a growing movement towards alternative analyses that overcome some of the shortcomings of null-hypothesis significance testing (NHST) and the associated *p*-values (Dienes, 2011; Gallistel, 2009; Johnson, 2013; Nuzzo, 2014). In particular, Bayesian methods, and Bayes factors, have been suggested as an excellent alternative (see Wagenmakers, 2007 for a thorough discussion of this alternative). In a similar vein to Masson (2011), the main goal of this paper is to provide a practical, procedural outline of how to estimate Bayes factors for both *t*-tests and regression analyses with output from common statistical programs, such as SPSS, using previously published findings from research on problem solving as examples. The focus will be on estimating Bayes factors using the BIC method, which allows for computations that are mathematically straightforward, and do not require specialized statistical programs or knowledge. That said, there are many more sophisticated approaches that should be considered by researchers interested in using these measures (Liang, Paulo, Molina, Clyde, & Berger, 2008; Rouder & Morey, 2012; Rouder, Morey, Speckman, & Province, 2012).

ISSUES WITH NULL-HYPOTHESIS SIGNIFICANCE TESTS

Given the importance of *p*-values to psychological studies, it is somewhat strange how little thought is given to their peculiarities. Everyone is taught the constraints of NHST's in their first statistics class, and yet little is done to account for them in practice. For example, it is common knowledge that running more participants increases the likelihood of finding a significant result, but few remember that it violates an assumption of NHST to "just run an extra five participants" as they chase significance (Wagenmakers, 2007). Likewise, we are all taught that .05 is the cutoff most often used to demonstrate significant results, but often forget that this is an arbitrary number, and that it may not always be the appropriate cutoff for a given study (nor may it be appropriate at all; see Johnson, 2013). Wagenmakers (2007) provides an excellent summary of some of the overarching issues of commonly used NHST, also referred to as the "frequentist approach." His points are summarized below:

1. *The p-value depends on hypothetical data.* That is to say, the sampling distribution represents many iterations of the same experiment, assuming the null hypothesis is true. These data are never actually observed, and this can lead to some logical quandaries.
2. *The p-value depends on the intentions of the researcher.* NHST is based on the idea that the experimenter has created a sampling plan with fixed stopping conditions regarding data collection, something that in practice may not always be the case. It has been demonstrated that, should an experimenter employ optional stopping,

http://dx.doi.org/10.7771/1932-6246.1167

they are virtually guaranteed to find a significant result eventually (Jennison & Turnbull, 1990).

3. *The p-value does not grant statistical evidence*. A *p*-value of .05 in a study with 20 people does not imply the same amount of evidence against the null hypothesis as the same *p*-value in a study with 200 people (or, to further emphasize this point, 2000 people). Large sample sizes will find tiny differences between groups to result in a "significant difference," whether or not such a difference has any "practical" significance. In terms of probability, it is incorrect to assume that two *p*-values from studies of different sample sizes carry the same statistical weight of evidence.

Perhaps the most salient point made by Wagenmakers (2007) is a consideration of what is being compared when using a frequentist approach. That is to say, *nothing* is being compared.

4. *A frequentist approach to NHST considers only the extremeness of the data under the null hypothesis, with no consideration of the alternative hypothesis.* In reality, it may be that neither the null nor the alternative hypothesis are good fits for the current data, suggesting that a comparative approach may result in a clearer picture.

Given the above, it is clear that the *p*-value does not deliver all that we expect it to. Indeed, many of the assumptions made about *p*-values are incorrect (Wagenmakers, 2007).

THE ADVANTAGES OF THE BAYESIAN APPROACH

As opposed to a frequentist approach, a Bayesian approach to hypothesis testing is comparative in nature. That is, the likelihood of the data is considered under both the null and alternative hypotheses, and these probabilities are compared via the Bayes factor. The Bayes factor is a ratio that contrasts the likelihood of the data fitting under the null hypothesis with the likelihood of fitting under the alternative hypothesis. One simplified way to express this is:

$$BF_{01} = \frac{\text{likelihood of data given } H_0}{\text{likelihood of data given } H_1}$$

Therefore, as BF_{01} increases, there is more evidence in support of the null hypothesis, and less in favor of the alternative hypothesis. Taking the inverse yields the opposite; if $1 / BF_{01} = 5$, that suggests that the data are five times more likely to occur under the alternative hypothesis compared to the null hypothesis.

This method addresses each of the previously listed concerns, as discussed by Wagenmakers (2007):

1. *The p-value depends on hypothetical data*. The Bayesian approach considers only the observed data, and how those data relate to the null and alternative hypotheses.
2. *The p-value depends on the intentions of the researcher*. Bayesian approaches are not altered by stopping or measurement criteria (Rouder, 2014). Indeed, in a full Bayesian analysis, analyzing one batch of data can be used to inform the analysis of the next batch.
3. *The p-value does not grant statistical evidence*. Because Bayes factors are ratios of probabilities, two Bayes factors of equal value represent the same amount of evidence in favor of the alternative hypothesis, regardless of sample size or other extraneous factors.
4. *A frequentist approach to NHST considers only the extremeness of the data under the null hypothesis, with no consideration of the alternative hypothesis*. All Bayesian approaches are comparisons of models. This means that a Bayes factor considers the likelihood of both the null and the alternative hypothesis. From the researcher's standpoint, this is likely closer to their overall goal than simply rejecting the null hypothesis.

One issue that is often raised about Bayesian analyses is that they require a "prior"; that is, a prior probability distribution for the model parameters. Coming up with such an estimate can be problematic as it can require a certain amount of subjectivity and/or prior knowledge about the effect that is to be studied. However, a number of fairly objective priors have been developed, that make relatively few assumptions about the parameters. By using Bayesian information criteria (BIC) to estimate Bayes factors using the following equation, a "unit information prior" is assumed (Masson, 2011; Wagenmakers, 2007). This is a prior probability distribution assuming probable values for the effect size are represented by a normal distribution, centered on the effect size observed in the data, and providing a standardized, conservative prior probability for the effect size in the analysis. The equation for estimating Bayes factors using BIC is as follows:

$$BF_{01} = e^{\Delta \text{BIC}_{10}/2}$$

This equation estimates Bayes factors using the difference between the two BICs for the null and alternative hypotheses. Furthermore, these BICs can be calculated using output received during a normal frequentist analysis, making analyses possible even without extensive statistical knowledge. In the case of regression (and correlation), the BICs can be computed using just the R^2 value, the sample size, and the number of predictor variables. In the case of ANOVA (and *t*-tests), the BICs can be computed using the sum of squares for each experimental effect (which may include interactions), the error term, and the total, as well as the sample size and the number of independent variables (and interactions).

COMPUTING THE BAYES FACTOR FOR A REGRESSION ANALYSIS

A study on the role of distraction in performance on a problem solving task can be used to illustrate the computation of BICs from regression analyses. In this study, Jarosz and

Table 1.
Estimating Bayes Factor from Regression (Jarosz & Wiley, 2012)

Calculation	Parameters	Data
Estimating BIC for H_0	$BIC = n * \ln(1 - R^2) + k * \ln(n)$	$BIC = 64 * \ln(1 - .04) + 2 * \ln(64) = 5.71$
Estimating BIC for H_1	$BIC = n * \ln(1 - R^2) + k * \ln(n)$	$BIC = 64 * \ln(1 - .21) + 3 * \ln(64) = -2.61$
Change in BIC	$\Delta BIC_{10} = BIC_{H1} - BIC_{H0}$	$\Delta BIC_{10} = -2.61 - 5.71 = -8.32$
Bayes Factor	$BF_{01} = e^{\Delta BIC_{10} / 2}$	$BF_{01} = e^{-.8.32 / 2} = .016$

Wiley (2012) created two versions of the Raven Advanced Progressive Matrices (RAPM). In one condition, the most commonly selected incorrect answer (the "salient distracter") was excluded from the response bank of each item, while in the other condition a different, less salient response was removed and the salient distracter remained. The condition in which the high salience distracters appeared was referred to as the high salience item condition, while the condition where only low salience distracters appeared was referred to as the low salience item condition. The results of the hierarchical regression predicting working memory capacity were reported as described below:

> A hierarchical regression predicting composite span score was performed with low salience item accuracy as a predictor in the first step, and high salience item accuracy as a predictor in the second step. While the initial model was marginally significant, $F(1, 62) = 2.86$, $p = .10$, the addition of high salience item performance resulted in a significant model, with a significant change in the R^2 value, $R = .46$, $\Delta R^2 = .17$, $\Delta F(1, 61) = 12.79$, $p = .001$. In the final model, low salience item performance did not predict composite span score ($\beta = -.06$, $t(61) = -.46$, *ns*), while high salience item performance did ($\beta = .49$, $t(61) = 3.58$, $p = .001$, $sr^2 = .41$). (Jarosz & Wiley, 2012, p. 432)

In this regression, there was a significant increase in model fit for performance on items that included the salient distracter over performance on items that excluded the salient distracter. These results were interpreted as showing a unique role for distraction in explaining the WMC–RAPM relationship:

> The results of this study strongly support the idea that salient distracters among response options contribute to the WMC–RAPM correlation. . . . When placed hierarchically into a regression, performance on the high salience items predicted variance in the composite span score above and beyond low salience item performance, and remained the only unique predictor. This follows the prediction of the attentional control account, suggesting that high WMC individuals are better able to avoid distraction from the highly salient incorrect option within the response bank. (Jarosz & Wiley, 2012, p. 432)

These data can be re-examined using Bayesian methods to compare the initial model to the final model (Table 1). The first step is to find the unexplained variance for both the model representing the null hypothesis (in this case, a the model including only performance on items with low salience distracters, or the first step of the hierarchical regression) and the alternative hypothesis (a model with performance on both low and high salience distracter items, or the second step of the regression). The unexplained variance for the alternative hypothesis can be computed as (1 – total variance explained in the second step). The R for the second step was .46, which makes the $R^2 = .21$ and the unexplained variance is $(1 - .21) = .79$. For the null hypothesis, we need to compute the unexplained variance for the first step of the regression. Since the change in variance explained by the second step was .17, this makes the variance explained by the first step .21 – .17 = .04. Thus, the total unexplained variance for the first step is $(1 - .04) = .96$. One could also find the variance explained (R^2) for each model directly from the SPSS regression output.

From this information it is possible to calculate a BIC for each model. The BIC for a regression model (Wagenmakers, 2007) is equivalent to

$$BIC = n \times \ln(1 - R^2) + k \times \ln(n)$$

where k is the number of free parameters or predictors (in this case, the number of regressors plus the intercept), and n is the total sample size (although see Masson, 2011, with regards to the value of n in within-subjects designs). The BIC for the model of the null hypothesis is represented by $BIC = 64 \times \ln(.96) + 2 \times \ln(64) = 5.71$, ($k$ has a value of 2 in the null hypothesis model, as there are two predictors, low salience item performance and the intercept). For the alternative hypothesis, the model has a BIC of $BIC = 64 \times \ln(.79) + 3 \times \ln(64) = -2.61$ (where $k = 3$ because there are three predictor variables).

The next step is to compare the difference between the two BICs by inserting them into this equation:

$$\Delta BIC_{10} = BIC_{H1} - BIC_{H0}$$

And finally, a transformation converts the change in BICs into a Bayes factor estimate:

$$BF_{01} = e^{\Delta BIC_{10}/2}$$

This leads to $\Delta BIC_{10} = -2.61 - 5.71 = -8.32$, and a Bayes factor of $BF_{01} = e^{-.8.32 / 2} = .016$.

Table 2.
Estimating Bayes Factor From Between-Subjects *t*-tests/ANOVAs (Jarosz et al., 2012; Insight Rating Findings)

Calculation	Parameters	Data
Unexplained variance H_0	$1 - R^2 = (SS_{error} + SS_{condition}) / SS_{total}$	$1 - R^2 = (48.43 + 4.05) / 587.82 = .089$
Unexplained variance H_1	$1 - R^2 = SS_{error} / SS_{total}$	$1 - R^2 = 48.43 / 587.82 = .082$
Estimating BIC for H_0	$BIC = n * \ln(1 - R^2) + k * \ln(n)$	$BIC = 40 * \ln(.089) + 1 * \ln(40) = -93.08$
Estimating BIC for H_1	$BIC = n * \ln(1 - R^2) + k * \ln(n)$	$BIC = 40 * \ln(.082) + 2 * \ln(40) = -92.66$
Change in BIC	$\Delta BIC_{10} = BIC_{H1} - BIC_{H0}$	$\Delta BIC_{10} = -92.66 - (-93.08) = .42$
Bayes Factor	$BF_{01} = e^{\Delta BIC_{10} / 2}$	$BF_{01} = e^{.42 / 2} = 1.23$

This Bayes factor suggests that the data are .016 times more likely to occur under the null hypothesis than under the alternative hypothesis. Alternatively, taking the inverse puts this value in terms of the alternative hypothesis, 1 / .016 = 62.50. This means that the data are 62.50 times more likely to occur under the alternative hypothesis than under the null hypothesis. Using this new information we can expand the previously published results to include the Bayes factor (with added text in brackets and italics):

> A hierarchical regression predicting composite span score was performed with low salience item accuracy as a predictor in the first step, and high salience item accuracy as a predictor in the second step. While the initial model was marginally significant, $F(1, 62) = 2.86$, $p = .10$, the addition of high salience item performance resulted in a significant model, with a significant change in the R^2 value, $R = .46$, $\Delta R^2 = .17$, $\Delta F(1, 61) = 12.79$, $p = .001$. In the final model, low salience item performance did not predict composite span score ($\beta = -.06$, $t(61) = -.46$, ns), while high salience item performance did ($\beta = .49$, $t(61) = 3.58$, $p = .001$, $sr^2 = .41$). *[In addition, the data were examined by estimating a Bayes factor using Bayesian Information Criteria (Wagenmakers, 2007). This compares the fit of the data under the null hypothesis, compared to the alternative hypothesis. An estimated Bayes factor (null/alternative) suggested that the data were .016:1 in favor of the alternative hypothesis, or rather, 62.50 times more likely to occur under a model including an effect for salient distracters than a model without it.]* (Jarosz & Wiley, 2012, p. 432)

For computation of the Bayes factor for correlations, one can use the same approach as outlined above, comparing the variance explained (correlation squared) for two predictors versus one predictor (the intercept).

COMPUTING THE BAYES FACTOR FOR ANOVAS AND T-TESTS

BIC analyses can also be applied to analysis of variance (ANOVA) and *t*-tests, using sum of squares to compute the unexplained variance. This section will focus primarily on the computations for between-subjects *t*-tests, though the analyses and discussion can be applied to between-subjects ANOVA as well. The unexplained variance for the model containing the alternative hypothesis involving an independent variable is represented by SS_{error} / SS_{total}, while the unexplained variance for the null hypothesis is represented by $(SS_{error} + SS_{independentvarianble}) / SS_{total}$—that is, a model where the variance explained by the independent variable is included as part of the unexplained variance.

To illustrate the estimation of a Bayes factor from an ANOVA/*t*-test, we can revisit the findings of Jarosz, Colflesh, and Wiley (2012), who explored the impact of moderate intoxication due to alcohol on creative problem solving. Using an alcohol intoxication condition and a control condition, each with 20 participants, they had participants solve a number of remote associates test (RAT) problems while rating whether they felt they had solved the problems insightfully, or analytically. Using a frequentist approach, they first reported a marginal difference in feelings of insight:

> On average, intoxicated individuals tended to rate their experience of problem solving as being more insightful ($M = 3.98$) than the sober participants ($M = 3.35$, $t(38) = 1.78$, $p < .08$). (Jarosz, et al., 2012, p. 490)

Re-analyzing these data from a Bayesian perspective requires several steps. First, sums of squares must be calculated. This can easily be accomplished by re-analyzing the data using an ANOVA in any common statistical program (although note that the total sum of squares is not always displayed, depending on the ANOVA performed—this may need to be calculated by adding together the other sums of squares). Doing so yields 4 sums of squares: the intercept, at 535.34; the alcohol condition variable, at 4.05; the error term, at 48.43; and the total sum of squares, at 587.82. The unexplained variance for the model containing the alternative hypothesis (that intoxicated individuals would differ in their problem ratings from sober individuals) is represented by SS_{error} / SS_{total}, while the null hypothesis would represent unexplained error by $(SS_{error} + SS_{alcohol}) / SS_{total}$. Thus, the unexplained variance for

the model representing the alternative hypothesis is 48.43 / 587.82 = .082, while the unexplained variance for the model representing the null hypothesis is (48.43 + 4.05) / 587.82 = .089. These values are plugged into the equations as shown in Table 2. For an ANOVA, *k* includes the intercept and the independent variable. The Bayes factor is found to be BF_{01} = 1.23, with an inverse of 1 / 1.23 = .81.

This suggests that the data actually provide more support for the null hypothesis, being 1.23 times more likely to occur under the null hypothesis, compared to the alternative hypothesis. Updating the results to include the Bayes factor (again with new text in brackets and italics) leads to:

> On average, intoxicated individuals tended to rate their experience of problem solving as being more insightful (M = 3.98) than the sober participants (M = 3.35, $t(38)$ = 1.78, $p < .08$). *[However, the data were also examined by estimating a Bayes factor using Bayesian Information Criteria (Wagenmakers, 2007), comparing the fit of the data under the null hypothesis and the alternative hypothesis. An estimated Bayes factor (null/alternative) suggested that the data were 1.23:1 in favor of the null hypothesis, or rather, 1.23 times more likely to occur under a model without including an effect of moderate alcohol intoxication, rather than a model with it.]* (Jarosz et al., 2012, p. 490)

Jarosz et al. (2012) also examined differences in problem solving accuracy between sober and intoxicated individuals. Here, they found a significant difference in creative performance using the Remote Associates Task:

> More importantly, a second set of analyses examined whether intoxication affected the actual solution of these creative problems. On average, intoxicated participants solved significantly more RAT problems (M = .58, SD = .13) than their sober counterparts (M = .42, SD = .16), $t(38)$ = 3.43, p = .001, d = 1.08. (Jarosz et al., 2012, p. 490)

Once again, re-analyzing these data from a Bayesian perspective requires several steps. The sums of squares were: .25 for the alcohol condition, .79 for the error term, and 11.11 for the total. The unexplained variance for the model containing the alternative hypothesis (that intoxication affects problem solving) would be represented by SS_{error} / SS_{total}, while the null hypothesis would represent unexplained error by (SS_{error} + $SS_{alcohol}$ / SS_{total}). For the alternative hypothesis, the unexplained variance is .79 / 11.11 = .071. For the null hypothesis, the unexplained variance is equivalent to (.25 + .79) / 11.11 = .094. As seen in Table 3, the Bayes factor is BF_{01} = .023. This suggests that the data are far less likely under the null hypothesis than the alternative hypothesis. Taking the inverse, 1 / .023 = 43.48 shows that the data are 43.48 times more likely to occur under the alternative hypothesis than under the null hypothesis. Thus, the results section could be updated with the new information (in brackets and italics) accordingly:

> More importantly, a second set of analyses examined whether intoxication affected the actual solution of these creative problems. On average, intoxicated participants solved significantly more RAT problems (M = .58, SD = .13) than their sober counterparts (M = .42, SD = .16), $t(38)$ = 3.43, p = .001, d = 1.08. *[The estimated Bayes factor (null/alternative) suggested that the data were .023:1 in favor of the alternative hypothesis, or rather, 43.48 times more likely to occur under the model including an effect for alcohol, rather than the model without it.]* (Jarosz et al., 2012, p. 490)

Some adjustments may be needed for within-subjects or repeated measures *t*-tests and ANOVAs, although there remains some debate as to the best way to calculate *n* for a repeated-measures ANOVA. While Wagenmakers (2007) suggests that treating this value as the number of subjects is fine, Masson (2011) suggests that treating this as the number of independent observations is more appropriate, Thus, Masson suggests adjusting *n* to be the number of subjects, multiplied by (number of conditions – 1). For a detailed review of how to calculate Bayes factors in ANOVA (and in particular, ANOVA in a within-subjects design), please refer to Masson (2011).

Table 3.

Estimating the Bayes Factor From Between-Subjects *t*-Tests/ANOVAs (Jarosz et al., 2012, Creative Problem Solving Performance Findings)

Calculation	Parameters	Data
Unexplained variance H_0	$1 - R^2 = (SS_{error} + SS_{condition}) / SS_{total}$	$1 - R^2 = (.25 + .79) / 11.11 = .094$
Unexplained variance H_1	$1 - R^2$ = SSerror / SStotal	$1 - R^2 = .79 / 11.11 = .071$
Estimating BIC for H_0	$BIC = n * \ln(1 - R^2) + k * \ln(n)$	$BIC = 40 * \ln(.094) + 1 * \ln(40) = -90.89$
Estimating BIC for H_1	$BIC = n * \ln(1 - R^2) + k * \ln(n)$	$BIC = 40 * \ln(.071) + 2 * \ln(40) = -98.43$
Change in BIC	$\Delta BIC_{10} = BIC_{H1} - BIC_{H0}$	$\Delta BIC_{10} = -98.43 - (-90.89) = -7.54$
Bayes Factor	$BF_{01} = e^{\Delta BIC_{10} / 2}$	$BF_{01} = e^{-7.54 / 2} = .023$

OTHER APPROACHES TO ESTIMATING BAYES FACTORS

While the BIC provides an easy way to estimate the Bayes factor based on output from more familiar NHST approaches, it is important to remember that the method outlined above is a fairly rough approximation of the Bayes factor. While it certainly gives a much better idea of the evidence for and against one's hypotheses than does the *p*-value, in recent years several mathematical psychologists and statisticians have worked on developing better methods for calculating Bayes factors (Liang et al., 2008; Rouder & Morey, 2012; Rouder et al., 2012). In particular, the JZS approach (advocated for by Rouder, Morey, and Wagenmakers, among others; Rouder, Speckman, Sun, Morey, & Iverson, 2009) deserves mention. This method employs a prior based on work by Jeffreys (1961) and Zellner and Siow (1980). Rouder and colleagues highlight several issues with the BIC method of approximation. First, the variance of the prior using the BIC method is based on the observed sample variance; second, the unit information prior is more informative than the JZS prior, making the BIC method a less conservative alternative with respect to the alternative hypothesis; and third, the BIC method may not be well suited for mixed models in ANOVA. While the BIC method approximates the JZS method for larger sample sizes, those using smaller samples or mixed methods may be better served by employing the JZS method to compute Bayes factors.

The calculations for deriving and employing the JZS method are beyond the scope of this paper. Thankfully, there is an online calculator available at http://pcl.missouri.edu/bayesfactor that can be used to estimate Bayes factors based on *t* values (in the case of *t*-tests) and R^2 values (in the case of regression), as well as sample size (Liang et al., 2008; Rouder & Morey, 2012; Rouder et al., 2009). For *t*-tests, this supplies both a BIC estimated Bayes factor, as well as the JZS Bayes factor (Rouder et al., 2009). It should be noted that the webpages for calculating Bayes factors based on *t*-tests contain an additional parameter, *r*. This factor is intended to be used to scale the prior distribution. Leaving it as 1 does not scale the distribution, while decreasing or increasing this value will scale the prior to represent smaller or larger effect sizes, respectively. This may be appropriate if one expects smaller or larger effects in a study, however, the value of *r* should be determined a priori, and it is generally recommended that this value be left as 1 (Rouder et al., 2009).

Finally, it must be noted that various researchers have begun implementing packages for full Bayesian analysis in programs such as R, capable of handling most traditional analyses (Morey & Rouder, 2011; Rouder & Morey, 2012; Rouder et al., 2012; Rouder et al., 2009). These allow the R-savvy researcher to complete Bayesian analyses without having to transform results from other traditional statistics programs. In addition, there is currently an effort to provide an open source Bayes factor alternative to popular statistical programs ("JASP", 2014). Together, these provide a wide variety of more advanced options with regards to computing Bayes factors.

INTERPRETING BAYES FACTORS

The advantage of the Bayes factor is that it is not just a measure of how unlikely the null hypothesis is, but rather, a comparison of how likely the null is compared to the alternative. That is, instead of simply saying "It is unlikely that there is no relationship between these variables," the researcher is able to say "this alternative model is considerably better than the null, and I have the probabilities to prove it!" The Bayes factor allows for the inclusion of a statement in the results of how much more likely the data are to occur if the null hypothesis is true, compared to if the alternative hypothesis is true. If the prior odds are assumed to be 1, then taking the inverse allows one to speak to the likelihood of the alternative hypothesis, compared to the null.

For example, imagine a scenario where $BF_{01} = .5$. In this case, the data are half as likely under the null hypothesis as they are under the alternative hypothesis. Taking the inverse demonstrates that the data are twice as likely under the alternative hypothesis. Thus, the easiest interpretation of a Bayes factor is simply taking it at face value, and considering those odds.

Alternatively, there are several authors (Jeffreys, 1961, Appendix B; Raftery (1995); Wetzels et al., 2011) who have each developed some guidelines for language that may be used to discuss and interpret Bayes factors. Their suggested terminology is shown in Table 4. According to these suggestions, the discussion of the results for Study 1 from Jarosz and Wiley (2012) could be updated to include the claim that the results provided strong or very strong evidence for the alternative hypothesis (with new text in brackets and italics).

> The results of this study strongly support the idea that salient distracters among response options contribute to the WMC–RAPM correlation. . . . When placed hierarchically into a regression, performance on the high salience items predicted variance in the composite span score above and beyond low salience item performance, and remained the only unique predictor. *[Further, the Bayes factor suggested strong evidence for the role of salient distracters in the RAPM–WMC relationship.]* This follows the prediction of the attentional control account, suggesting that high WMC individuals are better able to avoid distraction from the highly salient incorrect option within the response bank. (Jarosz & Wiley, 2012, p. 432)

Table 4.
Interpretation of Bayes Factors as Evidence for Alternative Hypotheses

Statistic		Support for H_1	
Bayes Factor	**Inverse of Bayes Factor**	**Raftery**	**Jeffreys**
1–.33	1–3	Weak	Anecdotal
.33–.10	3–10	Positive	Substantial
.10–.05	10–20	Positive	Strong
.05–.03	20–30	Strong	Strong
.03–.01	30–100	Strong	Very Strong
.01–.0067	100–150	Strong	Decisive
<.0067	>150	Very Strong	Decisive

The results of the insight rating analysis in Jarosz et al. (2012) could be said to provide weak or anecdotal evidence for the null hypothesis, while the analysis of problem solving performance would be described as strong or very strong in favor of the alternative hypothesis. For an example of a paper using Bayes factors alongside traditional NHST, the authors recommend a recent study by Zwaan and Pecher (2012) examining mental simulation and language comprehension.

Another approach to estimating Bayes factors is to rely on prior work that has explored the relation between *p*-values and Bayes factors. Wetzels and colleagues (2011) used data from 855 *t*-tests in popular psychology journals to compare evidence from *p*-values, Bayes factors, and effect sizes (Wetzels et al., 2011). In general, studies that reported *p*-values of .05 provided only anecdotal evidence for findings according to a Bayesian analysis. It is only as *p*-values approach .01 that evidence starts becoming substantial, according to the calculated Bayes factors (Johnson, 2013; Nuzzo, 2014). Thus, in lieu of computing Bayes Factors, when *p* values are less than 0.01 one could cite Wetzels et al. (2011) as evidence that such *p*-values are likely to represent what Bayesians would call substantial evidence for an alternative hypothesis. Alternatively, *p*-values greater than 0.01 should be interpreted as representing "anecdotal evidence" according to Wetzels et al. (2011). However, it should be noted that this method is by no means foolproof, as large sample sizes tend to skew the evidence in favor of rejecting the null hypothesis when using only *p*-values (Wagenmakers, 2007).

A final point made by many who advocate a move away from NHST is the need for researchers to engage in deeper statistical thinking. Gigerenzer (1998) considers NHST to be no better than "ritual handwashing," a habit followed by researchers, often with little understanding of why they do what they do. He suggests that current statistical protocols followed by many researchers obviate the need for deeper consideration of alternative models, themes, or positions, indeed allowing many to avoid specifying hypotheses altogether. In suggesting Bayes factors to support (or even replace) NHST, the goal is not to simply replace one mindless algorithm with another. Several papers (Edwards et al., 1963; Gallistel, 2009; Gigerenzer, 1998; Myung & Pitt, 1997) discuss how obtaining a value beyond some threshold is not the point of the analysis, and suggest that one of the inherent problems with NHST is that it allows analysis without throughtful reflection on alternative hypotheses. Rather, consideration must be given to the calculated probabilities themselves, what those probabilities say about the relative strengths of the null and alternative hypotheses, and how those probabilities inform the greater research question at hand before any conclusions can be drawn. Likewise, care must be taken to specify hypotheses in advance, and to consider that the models being compared are specific to those null and alternative hypotheses. Bayes factors are considerably more conducive to this line of thinking when compared to traditional NHST. In short, while this paper focuses on procedure, a procedural shift is only one small part in the necessary transformation in the way that researchers think about their data.

SUMMARY

The Bayes factor provides information with a similar purpose to the *p*-value—to allow the researcher to make a statistical inference about the evidence in an experiment. While the *p*-value is widely reported, the Bayes factor provides several advantages, particularly in that it allows the researcher to make a statement about the alternative hypothesis, rather than just the null hypothesis. In addition, it provides a clearer estimate of the amount of evidence present in the data. The BIC approximation, while only a rough estimate of a Bayes factor, provides a simple way to gain the benefits of Bayes factors without requiring a statistical background or additional statistical programs. Other more advanced methods of computation are becoming available, such as the JZS method or methods for engaging in full Bayesian analyses using R. While not yet widely used, it is the goal of this paper

to increase the odds that researchers include some approximation of Bayes factors when reporting the results of their experiments, particularly in the pages of the Journal of Problem Solving.

REFERENCES

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science, 6*(3), 274–290. http://dx.doi.org/10.1177/1745691611406920

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review, 70*(3), 193–242. http://dx.doi.org/10.1037/h0044139

Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review*, *116*(2), 439–453. http://dx.doi.org/10.1037/a0015251

Gigerenzer, G. (1998). We need statistical thinking, not statistical rituals. *Behavioral and Brain Sciences, 21*(2), 199–200. http://dx.doi.org/10.1017/S0140525X98281167

JASP. (2014). Retrieved from http://jasp-stats.org/

Jarosz, A. F., Colflesh, G. J. H., & Wiley, J. (2012). Uncorking the muse: Alcohol intoxication facilitates creative problem solving. *Consciousness & Cognition, 21*(1), 487–493. http://dx.doi.org/10.1016/j.concog.2012.01.002

Jarosz, A. F., & Wiley, J. (2012). Why does working memory capacity predict RAPM performance? A possible role of distraction. *Intelligence, 40*(5), 427–438. http://dx.doi.org/10.1016/j.intell.2012.06.001

Jeffreys, H. (1961). *Theory of probability* (3rd Ed.). Oxford, UK: Oxford University Press.

Jennison, C., & Turnbull, B. W. (1990). Statistical approaches to interim monitoring of medical trials: A review and commentary. *Statistical Science, 5*(3), 299–317. http://dx.doi.org/10.1214/ss/1177012099

Johnson, V. E. (2013). Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences of the United States of America, 110*(48), 19313–19317. http://dx.doi.org/10.1073/pnas.1313476110

Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2008). Mixtures of *g* priors for Bayesian variable selection. *Journal of the American Statistical Association*, *103*(481), 410–423. http://dx.doi.org/10.1198/016214507000001337

Masson, M. E. J. (2011). A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. *Behavior Research, 43*(3), 679–690. http://dx.doi.org/10.3758/s13428-010-0049-5

Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for interval null hypotheses. *Psychological Methods*, *16*(4), 406–419. http://dx.doi.org/10.1037/a0024377

Myung, J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review, 4*(1), 79–95. http://dx.doi.org/10.3758/BF03210778

Nuzzo, R. (2014). Statistical errors. *Nature, 506*, 150–152. http://dx.doi.org/10.1038/506150a

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

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Pyschonomic Bulletin and Review, 21*(2), 301–308. http://dx.doi.org/10.3758/s13423-014-0595-4

Rouder, J. N., & Morey, R. D. (2012). Default Bayes factors for model selection in regression. *Multivariate Behavioral Research, 47*(6), 877–903. http://dx.doi.org/10.1080/00273171.2012.734737

Rouder, J. N., Morey R. D., Speckman P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology, 56*(5), 356–374. http://dx.doi.org/10.1016/j.jmp.2012.08.001

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 and Review, 16*(2), 225–237. http://dx.doi.org/10.3758/PBR.16.2.225

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin and Review, 14*(5), 779–804. http://dx.doi.org/10.3758/BF03194105

Wetzels,, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E. J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 *t* tests. *Perspectives on Psychological Science, 6*(3), 291–298. http://dx.doi.org/10.1177/1745691611406923

Zellner, A., & Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. In J. M. Bernardo, M. H. Degroot, D. V. Lindeley, & A. F. M. Smith (Eds.), *Bayesian Statistics: Proceedings of the first international meeting held in Valencia (Spain)* (pp. 585–603), Valencia, Spain: University of Valencia.

Zwaan, R. A., & Pecher, D. (2012). Revisiting mental simulation in language comprehension: Six replication attempts. *PLoS ONE, 7*(12), e51382. http://dx.doi.org/10.1371/journal.pone.0051382