Behav Res (2012) 44:1255–1265
DOI 10.3758/s13428-012-0186-0

MorePower 6.0 for ANOVA with relational confidence intervals and Bayesian analysis

Jamie I. D. Campbell · Valerie A. Thompson

Published online: 22 March 2012

Abstract MorePower 6.0 is a flexible freeware statistical calculator that computes sample size, effect size, and power statistics for factorial ANOVA designs. It also calculates relational confidence intervals for ANOVA effects based on formulas from Jarmasz and Hollands (Canadian Journal of Experimental Psychology 63:124–138, 2009), as well as Bayesian posterior probabilities for the null and alternative hypotheses based on formulas in Masson (Behavior Research Methods 43:679–690, 2011). The program is unique in affording direct comparison of these three approaches to the interpretation of ANOVA tests. Its high numerical precision and ability to work with complex ANOVA designs could facilitate researchers' attention to issues of statistical power, Bayesian analysis, and the use of confidence intervals for data interpretation. MorePower 6.0 is available at https://wiki.usask.ca/pages/viewpageattachments.action?pageId=420413544.

Keywords Power for ANOVA · Relational confidence intervals · Bayesian analysis

Despite criticism on a variety of grounds (e.g., Dienes, 2011; Dixon, 2003; Dixon & O'Reilly, 1999; Masson, 2011; Rozeboom, 1960; Wagenmakers, 2007), null-hypothesis significance testing (NHST) remains the dominant method of data analysis in the psychological sciences. Presumably, this is because it continues to be widely offered as standard training for psychology students and because of the perception that alternative approaches are not readily accessible. In this article, we describe a statistical calculator, MorePower 6.0, that calculates power-related statistics (sample size, effect size, and power) and relational confidence intervals (CIs) for ANOVA effects, and that performs Bayesian analysis of the null hypothesis (H_0) versus the alternative hypothesis (H_1). Thus, the calculator provides three alternative approaches to interpretation of ANOVA effects. Power analysis quantifies the sensitivity of a statistical test to detect an effect of a specific size (Faul, Erdfelder, Lang, & Buchner, 2007). The use of relational CIs can reduce reliance on NHST by affording interpretation of a pattern of means without requiring an inference about the statistical significance of the difference between a given pair of the means (Jarmasz & Hollands, 2009; Masson & Loftus, 2003). Bayesian analysis affords direct comparison of the probabilistic evidence provided by the data for the null versus the alternative hypothesis, whereas NHST provides only a binary decision whether or not to reject H_0 (Berger, 1985; Masson, 2011). MorePower permits researchers who normally rely on NHST methods to assess how a Bayesian approach might alter their conclusions about data.

Many online power and sample-size applets are available, as well as standalone programs for power analysis (e.g., G*Power 3; Faul et al., 2007), but MorePower provides numerous unique features. Complex designs and effects—including repeated measures (RM; i.e., within-subjects) factors, independent measures (IM; i.e., between-subjects) factors, or combinations of both types of factors—can be specified easily using drop-down menus. An ANOVA effect size may be specified in terms of the effect-related variance explained (partial eta-squared [η_p^2]) or in terms of a test statistic (F, mean square treatment [MST], or t). For tests with one degree of freedom (including interactions), effect size may be specified or calculated as the *difference* in the original units of measurement, with variability specified in terms of mean square error (MSE), standard deviation (S), variance (S^2), or standard error (SE). Along with its Bayesian

J. I. D. Campbell (✉) · V. A. Thompson
Department of Psychology, University of Saskatchewan,
9 Campus Drive,
Saskatoon, Saskatchewan S7N 5A5, Canada
e-mail: jamie.campbell@usask.ca

and CI functions for ANOVA designs, these features make MorePower 6.0 a unique and powerful analytical tool.

Power analysis for ANOVA

In NHST, a researcher assumes H_0—usually that there is no difference between the population parameters—to be true, and rejects H_0 if the observed result has a probability at least as extreme as the critical p value (typically, .05) given that H_0 is true (see, e.g., Rozeboom, 1960; Wagenmakers, 2007). NHST thus provides a decision criterion without the researcher having to identify a specific quantitative alternative hypothesis (H_1). For this reason, one may reject or fail to reject H_0, but one cannot "accept" it: H_0 could be false, but the statistical test might not lead to the correct decision to reject H_0. The probability of correctly rejecting H_0 is the *power* of the test to detect a specific H_1. In general, a power of .8 is considered to be good or adequate (Cohen, 1988), but higher levels of power are desirable if a high priority is to detect an effect if it exists.

Power analysis in the context of NHST may be useful in several ways (Faul et al., 2007). First, power analysis enables planning of experiments by allowing the researcher to estimate a priori the sample size required for a given population effect size. Second, post hoc analysis of power may be useful after a study is completed in order to evaluate whether there was adequate power to detect a relevant population effect size (Faul et al., 2007; Yuan & Maxwell, 2005, p. 142). The effect size may be specified on a priori grounds or may be based on common conventions (Cohen, 1988). Third, *sensitivity* analysis of power provides statements about the precision of statistical tests (i.e., the p of detecting a specified effect of a given size; Faul et al., 2007). For example, one can calculate the minimum population effect size detectable with a power of .8. Both post hoc power and sensitivity analysis, as defined here, are equivalent to a prospective power analysis, because they do not depend on the specific results of a study (Yuan & Maxwell, 2005). Specifically, they are calculated using only the relevant population effect size, α, the power, and N, given the ANOVA design and effect of interest (i.e., a specific main or interaction effect). Consequently, they avoid the pitfalls of "observed" power that is calculated as a sample-based estimate of the population power (see, e.g., Hoenig & Heisey, 2001; Yuan & Maxwell, 2005). Post hoc power and sensitivity analysis are valid only if researchers specify population effect sizes on a priori grounds (Faul et al., 2007).

Bayesian analysis for ANOVA

Bayesian analysis of posterior probabilities for H_0 versus H_1 is poised to emerge as a widely used alternative to NHST in psychological research (e.g., Dienes, 2011; Kruschke, 2011; Masson, 2011; Raftery, 1995; Wagenmakers, 2007; Wetzels et al., 2011; see also Dixon, 2003; Glover & Dixon, 2004, for related alternatives). In NHST, statistical inference is based on the probability of observing a certain effect size or difference (D) if the null hypothesis is true; that is, if $p(D \mid H_0)$ is less than .05, then reject H_0. Of more value, however, is knowing about the likelihood that a hypothesis is true given the data. The probability that H_0 is true given D is the posterior probability $p(H_0 \mid D)$. It may seem intuitive that $p(D \mid H_0)$ and $p(H_0 \mid D)$ will be closely linked, but this is not necessarily the case (Berger, 1985; Wagenmakers, 2007). In fact, conditions under which $p(D \mid H_0)$ is less than .05, which would lead to rejection of H_0, can correspond to high values of $p(H_0 \mid D)$ that would suggest that H_0 was in fact true (Berger, 1985; Masson, 2011).

This dissociation reflects a difference between Bayesian analysis and NHST in the effect of sample size on the evidence for H_0 versus H_1 provided by the data (Masson, 2011; Wagenmakers, 2007). In Bayesian analysis, the posterior probability favoring the null hypothesis grows as sample size grows. In contrast, the NHST p value is not affected by sample size. Consequently, NHST p values tend to overestimate the evidence for H_1 relative to a Bayesian analysis, and this tendency increases with the number of observations (Masson, 2011, p. 688; Wagenmakers, 2007, p. 796). Additionally, when NHST p values are close to the rejection region (e.g., .01 to .05), Bayesian analysis often indicates only weak evidence favoring H_1 (Wetzels et al., 2011). Unlike Bayesian analysis, likelihood ratio calculations based on the Akaike information criterion, or AIC (see, e.g., Akaike, 1974), assume that p is not affected by sample size (Wagenmakers, 2007, p. 796), but like NHST, approaches based on the AIC also have a bias to favor H_1 (Rouder, Speckman, Sun, Morey, & Iverson, 2009, p. 228).

Despite concerns about the validity of NHST p values as evidence (see also Dixon, 2003), alternative approaches such as Bayesian analysis are not yet routine in connection with ANOVA, perhaps because their calculations seem to be complicated or ambiguous and their interpretation is unfamiliar to many (Wagenmakers, 2007). Following Wagenmakers (2007), Masson (2011) presented a straightforward approach to calculate and interpret the posterior probabilities $p(H_0 \mid D)$ and $p(H_1 \mid D)$ in the context of standard ANOVA. This approach is incorporated into MorePower 6.0, thereby permitting the user to calculate posterior probabilities for ANOVA effects by specifying only the design, sample size, and effect size (e.g., F or η_p^2) (see also Rouder et al., 2009, for a Bayesian analysis of t tests, with an online applet at http://pcl.missouri.edu/bayesfactor).

Calculation of the posterior probability for H_0 is based on the Bayes factor (BF), which is the odds ratio $p(D|H_0)/p(D|H_1)$. Following Masson (2011) and

Wagenmakers (2007), estimation of BF in MorePower is based on the relation $BF \approx p_{BIC}(D|H_0)/p_{BIC}(D|H_1) = e^{\Delta BIC/2}$, which uses the Bayesian information criterion (BIC) to approximate BF (Raftery, 1995, 1999; see also Rouder et al., 2009). The BIC provides "an objective baseline reference for automatic Bayesian hypothesis testing" (Wagenmakers, 2007, p. 797) and provides a good approximation of BF when the unit-information prior is assumed (Raftery, 1995, 1999; Wagenmakers, 2007, Appx. B). The BIC is somewhat conservative with respect to providing evidence for H_1 as compared to some other objective priors (Raftery, 1999; Wagenmakers, 2007). Raftery (1999, p. 412) noted that this conservative behavior favors "reporting BIC as a baseline reference analysis even if final conclusions are drawn using a different prior" (see, e.g., Rouder et al., 2009).

Given an estimate of BF, the posterior probability for H_0 is given by $p_{BIC}(H_0|D) \approx BF/(BF+1)$, and the posterior probability for H_1 is the complement $p_{BIC}(H_1|D) = 1 - p_{BIC}(H_0|D)$.

Calculation of ΔBIC, which is the difference in BIC values for the null and alternative hypothesis models, is given by Eq. 1 for ANOVA (Masson, 2011; Wagenmakers, 2007):

$$\Delta BIC = n \ln(SSE_1/SSE_0) + (k_1 - k_0)\ln(n). \quad (1)$$

In Eq. 1, n is the number of independent observations contributing to an effect, SSE_1/SSE_0 is the ratio of the error sums of squares for the alternative and null hypothesis models, and $k_1 - k_0$ is the difference between the models in the number of free parameters. For effects composed only of IM factors, n is the number of subjects. When the effect of interest includes RM factors, n is equal to the number of subjects multiplied by the degrees of freedom associated with the RM factor(s) (Masson, 2011).[1] MorePower calculates ΔBIC on the basis of two equalities pointed out by Masson (2011, p. 682). First, the ratio SSE_1/SSE_0 equals the complement of partial eta-squared, $(1 - \eta_p^2)$. η_p^2 is a common measure of effect size that can be entered directly or calculated by MorePower from the observed F value or MST from an ANOVA. Second, the quantity $k_1 - k_0$ equals the degrees of freedom for the effect of interest when ΔBIC contrasts the null and alternative hypothesis models. These substitutions produce Eq. 2, used by MorePower to calculate ΔBIC:

$$\Delta BIC = n \ln\left(1 - \eta_p^2\right) + df \cdot \ln(n). \quad (2)$$

This quantity is then used to compute the Bayes factor and the posterior probability using the formulas presented previously.

[1] Masson (2011, p. 682) notes that whether n should be adjusted for RM factors remains an open issue (see also Rouder et al., 2009; Wagenmakers, 2007).

A fundamental advantage of the Bayesian approach is that it affords a graded comparison of $p(H_0 \mid D)$ and $p(H_1 \mid D)$, rather than only a binary decision to reject or not reject H_0, as in NHST (Masson, 2011; Wagenmakers, 2007). Raftery (1995; see also Masson, 2011) proposed a graded interpretation of the posterior probability. Values of $p(H_0 \mid D)$ from .5 to .75 may be classified as "weak evidence," from .75 to .95 as "positive evidence," from .95 to .99 as "strong evidence," and >.99 as "very strong evidence." Wetzels et al. (2011, p. 293) presented a more finely graduated evidence scale for the interpretation of BF, adapted from Jeffreys (1961) (see Table 1). These values provide approximate descriptive rules of thumb to summarize the results of Bayesian analysis.

Relational CIs for ANOVA

The use of relational CIs for ANOVA is an influential alternative or augmentation to NHST (Cumming & Finch, 2005; Fidler, Thomason, Cumming, Finch, & Leeman, 2004; Hollands & Jarmasz, 2010; Jarmasz & Hollands, 2009; Masson & Loftus, 2003). Unlike CIs calculated for an individual mean, relational CIs are computed to provide a visual basis to identify a pattern of relations among means (Loftus & Masson, 1994; Masson & Loftus, 2003). Statistical inference based on CIs is subject to the same criticisms as NHST (see, e.g., Rouder & Morey, 2005; Rouder et al., 2009). Nonetheless, relational CIs convey the degree of precision in the measurement of an effect and provide a graphic index of the replicability of effect sizes that is not

Table 1 Evidence categories for Bayes factors (adapted from Wetzels et al., 2011, p. 293)

Bayes Factor	Interpretation
>100	Decisive evidence for H_0
30–100	Very strong evidence for H_0
10–30	Strong evidence for H_0
3–10	Substantial evidence for H_0
1–3	Anecdotal evidence for H_0
1	No evidence
0.333–1	Anecdotal evidence for H_1
0.1–0.333	Substantial evidence for H_1
0.0333–0.1	Strong evidence for H_1
0.01–0.0333	Very strong evidence for H_1
<0.01	Decisive evidence for H_1

Following Masson (2011) and Wagenmakers (2007), the Bayes factor was computed to represent the odds ratio for H_0 over H_1. High values favor H_0, and low values favor H_1. Wetzels et al. (2011) computed BF to represent the odds of H_1 over H_0; consequently, their evidence categories for H_0 over H_1 were opposite those in Table 1 (i.e., substitute H_0 for H_1)

inherent in standard hypothesis testing (Rouder & Morey, 2005). Psychological journals often require error bars in graphs that present means. Consequently, an attractive aspect of this approach is that it provides explicit rules for CIs to standardize how such error bars should be constructed (cf. Rouder & Morey, 2005). Furthermore, relational CIs for two means bear a simple relation to NHST: The difference between two means will be significant by an ANOVA or *t* test if it is greater than the CI's margin of error (i.e., half of the width of the CI) multiplied by a factor of $\sqrt{2}$ (Jarmasz & Hollands, 2009; Loftus & Masson, 1994).

For all ANOVA tests, MorePower calculates relational CIs for the effect of interest on the basis of formulas proposed by Loftus and Masson (1994) and Masson and Loftus (2003) and revised for RM effects by Jarmasz and Hollands (2009). Here, we focus on how MorePower calculates the margin of error for the CI. For example graphs and interpretative strategies, see Cumming and Finch (2005), Jarmasz and Hollands (2009), and Masson and Loftus (2003). The general form of the formulas presented by Jarmasz and Hollands is shown in Eq. 3.

$$M_i \pm t_{critical} \cdot \sqrt{\frac{MSE}{n \cdot L/r}} \tag{3}$$

M_i represents the mean of any cell in the effect of interest, and the right side of the equation is the margin of error for the CI. The critical *t* value is based on the *df* for the *MSE* of the effect of interest. MorePower follows the practice of constructing CIs using the *MSE* from the corresponding fixed-effects ANOVA (Jarmasz & Hollands, 2009; Loftus & Masson, 1994; but see Blouin & Riopelle, 2005). In the denominator, *n* is the number of participants in each IM cell of the effect of interest, or if the effect has only RM factors, then *n* equals the total sample (*N*). *L* is the product of the number of levels of all RM factors in the design, and *r* is the product of the number of levels of the RM factors in the effect of interest. In MorePower, *L* is set to 1 when there are no RM factors. Similarly, *r* is set to 1 when there are no RM factors in the effect. Consequently, $L/r = 1$ when there are no RM factors (i.e., this term drops out of the equation), and $L/r = L$ when there are RM factors in the design but not in the effect. The denominator term $n \cdot L/r$ is the number of observations contributing to each mean being compared in the effect (Jarmasz & Hollands, 2009).

The relational CI function in MorePower 6.0 implements Formulas 1–5 in Table 4 of Jarmasz and Hollands (2009, p. 130). These apply, respectively, to (1) an IM main effect (IM designs), (2) an RM main effect (RM and mixed IM–RM designs), (3) an RM interaction (RM and mixed IM–RM designs), 4) an IM main effect (mixed IM–RM designs), and (5) a mixed IM–RM interaction (comparing RM conditions).

Overview of MorePower 6.0

The following sections provide an introduction to MorePower 6.0 and explain a series of built-in examples that illustrate its application to ANOVA.[2] MorePower was developed under Microsoft Windows, but it can be installed on other platforms that will run MS Windows as a virtual machine or an alternative OS. The calculator uses high-precision algorithms, but the validity of its calculations requires that the relevant statistical assumptions for ANOVA be satisfied (i.e., normal distributions, sphericity, and homogeneity of the variances and covariances; see, e.g., Hays, 1994). Probabilities for the noncentral *F* distribution are calculated to eight digits of precision using the CDFDNF algorithm (Reeve, 1986; see also Bradley, Russell, & Reeve, 1996). To obtain the calculator, go to https://wiki.usask.ca/pages/viewpageattachments.action?pageId=420413544 and download MorePower6_setup.zip. Unzip the file in a temporary directory, open the Package folder, and run setup.exe. This will install the latest version of MorePower.

The MorePower interface (see Fig. 1) consists of a collection of framed sections allowing the user to select from various options and to enter numerical input into a text field. Placing the cursor over any of these objects displays a short description of it. There is also a main text output window in which a detailed summary (described later) of the last calculation is presented.

The radio buttons in the Analysis section allow the user to select from a set of six analysis types, including ANOVA, one- or two-sample *t* test, one or two-sample *z* test of proportions, and simple correlation. The Design Factors and Effect of Interest sections are used to specify the ANOVA design and effect (see the following section for details). The text field in the Alpha section specifies the desired Type I error rate and can be toggled between tests for one side or two sides, which is mandatory for ANOVA. The Sample and Power text fields contain the specified or calculated total sample size and power, respectively.

The Effect Size section provides several options for specification of the desired or computed effect size. When the upper radio button is selected, effect size may be specified in terms of η_p^2 or the difference in original units when treatment $df = 1$ (i.e., when the effect of interest is composed exclusively of factors with two levels). The upper radio button is automatically selected when solving for sample size. Selection of the lower radio button in the Effect Size

[2] The focus here is on ANOVA, but MorePower 6.0 includes power calculations for *t* tests of means and simple correlation and *z* tests of binomial proportions. These use the power formulas for $df = 1$ tests described by Campbell and Thompson (2002). Supplementary documentation of these procedures is included in the distribution software for MorePower 6.0.

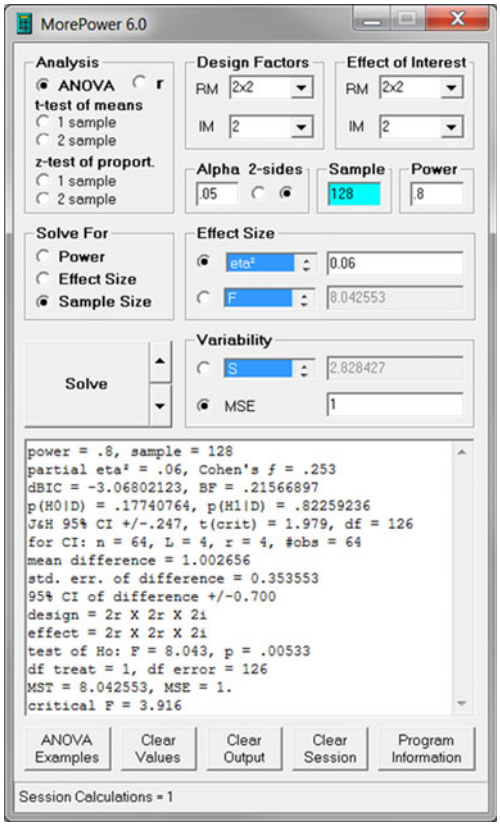

Fig. 1 MorePower 6.0 interface and output window (showing ANOVA Example 1)

section allows the user to specify (or request) effect size in terms of a test statistic, including the *F* ratio, *MST*, or *t*.

In the Variability section, the user may select the lower button for *MSE*, which is the default for ANOVA. Alternatively, for tests with treatment $df = 1$, the user may select the upper radio button in the Variability section, which allows the user to scroll and click to select standard deviation (*S*, the default), variance (S^2), or standard error (*SE*). The *SE* option is not available when solving for sample size. If eta^2 is selected for effect size (the default), the value specified for variability has no bearing on the calculation of sample size, effect size, or power. If there is no value in the appropriate Variability field when an analysis type is selected, then a value of 1 is assigned. For tests with $df = 1$, the value specified in Variability is used for the calculation of the mean difference in original units and its variability (see Eqs. 4 and 5, discussed later).[3]

The Solve For section allows the user to toggle among solving for power, effect size, or sample size. The user selects one of these to be solved for and supplies values for the other two. Once selected, the corresponding Sample, Power, or Effect Size text field turns light blue, indicating where the result of the calculation will appear. Click the Solve button to run the calculator using the current information in the relevant input fields. The up–down arrows next to the Solve button scroll forward or backward through session calculations, with the current total number of calculations displayed at the bottom left corner of the calculator.

Finally, the five buttons at the bottom of the calculator provide the following functions: Clicking on ANOVA Examples displays a list of examples that are discussed later in this document. Enter the number for the desired example and press return. The result of running an example is equivalent to entering the example input information into the appropriate fields and clicking Solve. The Clear Values button initializes all of the input fields. Clear Output clears the main output window. Clear Session deletes the record of all calculations since the calculator was started and resets the session calculation count to 0. Program Information displays the current version information and contact information.

Specification of ANOVA design factors and the effect of interest

The overall design of the experiment is specified using the IM and RM fields in the Design Factors section (see Fig. 1). To do this, the user enters the number of levels for each of the IM factors in the IM field and the number of levels for RM factors in the RM field. For example, if the experiment consisted of two IM and two RM factors, each with two levels, enter 22 for IM and 22 for RM. For a 2 × 2 × 3 RM design, enter the sequence 223 (or 232 or 322) for RM and leave IM blank. To simplify this process, there is a drop down menu for each of the IM and RM fields that lists common designs. Any design may be specified, however, by typing a series of single digits numbers in the RM and IM fields (a factor can have a maximum of nine levels). Once the design factors have been entered, specify the effect using the corresponding RM and IM fields in the Effect of Interest section. The user may enter a digit sequence representing the levels of factors in each field, but the down arrows provide a drop-down list of relevant effects, given the design factors already specified.

Specification of effect size for ANOVA

MorePower 6.0 provides several options for entering or computing effect size for ANOVA. Either the upper Effect Size field (eta^2 or *difference*) or the lower Effect Size field (*F*, *MST*, or *t*) may be selected. The default is partial eta-squared. Denoted η_p^2, it is based on the sums of squares for

[3] If *difference* is selected rather than eta^2 for effect size, then with $S = 1$, the difference specified or calculated is in standard deviation units (d_z). This is another common measure of effect size (see, e.g., the one-sample *t* tests of means in G*Power 3; Faul et al., 2007).

the effect of interest: $SS_{treatment}/(SS_{treatment} + SS_{error})$. η_p^2 therefore ranges between 0 and 1. This quantity represents the amount of total variability in the effect attributable to the independent variable(s). In other words, η_p^2 is the amount of unique variance explained by the effect of interest, divided by the proportion of variance unexplained by all of the other main and interaction effects in the design (Levine & Hullett, 2002).[4] As a measure of ANOVA effect size, η_p^2 is frequently used by psychological researchers, and it is the effect size measure reported in example manuscripts presented in the *Publication Manual of the American Psychological Association, Sixth Edition* (2010; see, e.g., p. 46). The general linear model (GLM) procedure for repeated measures ANOVA in IBM SPSS 19.0 also computes η_p^2 to represent the observed effect size. Thus, η_p^2 is a common effect size measure for ANOVA. Cohen's *f*, another common effect measure for ANOVA, can be derived from η_p^2 by the relation $f = \sqrt{\eta^2 \div (1 - \eta^2)}$ (see Cohen, 1988, p. 281). MorePower includes the conversion of η_p^2 to *f* in its ANOVA output. By convention, values of .10, .25, and .40 of Cohen's *f* correspond to small, medium, and large effect sizes, respectively (Cohen, 1988, p. 355). For η_p^2, these conventional values convert to .01, .06, and .14.

Effects with one degree of freedom For ANOVA effects with numerator *df* = 1, effect size may be specified (or calculated by MorePower) as the mean difference in the original units (percentage correct, milliseconds, etc.). This allows the user to calculate sample size or power for an effect of a particular size specified in familiar units (e.g., the sample size required to detect a 10-ms difference with a power of .8). Selecting the upper radio button in the Effect Size section and clicking on *difference* enables this option. This can be used for any 2^k effect in the design, where *k* is the number of two-level factors in the effect of interest. Any such effect will have numerator *df* = 1 and may consist of RM, IM, or combined RM × IM effects, even when the 2^k effect is nested in a design that includes factors with more than two levels (Campbell & Thompson, 2002). The difference (*d*) is the size of the observed or specified difference that one is interested in. For a two-level main effect, *d* is the absolute difference between the means for the two levels of the factor of interest, averaged over the levels of any other factors. For a 2 × 2 interaction, *d* is the mean difference of differences $|M_{11} - M_{12}| - |M_{21} - M_{22}|$ averaged over the levels of other factors, and for a 2 × 2 × 2 interaction it is the mean difference between the mean differences of differences $[|M_{111} - M_{112}| - |M_{121} - M_{122}|] - [|M_{211} - M_{212}| - |M_{221} - M_{222}|]$, and so on for all 2^k effects. For example, if M_{11}, M_{12}, M_{21}, and M_{22} were equal to 60, 30, 20, and 10, respectively, then *d* for the 2 × 2 interaction would equal |60 – 30| – |20 – 10| or 20.

In MorePower, *d* is calculated using Eq. 4, developed by Campbell and Thompson (2002). *B* is the number of two-level IM (i.e., between-subjects) factors in the effect of interest, *W* is the number of two-level RM (i.e., within-subjects) factors in the effect, and *L* is the total number of RM cells in the design (i.e., the product of all RM factor levels). The quantity *n* is the number of observations for each treatment level in the effect of interest. For a 2^k effect composed entirely of RM factors (i.e., no IM factors), *n* is the total number of participants (*N*). If the 2^k effect includes one IM factor, then *n* is the number of observations contributing to each level of that factor ($n = N/2$); if the effect includes two IM factors (e.g., a 2 × 2 × 2 interaction involving two IM factors and one RM factor), then *n* is the number of observations in each of the four IM cells ($n = N/4$), and so forth. The value of *n* is derived by MorePower from the specification of the design and the effect, and from the total *N* entered in the Sample field.

$$d = \sqrt{\left(MST \cdot 2^B \cdot \frac{2^{W \cdot 2}}{L}\right)/n} \tag{4}$$

The *MST* value used by MorePower to calculate *d* is based on the relation $MST = MSE \times F$. *F* may be input directly if the lower radio button in the Effect Size section is selected, but if the upper radio button in that section is selected, *F* is calculated from η^2 and the *MSE* value specified in the box in the Variability section. Instead of *F*, however, the user can alternatively select *MST* or *t*. The *t* option reflects the fact that *F* with numerator *df* = 1 is equal to t^2. Similarly, instead of *MSE*, one may select the upper radio button in the Variability section, which allows the user to alternatively specify the variability of *d* in terms of variance, standard deviation, or standard error. This gives great flexibility in specifying the variability for 2^k effects. These options are based on Eq. 5, derived by Campbell and Thompson (2002). This equation calculates the variance of the difference (s_d^2) from the *MSE* for any 2^k effect, with *B*, *W*, and *L* as defined previously. The standard deviation (s_d) is the square root of s_d^2. The standard error (*SE*) is $S_d/\sqrt{n}$.

$$s_d^2 = MSE \cdot 2^B \cdot \frac{2^{W \cdot 2}}{L} \tag{5}$$

Contents of the output window for ANOVA calculations

Figure 1 shows the MorePower interface after running the built-in Example 1, which calculates the sample size required (with power = .8) for a 2 × 2 × 2 interaction in a

[4] It is important to distinguish η_p^2 from η^2, which is the proportion of explained variance relative to the total sums of squares including all experimental effects. These quantities are the same if the design has only one factor, but η_p^2 is larger than η^2 in multifactor designs (Levine & Hullett, 2002; Pierce, Block, & Aguinis, 2004).

design with two RM factors and one IM factor (the built-in examples are described in detail in the next section). Here, we describe the contents of the output window for ANOVA calculations. MorePower's main text output window first presents power and total sample (*N*), followed by η_p^2 and Cohen's *f*. The following two lines contain output associated with the Bayesian analysis, including ΔBIC (dBIC in the display), the estimated Bayes factor (BF), and estimated posterior probabilities $p_{BIC}(H_0 \mid D)$ and $p_{BIC}(H_1 \mid D)$. The next two lines report information about the Jarmasz and Hollands (2009) CI for this interaction. Examples related to the Bayesian analysis and CI functions in MorePower are discussed in the subsequent sections. As the 2 × 2 × 2 interaction has *df* treatment = 1, the mean difference (*d*) in original units (1.003) is presented on the basis of Eq. 4. This number represents the magnitude of the interaction effect in original units. The *SE* of the difference (*d*) based on Eq. 5 is included, as well as the 95% [i.e., 100(1 – α)] CI based on the *SE* and *t*, with *df* equal to the df_{error} for the interaction. Next, the output displays the design and effect in factorial format (e.g., 2r is a two-level RM factor; 3i is a three-level IM factor). The remaining lines in the output present the standard test of the null hypothesis for the interaction, including the observed *F* ratio, the observed significance level (p), the error and treatment *df* for the effect of interest, the *MST* and *MSE*, and the critical *F* ratio. Of course, not all of this information will always be relevant, but any part or all of the output may be copied and pasted to be used as needed.

Built-in examples 1–4: calculating sample size and effect size for ANOVA designs

MorePower calculates the required sample size given the current values for the design and effect of interest, effect size, α, and power. Alpha is set by default to .05 two-sided and does not need to be entered except to change it from the default. The calculator solves for *n* per IM cell and multiplies this result by the number of IM cells to obtain the total sample size (*N*). Calculation of the sample size for ANOVA has a maximum limit of *n* = 2,500 per IM cell or a total *N* = 2,500 for an all-RM design. This largely avoids long waits (e.g., 30 s) owing to slow processing by CDFDNF. If the limit is reached, a warning appears, and the result is suspect. This limit does not apply to calculation of power or effect size for ANOVA. Sample size is a discrete variable in MorePower, whereas power is a continuous variable; consequently, the calculated sample size will not correspond exactly to the power specified. To find the exact power for a calculated sample size, the user can select Power under Solve For and click Solve, which displays the exact power given the calculated *N*.

Built-in Example 1 (see Fig. 1) illustrates the use of MorePower to calculate the sample size required for a 2 × 2 × 2 interaction in a design with two RM factors and one IM factor. The design and effect were specified by entering 22 in the RM field and 2 in the IM field of the Design Factors and Effect of Interest sections; alternatively, the drop-down lists could be used. The desired power of .80 was entered in the Power field, and a medium effect size of .06 for eta^2 was specified in the Effect Size field. An entry in the Variability section (e.g., *MSE*) is required, although this value does not affect the calculation of the required sample size. This is because error variability is implicit in η_p^2 [i.e., it is *SST*/(*SST* + *SSE*)]. Enter a value if prompted (the calculator will initialize to a default value of 1). The required sample of 128 appears in the Sample field after a second or two. Thus, a total sample of at least 128 is required to have a .8 probability to correctly reject the null hypothesis (i.e., power = .8) given a medium effect size (η_p^2 = .06) for the 2 × 2 × 2 mixed RM–IM interaction.

Built-in Examples 2 and 3 specify a five-factor mixed RM–IM design with 2 × 2 RM (within-subjects) factors combined with 2 × 2 × 3 IM (between-subjects) factors. In Example 2, MorePower calculates the sample size required with η_p^2 = .06 and power ≥ .8 for an IM–RM 2 × 2 interaction (required $N \geq 144$). Example 3 illustrates a sensitivity analysis and calculates the minimum population effect size required ($\eta_p^2 > .144$) for the five-way interaction, given power = . 8 and sample size = 72. Example 4 illustrates a 3 × 3 RM design and calculates the sample size required ($N \geq 26$) for a large-effect-size interaction (η_p^2 = .14) and power = .9.

Built-in examples 5–8: Bayesian analysis

ANOVA Examples 5 through 8 are included to illustrate MorePower's calculations for the estimated Bayesian posterior probabilities. They are based on previously published examples from Masson (2011) and Wagenmakers (2007). Example 5 refers to the ANOVA design with one three-level RM factor in Table 1 of Masson (2011, p. 683). To create this example, only three quantities were required: the design (three-level RM factor in the Design Factors and Effect of Interest fields), the sample size (40 in the Sample field), and the observed *F* value from Masson's Table 1 (*F* = 12.9, although the *MST* of .178 could be used instead).[5]

Figure 2 shows MorePower 6.0 after running Example 5. Masson (2011, p. 683) reported ΔBIC = –14.21, BF =

[5] In this example, *MSE* affects only the value of *MST* in the output; consequently, any positive value may be entered for *MSE* and the rest of the output will be unaffected. The *MSE* = .014 used in Example 5 is from Table 1 in Masson (2011).

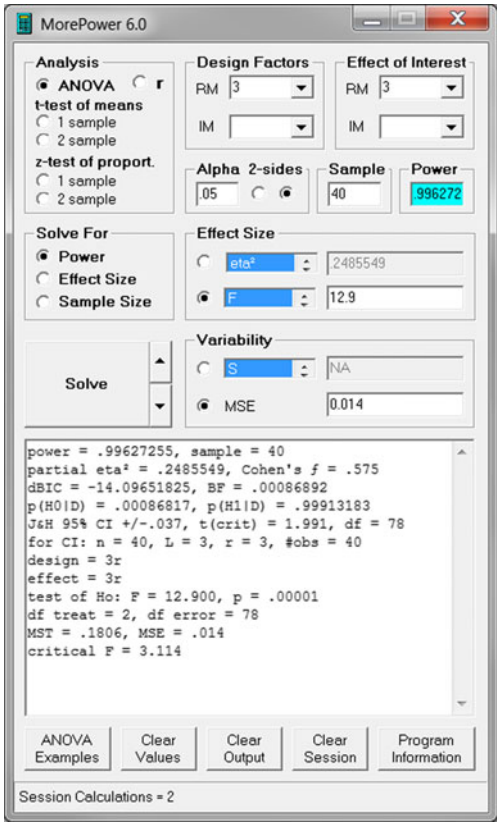

Fig. 2 ANOVA Example 5 refers to the Bayesian analysis of a three-level RM ANOVA (based on Table 1 in Masson, 2011)

0.000859, $p_{BIC}(H_o \mid D) = .00086$, and $p_{BIC}(H_1 \mid D) = .9991$. The corresponding quantities calculated by MorePower are $\Delta BIC = -14.10$, BF = 0.00087, $p_{BIC}(H_o \mid D) = .00086817$, and $p_{BIC}(H_1 \mid D) = .9991$. Using F to represent effect size, the Bayesian analysis calculated by MorePower is in very close agreement, even though F was specified to only one decimal place. Using Raftery's (1995) rules of thumb, the Bayesian results provided "very strong" evidence in favor of the alternative hypothesis.

Bayesian Examples 6 and 7 refer to the 2 × 2 RM ANOVA reported in Table 3 of Masson (2011, p. 685). Example 6 is the main effect of alignment for which Masson reports $\Delta BIC = -3.15$, BF = 0.2070, $p_{BIC}(H_0 \mid D) = .171$, and $p_{BIC}(H_1 \mid D) = .829$. This provides positive evidence for H_1. Example 7 is the 2 × 2 (Alignment × Delay) interaction for which Masson reports $\Delta BIC = -10.44$, BF = 0.0054, $p_{BIC}(H_0 \mid D) = .005$, and $p_{BIC}(H_1 \mid D) = .995$. This indicates "decisive" evidence for H_1 (Table 1 above). Finally, Example 8 refers to the Bayesian analysis of the 2 × 2 IM interaction reported in Table 4 of Wagenmakers (2007, p. 799), who reports $\Delta BIC = -4.31$, $p_{BIC}(H_0 \mid D) \approx .10$, and $p_{BIC}(H_1 \mid D) \approx .90$. These examples demonstrate that MorePower accurately reproduces the estimated Bayesian posterior probabilities based on Masson (2011) and Wagenmakers (2007) in each case.

Built-in examples 9–13: relational confidence intervals

The relational CIs calculated by MorePower correspond to Formulas 1–5 in Jarmasz and Hollands (2009, Table 4). The confidence level (e.g., 99%, 95%, or 90%) for the CI is controlled by setting alpha (e.g., .01, .05, or .10, respectively), which is .05 by default; thus, the 95% CI is calculated by default. To demonstrate MorePower's CI function, the built-in ANOVA examples include the illustrative cases presented by Jarmasz and Hollands. Their example design had one IM factor with four levels (6 participants in each) and four RM factors in a 4 × 2 × 2 × 2 design. Figure 3 shows MorePower after running Example 9. The Design Factors RM field contains 4 × 2 × 2 × 2, and the IM field contains 4, which, together, represent the five-factor mixed IM–RM design. The Effect of Interest section contains a 2 in the RM field, and the IM field is blank. Thus, the effect of interest is the main effect of a two-level RM factor (the within-subjects Direction factor; see Jarmasz & Hollands, 2009, p. 129). Apart from the design and effect, MorePower also required input of the total Sample size (24), the observed F ratio (11.8), and the MSE for this effect (1462) (see Jarmasz & Hollands, 2009, Table 5, p. 131).

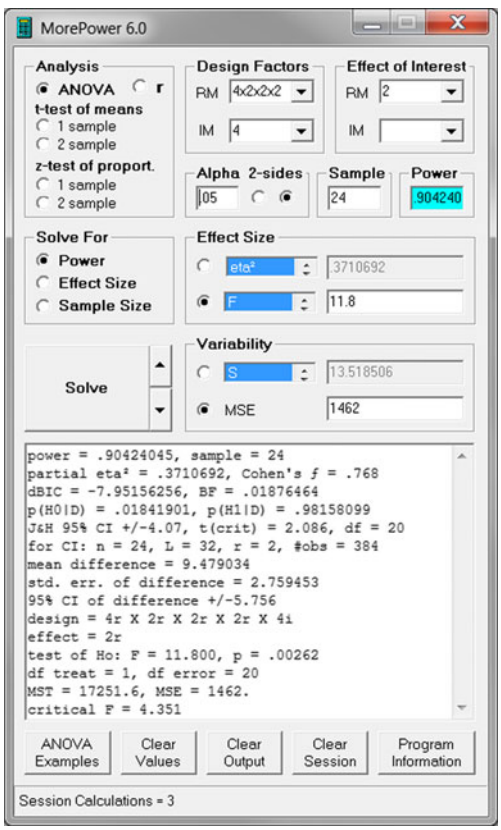

Fig. 3 ANOVA Example 9 refers to the relational confidence interval calculations for a repeated measures main effect in a mixed IM–RM design (Jarmasz & Hollands, 2009, p. 129)

The output window (see Fig. 3) presents the information discussed previously, but we will focus here on the Jarmasz and Hollands (2009) CI, which is reported as J&H 95% CI +/-4.07, with a critical value of $t(20) = 2.086$, MSE = 1,462, $n = 24$, $L = 32$ (i.e., 4 × 2 × 2 × 2), $r = 2$ (i.e., the effect of interest has two RM cells), and #obs (i.e., number of observations) = 384. These results correspond to the solution presented by Jarmasz and Hollands (2009, p. 129) as follows.

$$M_i \pm 2.086 \cdot \sqrt{\frac{1462}{24.32/2}}$$
$$M_i \pm 4.1$$

As this effect has one degree of freedom in the numerator, MorePower also displays the mean difference in original units (9.48% from Eq. 4), the standard error of the difference (2.76 from Eq. 5), and the 95% confidence interval of the difference (±5.76%). The difference between the CI constructed using Eq. 3 (±4.07%) and the 95% CI based on the SE (±5.76%) reflects the relation between relational CIs and significance tests described previously; specifically, differences between means greater than the CI margin of error multiplied by $\sqrt{2}$ are significant at the specified α level (.05 by default). Consistent with this, the relational CI margin of error (4.07) multiplied by $\sqrt{2}$ equals 5.76, the margin of error for the NHST-based 95% CI.

ANOVA Examples 10–12 further demonstrate the J&H CI function in MorePower using Jarmasz and Hollands's (2009) illustrative analyses. Example 10 corresponds to the CI for the 4 × 2 RM effect (±3.4 on p. 131). Example 11 is the CI to compare RM conditions within IM cells for the 4 × 2 mixed IM–RM effect (±3.2 on p. 133). Example 12 refers to the CI for the main effect of the four-level IM factor (±6.3 on p. 134). This experimental design included only one IM factor, and therefore did not afford a demonstration for interactions in multifactor IM designs with no RM factors. Example 13 confirms the application of MorePower's CI function for the case of a 2 × 2 IM interaction ($n = 12$ per group) taken from Masson and Loftus (2003; CI ±0.055 on p. 209). These built-in examples allow the reader to confirm that MorePower generates the appropriate quantities to reproduce the desired relational CI in each case.

MorePower`s CI function for mixed IM–RM effects does *not* apply, however, when the researcher wishes to compare IM means within levels of RM factors. The calculator uses the MSE for the IM–RM interaction, which is the error variability for the RM factor computed within IM conditions (Jarmasz & Hollands, 2009, p. 128). This fact is reflected in the source table of any mixed IM–RM ANOVA, which shows that both the MSE and df_{error} are the same for an RM effect and for any interactions of the RM factor(s) with IM factors. Consequently, in contrast to Example 11, where the CI based on the MSE affords comparison of RM means within an IM condition, the CI to compare IM means within an RM condition requires calculation of the "pooled mean square within cells"(MS_{wc}) and a corresponding df (MS_{wc}) (see Jarmasz & Hollands, 2009, pp. 128–129; Formula 6 on p. 130). One can enter the MS_{wc} rather than MSE in MorePower, but the df_{error} cannot be specified directly (it is derived from the design and effect information). Consequently, the calculator will compute the correct CI for IM-within-RM comparisons only when $df(MS_{wc}) = df(MSE)$. This will rarely be the case. Nonetheless, MorePower can be used to compute the number of observations (i.e., $n\ L/r$ in Eq. 3) required for the hand calculation of the CI, because the number of observations is based on the total sample size, design, and effect information, regardless of which type of comparison is desired.

MorePower 6.0 and reducing errors in statistical reports

Another useful application of MorePower 6.0 is to check for errors in reporting of statistical analyses. Bakker and Wicherts (2011) estimated that a high percentage (perhaps 18%) of statistical results in the psychological literature contain errors. In many instances, the errors are simple typos or oversights (e.g., copying an ANOVA description as a template for a similar analysis, but failing to make all of the necessary changes). MorePower allows the user to quickly specify an ANOVA effect of practically arbitrary complexity and then to review the p value, degrees of freedom, effect size (η_p^2 or original units for one-degree-of-freedom tests), and so forth. This makes it straightforward to confirm that components of reported statistical tests are internally consistent and correspond to the intended analysis. It should be noted that the APA manual stipulates reporting effect size measures such as η_p^2, but such measures provide no more information than the test statistic itself (e.g., F, t, or z). Reporting a measure of error variability (e.g., MSE or SE) provides valuable additional diagnostic information about the accuracy of statistical reports.

Error handling and help

If required input is missing or invalid, MorePower 6.0 identifies the problem item with a query, and the cursor appears in the corresponding input field. The program also generates a warning message on any internal error; simply follow the instructions, or click the Clear Values button and retry. If a persistent error occurs, please e-mail jamie.campbell@usask.ca with the error number and a screenshot of the calculator.

Summary and conclusions

MorePower 6.0 computes exact power, sample size, and effect size statistics for factorial ANOVA designs. The calculator also provides a straightforward process for estimating Bayesian posterior probabilities for the null and alternative hypotheses based on Masson (2011), and also computes CIs based on the Jarmasz and Hollands (2009) formulas. A unique feature of MorePower is that it affords direct comparison of these three approaches for interpreting ANOVA. Other functions include power analyses for one- or two-sample *t* tests or proportions, as well as simple correlation (see Campbell & Thompson, 2002, and the supplementary document included in the calculator download) and a probability calculator for the *F*, *t*, and *z* distributions. The program's high numerical precision for ANOVA and ability to work with complex ANOVA designs could further facilitate researchers' attention to issues of statistical power, the use of relational CIs for data interpretation, and Bayesian analysis for ANOVA. The program may be especially useful to researchers as they consider a transition from standard NHST to Bayesian analysis. Of course, the Bayesian analysis based on the BIC in MorePower is one approach among a variety of likelihood-ratio methods that offer practical alternatives to NHST (see, e.g., Dixon, 2003; Glover & Dixon, 2004; Rouder et al., 2009; see also Wagenmakers, 2007, p. 794).

Author Note This research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada.

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

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: American Psychological Association.

Bakker, M., & Wicherts, J. M. (2011). The (mis)reporting of results in psychology journals. *Behavioural Research, 43,* 666–678. doi:10.3758/s13428-011-0089-5

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis*. New York, NY: Springer.

Blouin, D. C., & Riopelle, A. J. (2005). On confidence intervals for within-subjects designs. *Psychological Methods, 10,* 397–412.

Bradley, D. R., Russell, R. L., & Reeve, C. P. (1996). Statistical power in complex experimental designs. *Behavior Research Methods, Instruments, & Computers, 28,* 319–326.

Campbell, J. I. D., & Thompson, V. A. (2002). More power to you: Simple power calculations for treatment effects with one degree of freedom. *Behavior Research Methods, Instrumentation, and Computers, 34,* 332–337.

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

Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist, 60,* 170–180.

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

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

Dixon, P., & O'Reilly, T. (1999). Scientific versus statistical inference. *Canadian Journal of Experimental Psychology, 53,* 133–149. doi:10.1037/h0087305

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behaviour Research Methods, 39,* 175–191. doi:10.3758/BF03193146

Fidler, F., Thomason, N., Cumming, G., Finch, S., & Leeman, J. (2004). Editors can lead researchers to confidence intervals, but can't make them think: Statistical reform lessons from medicine. *Psychological Science, 15,* 119–126. doi:10.1111/j.0963-7214.2004.01502008.x

Glover, S., & Dixon, P. (2004). Likelihood ratios: A simple and flexible statistic for empirical psychologists. *Psychonomic Bulletin & Review, 11,* 791–806. doi:10.3758/BF03195791

Hays, W. L. (1994). *Statistics* (5th ed.). Fort Worth, TX: Harcourt Brace.

Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *The American Statistician, 55,* 19–24.

Hollands, J. G., & Jarmasz, J. (2010). Revisiting confidence intervals for repeated measures designs. *Psychonomic Bulletin & Review, 17,* 135–138.

Jarmasz, J., & Hollands, J. G. (2009). Confidence intervals in repeated-measures designs: The number of observations principle. *Canadian Journal of Experimental Psychology, 63,* 124–138.

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

Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science, 6,* 299–312. doi:10.1177/1745691611406925

Levine, T. R., & Hullett, C. R. (2002). Eta squared, partial eta squared, and the misreporting of effect size in communication research. *Human Communication Research, 28,* 612–625.

Loftus, G. R., & Masson, M. E. J. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review, 1,* 476–490. doi:10.3758/BF03210951

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

Masson, M. E. J., & Loftus, G. R. (2003). Using confidence intervals for graphically based data interpretation. *Canadian Journal of Experimental Psychology, 57,* 203–220. doi:10.1037/h0087426

Pierce, C. A., Block, C. A., & Aguinis, H. (2004). Cautionary note on reporting eta-squared values from multifactor ANOVA designs. *Educational and Psychological Measurement, 64,* 916–924.

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

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

Reeve, C. P. (1986). *An algorithm for computing the doubly noncentral F C.D.F. to a specified accuracy.* [National Bureau of Standards Statistical Engineering Division Note 86–4, November]. Washington, DC: National Bureau of Standards.

Rouder, J. N., & Morey, R. D. (2005). Relational and a relational confidence intervals: A comment on Fidler, Thomason, Cumming, Finch, and Leeman (2004). *Psychological Science, 16,* 77–79. doi:10.1111/j.0956-7976.2005.00783.x

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

Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin, 57,* 416–428.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review, 14,* 779–804. doi: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,* 291–298. doi:10.1177/1745691611406923

Yuan, K., & Maxwell, S. (2005). On the post hoc power of testing mean differences. *Journal of Education and Behavioural Statistics, 30,* 141–167.