MorePower 6.0 for ANOVA with Relational Confidence Intervals and Bayesian Analysis

Jamie I. D. Campbell
Valerie A. Thompson
University of Saskatchewan

This research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada. Address correspondence to Jamie Campbell, Department of Psychology, University of Saskatchewan, 9 Campus Drive, Saskatoon, SK, Canada, S7N 5A5 (phone 306-966-6664, fax 306-966-1959, e-mail jamie.campbell@usask.ca. This version of the manuscript created January 1, 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 (2009), as well as Bayesian posterior probabilities for the null and alternative hypotheses based on formulas in Masson (2011). The program is unique in affording direct comparison of these three approaches to interpretation of ANOVA tests. It’s high numerical precision and ability to work with complex ANOVA designs could facilitate researchers’ attention to issues of statistical power, Bayesian analysis, and use of confidence intervals for data interpretation. *MorePower 6.0* is available at [https://wiki.usask.ca/pages/viewpageattachments.action?pageId=420413544](https://wiki.usask.ca/pages/viewpageattachments.action?pageId=420413544).
Despite criticism on a variety of grounds (e.g., Rozeboom, 1960; Dienes, 2011; Dixon, 2003; Dixon & O'Reilly, 1999; Masson, 2011; 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 the perception that alternative approaches are not readily accessible. In this paper, we describe a statistical calculator, MorePower 6.0, that calculates power-related statistics (sample size, effect size, and power), relational confidence intervals (CIs) for ANOVA effects, and Bayesian analysis of the null vs. alternative hypotheses. 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). 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 means (Jarmasz & Hollands, 2009; Masson & Loftus, 2003). Bayesian analysis affords direct comparison of the probabilistic evidence for the null vs. the alternative hypothesis provided by the data whereas NHST provides only a binary decision to reject or not 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.

There are many on-line power and sample size applets and also 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 RM and IM 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) or in terms of a test statistic ($F$, mean square treatment or $MST$, and $t$). For tests with one degree of freedom (including interactions), effect size may be specified or calculated as the difference in 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 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 a null hypothesis ($H_0$) to be true, usually that there is no difference between population parameters, 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 (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, power of .8 is considered to be good or adequate (Cohen, 1988), but higher levels of power are desirable if it is a high priority 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 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 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 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, *a*, 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 (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 vs. 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; i.e., if *p*(D|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 \( p(H_0|D) \). It may seem intuitive that \( p(D|H_0) \) and \( p(H_0|D) \) will be closely linked probabilities, but this is not necessarily the case (Berger, 1985; Wagenmakers, 2007). In fact, conditions under which \( p(D|H_0) \) is less than .05 and lead to rejection of \( H_0 \), can correspond to high values of \( p(H_0|D) \) that suggest that \( H_0 \) is 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 \( H_0 \) vs. \( H_1 \) 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 (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 et al., 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 interpretation unfamiliar to many (Wagenmakers, 2007). Following Wagenmakers (2007), Masson (2011) presented a straightforward approach to calculate and interpret the posterior probabilities \( p(H_0|D) \) and \( p(H_1|D) \) in the context of standard ANOVA. This approach is
Running head: MorePower for ANOVA

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 $\eta_p^2$) (see also Rouder, Speckman, Sun & Morey, 2009, for Bayesian analysis of $t$-tests with on-line 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 (Rafferty, 1995; 1999; Wagenmakers, 2007, Appendix B). The BIC is somewhat conservative with respect to providing evidence for $H_1$ compared to some other objective priors (Raftery, 1999; Wagenmakers, 2007). Raftery (1999, p. 412) notes that this conservative behaviour favors “reporting of 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, then 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 $\Delta BIC$, which is the difference in BIC values for the null and alternative hypothesis models, is given by Equation 1 for ANOVA (Masson, 2011; Wagenmakers, 2007).

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

In Equation 1, $n$ is the number of independent observations contributing to an effect, $SSE_1/SSE_0$ is the ratio of the sums of squares error 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 then \( n \) is equal to the number of subjects multiplied by the degrees of freedom associated with the RM factor(s) (Masson, 2011). MorePower calculates \( \Delta BIC \) based on two equalities pointed out by Masson (2011, p. 682). First, the ratio \( \frac{SSE_1}{SSE_0} \) equals the complement of partial eta-squared \( (1 - \eta_p^2) \). \( \eta_p^2 \) 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 \( \Delta BIC \) contrasts the null and alternative hypothesis models. These substitutions produce Equation 2 used by MorePower to calculate \( \Delta BIC \).

(2) \[
\Delta BIC = n \ln(1 - \eta_p^2) + df \cdot \ln(n)
\]

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

A fundamental advantage of the Bayesian approach is that it affords a graded comparison of \( p(H_0|D) \) and \( p(H_1|D) \), 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 \( p(H_0|D) \) from .5 to .75 may be classified as “weak evidence”, .75 to .95 as “positive evidence”, .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 interpretation of BF adapted from Jeffreys (1961) (see Table 1). These provide approximate, descriptive rules of thumb to summarize the results of Bayesian analysis.

[Table 1 about here]
Relational Confidence Intervals (CIs) for ANOVA

Use of relational CIs for ANOVA is an influential alternative or augmentation to NHST (Cumming & Finch, 2005; Fiddler, 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 (e.g., Rouder, et al., 2009; Rouder & Morey, 2005). 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 not 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 the width of the CI) multiplied by a factor of $\sqrt{Z}$ (Jarmasz & Hollands, 2009; Loftus & Masson, 1994).

For all ANOVA tests, MorePower calculates relational CIs for the effect of interest based on 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 Equation 3.
\( M_i \pm t_{\text{critical}} \cdot \sqrt{\frac{MSE}{n \cdot L / r}} \)

\( M_i \) 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 (Loftus & Masson, 1994; Jarmasz & Hollands, 2009; but see Blouin and Riopelle, 2005). \( n \) is the number of participants in each IM cell of the effect of interest, or \( n \) equals the total sample (\( N \)) if the effect has only RM factors. \( L \) 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 RM factors in the effect of interest. In \( MorePower \), \( L \) is set to 1 when there 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 to 5 in Table 4 of Jarmasz and Hollands (2009, p. 130). These apply, respectively, to 1) IM main effect (IM designs), 2) RM main effect (RM and mixed IM-RM designs), 3) RM interaction (RM and mixed IM-RM designs), 4) IM main effect (mixed IM-RM designs), and 5) 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. MorePower was developed under Microsoft Windows but 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 require that the relevant statistical assumptions for ANOVA are satisfied (normal distributions, sphericity, homogeneity of variances and covariances; see, for example, Hays, 1994). Probabilities for the non-central $F$ 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 Figure 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 that presents a detailed summary (described later) about the last calculation.

The radio buttons in the Analysis section 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 specify the ANOVA design and effect (see 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 partial $\eta^2$ 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 section allows the user to specify (or request) effect size in terms of a test statistic, including the $F$-ratio, mean square treatment ($MST$), 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 for Variability is used for the calculation of the mean difference in original units and its variability (see Equations 4 and 5 discussed later).

The Solve For section toggles 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
Running head: MorePower for ANOVA

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 initializes all 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 current version information and contact information.

Specification of ANOVA Design Factors and Effect of Interest

The overall design of the experiment is specified using the IM and RM fields in the Design Factors section (see Figure 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 RM factors each with two levels, enter 22 for IM and 22 for RM. For a $2 \times 2 \times 3$ 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 IM and RM fields (a factor can have a maximum of 9 levels). Once the design factors have been entered, specify the effect using the corresponding IM and RM 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 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 lower Effect Size field (F, MST, t) may be selected. The default is partial eta-squared. Denoted η$^2_p$, it is based on the sums of squares for the effect of interest: SS$^{treatment}$ / (SS$^{treatment}$ + SS$^{error}$). η$^2_p$ 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$^2_p$ is the amount of unique variance explained by the effect of interest divided by the proportion of variance unexplained by all the other main and interaction effects in the design (Levine & Hullett, 2002). As a measure of ANOVA effect size, η$^2_p$ is frequently used by psychological researchers and it is the effect-size measure reported in example manuscripts presented in the American Psychological Association publication manual 6th edition (2010, e.g., p. 46). The General Linear Model (GLM) procedure for repeated-measures ANOVA in IBM SPSS 19.0 also computes η$^2_p$ to represent the observed effect size. Thus, η$^2_p$ is a common effect size measure for ANOVA. Cohen’s $f$, another common effect measure for ANOVA, can be derived from η$^2_p$ by the relation $f = \sqrt{\eta^2_p / (1 - \eta^2_p)}$ (see Cohen, 1988, p. 281). *MorePower* includes the conversion of η$^2_p$ 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$^2_p$ 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 original units (e.g., 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., sample size required to detect a 10 ms difference with power of .8).

Selecting the upper radio button in the Effect Size section and clicking difference enables this option. This can be used for any \( 2^k \) effect in the design where \( k \) is the number of 2-level factors in the effect of interest. Any such effect will have numerator \( df = 1 \) and may consist of RM, IM, or combined RM \( \times \) 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 one is interested in. For a 2-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 \( \times \) 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 \( \times \) 2 \( \times \) 2 interaction it is the mean difference between the mean difference 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, then \( d \) for the 2 \( \times \) 2 interaction would equal \( |60 - 30| - |20 - 10| \) or 20.

In MorePower, \( d \) is calculated using Equation 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-subject) 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 the IM factor \((n = N/2)\); if the effect includes two IM factors (e.g., a \( 2 \times 2 \times 2 \) interaction involving two IM and one RM factor), then \( n \) is the number of observations in each of the four IM cells \((n = N/4)\), etc. The value of \( n \) is derived by MorePower from the specification of the design and effect and from the total \( N \) entered in the Sample field.

\[
d = \sqrt{\left( \frac{\text{MST} \cdot 2^B \cdot 2^W}{L} \right) / n}
\]

The \( \text{MST} \) value used by MorePower to calculate \( d \) is based on the relation \( \text{MST} = \text{MSE} \times F \). \( F \) may be input directly if the lower radio button in the Effect Size section is selected, but if the upper radio button in the Effect Size section is selected, \( F \) is calculated from \( \text{eta}^2 \) and the \( \text{MSE} \) value specified box in the Variability section. Instead of \( F \), however, the user can alternatively select \( \text{MST} \) or \( t \). The \( t \) option reflects the fact that \( F \) with numerator \( df = 1 \) is equal to \( t^2 \). Similarly, instead of \( \text{MSE} \), 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 Equation 5 derived by Campbell and Thompson (2002). This equation calculates the variance of the difference \( (s_d^2) \) from the \( \text{MSE} \) 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 \( s_d/\sqrt{n} \)

\[
s_d^2 = \text{MSE} \cdot 2^B \cdot \frac{2^W}{L}
\]
Contents of the Output Window for ANOVA calculations

Figure 1 shows the MorePower interface after running built-in Example 1, which calculates the sample size required (with power = .8) for a $2 \times 2 \times 2$ interaction in a 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 partial $\eta^2$ and Cohen’s $f$. The following two lines contain output associated with the Bayesian analysis, including $\Delta$BIC (dBIC in the display), the estimated Bayes Factor (BF), and estimated posterior probabilities $p_{BIC}(H_0|D)$ and $p_{BIC}(H_1|D)$. The next two lines report information about the Jarmasz and Holland (2009) CI for this interaction. Examples related to the Bayesian analysis and CI functions in MorePower are discussed in subsequent sections. As the $2 \times 2 \times 2$ interaction has $df$ treatment $= 1$, the mean difference ($d$) in original units (1.003) is presented based on Equation 4. This number represents the magnitude of the interaction effect in original units. The $SE$ of the difference ($d$) based on Equation 5 is included and also the 95% [i.e., 100(1 - $\alpha$)] confidence interval 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, 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, $\alpha$, and power. Alpha is set by default to .05 two-sides 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 by the number of IM cells to obtain the total sample size ($N$). Calculation of sample size for ANOVA has a maximum limit of $n = 2500$ per IM cell or total $N = 2500$ for an all RM design. This largely avoids long waits (e.g., 30 seconds) owing to slow processing by CDFDNF. If the limit is reached there is a warning 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 and 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 Figure 1) illustrates use of MorePower to calculate the sample size required for a $2 \times 2 \times 2$ 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 required sample size. This is because error variability is implicit in $\eta^2$ [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
Running head: MorePower for ANOVA

correctly reject the null hypothesis (i.e., power = .8) given a medium effect size (partial $\eta^2 = .06$) for the $2 \times 2 \times 2$ mixed RM-IM interaction.

Built-in Examples 2 and 3 specify a five-factor mixed RM-IM design with $2 \times 2$ RM (within-subjects) factors combined with $2 \times 2 \times 3$ IM (between-subjects) factors. In Example 2, MorePower calculates the sample size required with partial $\eta^2 = .06$ and power $\geq .8$ for a IM-RM $2 \times 2$ interaction (required $N \geq 144$). Example 3 illustrates a sensitivity analysis and calculates the minimum population effect size required (partial $\eta^2 > .144$) for the five-way interaction given power = .8 and sample size = 72. Example 4 illustrates a $3 \times 3$ RM design and calculates the sample size required ($N \geq 26$) for a large effect size interaction (partial $\eta^2 = .14$) 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 (3-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 Table 1 of Masson (2011) ($F = 12.9$, 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 $\Delta BIC = -14.21$, $BF = 0.000859$, $p_{BIC}(H_0|D) = .00086$, and $p_{BIC}(H_1|D) = .9991$. The corresponding quantities calculated by MorePower are $\Delta BIC = -14.10$, $BF = 0.00087$, $p_{BIC}(H_0|D) = .9999$, and $p_{BIC}(H_1|D) = .9991$. 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 favour of the alternative hypothesis.

Bayesian Examples 6 and 7 refer to the $2 \times 2$ RM ANOVA reported in Table 3 of Masson (2011, p. 685). Example 6 is the main effect of Alignment for which Masson reports $\Delta BIC = -3.15$, $BF = 0.2070$, $p_{BIC}(H_0|D) = .171$, and $p_{BIC}(H_1|D) = .829$. This provides positive evidence for $H_1$. Example 7 is the $2 \times 2$ (Alignment $\times$ Delay) interaction for which Masson reports $\Delta BIC = -10.44$, $BF = 0.0054$, $p_{BIC}(H_0|D) = .005$, and $p_{BIC}(H_1|D) = .995$. This indicates “decisive” evidence for $H_1$ (Table 1). Finally, Example 8 refers to the Bayesian analysis of a $2 \times 2$ IM interaction reported in Table 4 of Wagenmakers (2007, p. 799) who reports (p. 799) $\Delta BIC = -4.31$, $p_{BIC}(H_0|D) \approx .10$, and $p_{BIC}(H_1|D) \approx .90$. 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 to 5 in Jarmasz and Hollands (2009, Table 4). The confidence level (e.g., 99%, 95%, 90%) for the CI is controlled by setting alpha (e.g., .01, .05, .10), 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 $4 \times 2 \times 2 \times 2$ design. Figure 3 shows MorePower after running Example 9. The Design Factors’ RM field contains 4
Running head: *MorePower* for ANOVA

× 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 and Hollands, 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 and Hollands, Table 5, p. 131).

[Figure 3 about here]

The output window (see Figure 3) presents information as discussed previously, but we will focus here on the Jarmasz and Hollands CI, which is reported as J&H 95% ±4.07, with a critical value of $t(20) = 2.086$, $MSE = 1462$, $n = 24$, $L = 32$ (i.e., $4 \times 2 \times 2 \times 2$) and $r = 2$ (i.e., the effect of interest has two RM cells), #obs (i.e., number of observations) = 384. These results correspond to the solution presented by Jarmasz and Hollands (2009, p. 129) as follows.

$$M \pm 2.086 \cdot \frac{1462}{\sqrt{24 \cdot 32 / 2}}$$

$$M \pm 4.1$$

As this effect has one degree of freedom in the numerator, *MorePower* also displays the mean difference in original units (9.48% based on Equation 4), the standard error of the difference (2.76 based on Equation 5), and the 95% confidence interval of the difference (±5.76%). The difference between the CI constructed using Equation 3 (±4.07%) and the 95% confidence interval 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 $\alpha$ 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% confidence interval.

ANOVA Examples 10 to 12 further demonstrate the J&H CI function in *MorePower* using Jarmasz and Hollands’ (2009) illustrative analyses. Example 10 corresponds to the CI for the $4 \times 2$ RM effect ($\pm3.4$ on p. 131). Example 11 is the CI to compare RM conditions within IM cells for the $4 \times 2$ mixed IM-RM effect ($\pm3.2$ on p. 133). Example 12 refers to the CI for the main effect of the 4-level IM factor ($\pm6.3$ on p. 134). This experimental design included only one IM factor and therefore did not afford a demonstration for interactions in multi-factor IM designs with no RM factors. Example 13 confirms the application of *MorePower*’s CI function for the case of a $2 \times 2$ IM interaction ($n = 12$ per group) taken from Masson and Loftus (2003, CI $\pm0.055$ 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*<sub>error</sub> are the same for an RM effect and 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 & Holland, 2009, p. 128-129; Formula 6 on p. 130). One can enter the $MS_wc$ rather than *MSE* in
Running head: MorePower for ANOVA

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 \( df(\text{MS}_{\text{wc}}) = df(\text{MSE}) \). This will rarely be the case. Nonetheless, MorePower can be used to compute the number of observations (i.e., \( n \cdot L/r \) in Equation 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., copy an ANOVA description as a template for a similar analysis, but fail to make all the necessary changes). MorePower allows the user to quickly specify an ANOVA effect of practically arbitrary complexity and then review the \( p \)-value, degrees of freedom, effect size (\( \eta_{p}^2 \) or original units for one degree of freedom tests), etc. 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 to report effect size measures such as \( \eta_{p}^2 \), 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., \( \text{MSE} \) or \( \text{SE} \)) 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. Follow the instructions or click the Clear Values button and retry. If a persistent error occurs, please email jamie.campbell@usask.ca with the error number and a screen shot 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 Jarmsz and Hollands (2009) formulas. A unique feature of MorePower is that it affords direct comparison of these three approaches to interpreting ANOVA. Other functions include power analyses for one or two sample t-tests or proportions, and simple correlation (see Campbell & Thompson, 2002, and the supplementary document included in the calculator download), and it is a probability calculator for the $F$, $t$, and $z$ distributions. Its high numerical precision for ANOVA and ability to work with complex ANOVA designs could further facilitate researchers’ attention to issues of statistical power, use of relational confidence intervals 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 (e.g., Dixon, 2003; Glover and Dixon, 2004; Rouder et al., 2009; see also Wagenmakers, 2007, p. 795).
References


Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to


Running head: *MorePower* for ANOVA


Reeve, C. P. (1986). *An algorithm for computing the doubly noncentral F C.D.F. to a specified*
Running head: MorePower for ANOVA

_accuracy_. National Bureau of Standards Statistical Engineering Division Note 86-4, November.


Footnotes

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).

2. The focus here is on ANOVA, but MorePower 6.0 includes power calculations for \( t \)-test of means and simple correlation and \( z \)-test 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.

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 (\( d \)). This is another common measure of effect size (e.g., see one-sample \( t \)-tests of means in G*Power 3; Faul et al., 2007).

4. It is important to distinguish \( \eta_p^2 \) from \( \eta^2 \), 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 \( \eta_p^2 \) is larger than \( \eta^2 \) in multi-factor designs (Levine & Hullett 2002; Pierce, Block & Aguinis, 2004).

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 is unaffected. The \( MSE = .014 \) used in Example 5 is from Table 1 in Masson (2011).
Table 1. Evidence Categories for Bayes Factor (adapted from Wetzels et al., 2011, p. 293)

<table>
<thead>
<tr>
<th>Bayes Factor</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;100</td>
<td>Decisive evidence for $H_0$</td>
</tr>
<tr>
<td>30-100</td>
<td>Very strong evidence for $H_0$</td>
</tr>
<tr>
<td>10-30</td>
<td>Strong evidence for $H_0$</td>
</tr>
<tr>
<td>3-10</td>
<td>Substantial evidence for $H_0$</td>
</tr>
<tr>
<td>1-3</td>
<td>Anecdotal evidence for $H_0$</td>
</tr>
<tr>
<td>1</td>
<td>No evidence</td>
</tr>
<tr>
<td>.333-1</td>
<td>Anecdotal evidence for $H_1$</td>
</tr>
<tr>
<td>.1-.333</td>
<td>Substantial evidence for $H_1$</td>
</tr>
<tr>
<td>.0333-.1</td>
<td>Strong evidence for $H_1$</td>
</tr>
<tr>
<td>.01-.0333</td>
<td>Very strong evidence for $H_1$</td>
</tr>
<tr>
<td>&lt;.01</td>
<td>Decisive evidence for $H_1$</td>
</tr>
</tbody>
</table>

*Note.* Following Masson (2011) and Wagenmakers (2007), BF 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 $H_1$ are opposite to those in Table 1 (i.e., substitute $H_0$ for $H_1$).
Figure 1. *MorePower 6.0* interface and output window (ANOVA Example 1).
Figure 2. ANOVA Example 5 refers to the Bayesian analysis of a three-level RM ANOVA (based on Table 1 in Masson, 2011).
Figure 3. ANOVA Example 9 refers to the relational confidence interval calculations for a repeated-measures main effect in a mixed IM-RM design (Jarmasz and Hollands, 2009, p. 129).