Techniques and Solutions for Sample Size Determination in Psychology: A Critical Review

August 6, 2019

Christopher L. Aberson
Department of Psychology, Humboldt State University

Dries H. Bostyn
Department of Developmental, Personality and Social Psychology, Ghent University

Tom Carpenter
Department of Psychology, Seattle Pacific University

Beverly G. Conrique
Department of Psychology, University of Pittsburgh

Roger Giner-Sorolla
School of Psychology, University of Kent

Neil A. Lewis, Jr.
Department of Communication, Cornell University & Division of General Internal Medicine, Weill Cornell Medical College

Amanda K. Montoya
Department of Psychology, University of California - Los Angeles

Brandon W. Ng
Department of Psychological and Brain Sciences, Texas A&M University

Alan Reifman
Department of Human Development and Family Studies, Texas Tech University

Alexander M. Schoemann
Department of Psychology, East Carolina University

Courtney Soderberg
Center for Open Science

Author Note: The above authors, listed in alphabetical order, made equivalent contributions to writing this supplement and/or critical background material in the preprint article it references. As with that preprint, this work was facilitated by the Society of Personality and Social Psychology, who helped us organize as the Power Analysis Working Group at their 2019 meeting.

Techniques and Solutions for Sample Size Determination in Psychology: A Critical Review

This article is a supplement to a preprint manuscript by the same authors: Giner-Sorolla, Aberson, Bostyn, Carpenter … & Soderberg (2019). It starts from the assumption that readers know the basic premises and terminology of a number of commonly used statistical tests in psychology, as well as the basics of power analysis and other ways to determine and evaluate sample size. It seeks to give further guidance into software approaches to sample size determination for these tests, via precision analysis, optional stopping techniques, or power analysis of specific inferential tests. Further information on the first two methods, and on power analysis in general, can be found in the Giner-Sorolla et. al. (2019) article. This critical review seeks to define best practice in light of the strengths and weaknesses of each software product.

Specific Techniques for Precision Analysis

For many simple statistics (e.g. regression coefficients, standardized mean differences) the sample size needed for the AIPE approach can be computed analytically (Kelley & Maxwell, 2003; Kelley & Rausch, 2006). In these cases, the equation *desired width = criterion*standard error* can be solved for *N,* which is part of *standard error*. Analytic methods using AIPE can be found in the *MBESS* (Kelley, 2007) package in *R*. For more complex designs or when an interval estimate may not be computed analytically (e.g. bootstrapping), Monte Carlo simulations can be used (Beaujean, 2014).

Specific Techniques for Optional Stopping

For all procedures listed below, broadly known as sequential sampling rules (SSR), the false positive rate is only controlled at the nominal level if the procedures are planned before results have been observed. For this reason, we strongly encourage pre-registering sample collection and termination plans.[1]

One set of methods involves setting a lower and upper bound on *p*-values. A study is run collecting several cases at a time. After each collection, the study is stopped if the observed *p*-value is below the lower bound, or above the upper bound. Otherwise, collection continues. A number of different SSR methods have been developed for different statistical tests and minimum and maximum *N*s, including the COAST method (Frick, 1998), the CLAST method (Botella, Ximenez, Revuelta, & Suero, 2006), variable criteria sequential stopping rule (Fitts, 2010a; Fitts 2010b), and others (Ximenez & Revuelta, 2007).

Another set of techniques is group sequential analyses. In these designs, researchers set only a lower p-value bound and a maximum *N*, and stop the study early if the p-value at an interim analysis falls below the boundary. To keep the overall alpha level at the prespecified level, the total alpha is portioned out across the interim analyses, using one of a number of different boundary equations or spending functions (see Lakens, 2014; Lakens & Evers, 2014).

The alpha boundaries for these sequential designs can be calculated using a number of different programs, including the GroupSeq package in R or the WinDL software by Reboussin, DeMets, Kim, and Lan (2000). Tutorials on how to use both sets of software can be found at https://osf.io/uygrs/ (Lakens, 2016). The packages allow for the use of a number of different boundary formulas or alpha-spending functions.

[1] There is one optional stopping technique, p-augmented (Sagarin, Ambler, & Lee, 2014), that researchers can decide to implement after seeing a null result. However, this technique does not keep the false positive rate at .05. Instead it allows the researcher to calculate the true p-value of the final sample, given that a data-dependent increase in the sample size was made following an initial null result. This p-augmented value will always be more than .05, but if only one sample size increase was made will generally be under .1. Therefore, the technique allows researchers some flexibility in sample size while being transparent about the amount of potential false positive inflation this flexibility caused.

Types of Technique for Power Analysis

Effect size metrics. Power analysis, as we have noted, involves three different approaches, which either require or output effect size as a parameter. Effect size specification is thus critical for conducting or interpreting power analyses. The two most prominent approaches to effect size have come from Cohen (1988) and Rosenthal (e.g., Rosenthal, Rosnow, & Rubin, 2000). Cohen defined a plethora of effect size estimates depending on the statistical test design, using different Greek and Roman letters, whereas Rosenthal sought to express effects in the common metric of the correlation coefficient *r*. This document largely focuses on estimates consistent with Cohen, as these appear to be more commonly used in psychology publishing, and by analytic programs such as SPSS and G*Power.

Programs like G*Power rely on values such as Cohen's *d* for mean comparisons (i.e., t-tests), *r* for tests of correlations, and phi (defined as *w* in some sources) for chi-square. Estimates for multiple regression, ANOVA, and more advanced approaches often focus on estimates addressing proportion of explained variance, including R^2, η^2, partial η^2, and the squared semi-partial correlation (sr^2). Sensitivity analyses for many approaches provide effect sizes in terms of f or $f^{\,2}$ which are not commonly reported and may be better understood after converting to more prevalent metrics (e.g., *d*, *r*, R^2). Effect size converters can be found online (e.g., the implementation of Lin, 2019 at http://escal.site/).

Algorithmic approaches. Power estimation using an algorithmic approach, also known as "analytic," calculates a power function based on known parameters. Algorithmic analyses involve central and non-central distributions and a non-centrality parameter (NCP).

Common central distributions are t, F, and χ^2. The shape of these distributions are a function of degrees of freedom. Importantly, central distributions reflect the null hypothesis and decisions about whether or not to reject the null. Non-central distributions are distributions with shapes that vary based on both degrees of freedom and effect size. These distributions define the alternative distribution (i.e., the distribution reflecting the specified population effect size).

The relationship between central and non-central distributions determines power, and is quantified by the NCP. One simple way to think about the NCP (for two independent groups) is as the distance between the centers of the two distributions (i.e., how far the alternative distribution is from the null). The NCP allows for determination of how much of the alternative distribution corresponds to failing to reject (Beta error) and rejecting the null decisions (power), by calculating areas under curves. More broadly, the NCP is a function of effect size and sample size. Larger effect sizes and larger sample size make larger NCP values. Larger NCP values correspond to more power. Figure A1 demonstrates the influence of effect size and sample size on the NCP.

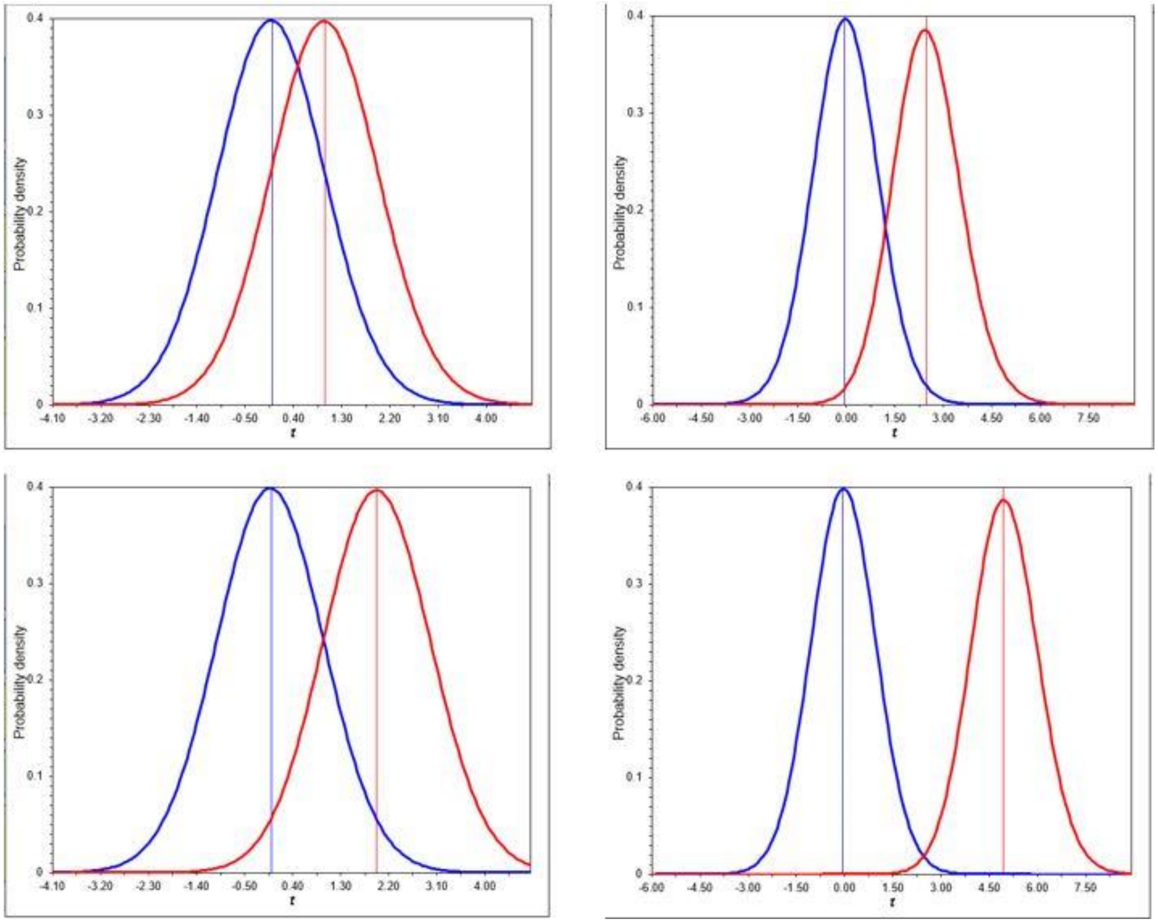


Figure A1. Visual representation of influence of effect size and sample size on noncentrality parameters. The center of each distribution on the right is the NCP. Top left panel, n = 50 per group, d = 0.20 yields δ =1.0. Top right panel, n = 50 per group, d = 0.50 yields δ =2.5. Bottom left panel, n = 200 per group, d = 0.20 yields δ =2.0. Bottom right panel, n = 200 per group, d = 0.50 yields δ =5.0.

Simulation approaches. Another approach to power analysis is Monte Carlo or simulation-based. This method involves specifying population effect size(s), sample size (n), and Type I error rate as before. Instead of determining relationships between central and noncentral distributions, simulations generate a population with the specified effect size parameter and then draw random samples (usually 1000s) of size n. After drawing samples, we run the analysis of interest on each sample and tally the proportion of results that allowed for rejecting the null hypothesis. This proportion constitutes power.

This procedure differs from the classic approach as it addresses the samples that actually allowed for rejection of the null rather than relying on assumptions required for the central and noncentral distributions. For simpler analyses (e.g., t-tests, ANOVA, correlation, chi-square) traditional and simulation approaches generally produce indistinguishable results. However, simulation approaches are often the most effective way to address analyses involving complex statistical models and situations where data are not expected to meet distribution assumptions. Details of simulation methods are outside the scope of the present paper but interested readers should see the paramtest (Hughes, 2017), simr (Green & MacLeod, 2016), simDesign (Sigal & Chalmers, 2016), and MonteCarlo (Leschinski, 2019) packages for R .

Power Analysis: Best Practices and Resources for the Most Commonly Used Tests

In the remainder of this article we will, one by one, cover power-analytic techniques pertaining to the most commonly used statistical tests in psychological research, including special considerations for using the popular application G*Power (Faul, Erdfelder, Lang & Buchner, 2007). Our list also might help guide developers of sample-size-determination tools to strategically fill the gaps in our coverage.

Simple correlation tests. The linear association between two ordered numerical variables is most commonly assessed using the Pearson correlation coefficient, represented by r in samples and rho (ρ) in populations. Power calculations for correlation tests are readily available in most power calculation software and use rho as an effect size. In G*Power, a test for the power of rho’s difference from zero is available under the “exact test” family (not the “point biserial” option, which is more obvious in the menu system but refers to the correlation of an ordered with a dichotomous variable).

To help show how power depends on effect size using a relatively simple statistical example, power curves for correlation tests with sample sizes ranging from 0 to 200 are displayed for various rho in Figure A2.

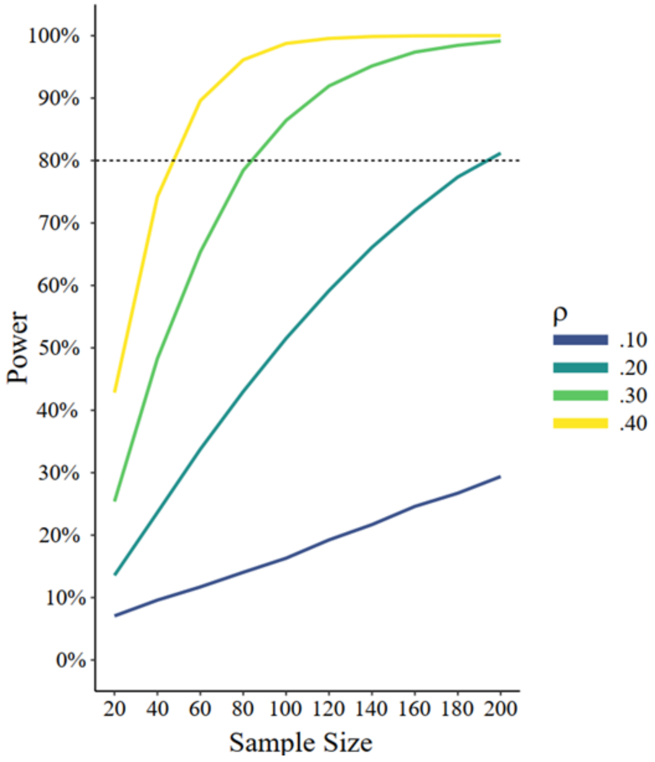


Figure A2. Power curves for a simple correlation test.

χ^2 and tests of Proportions. Chi-squared (χ^2) tests evaluate the likelihood that observed data--such as categorical frequencies, contingency tables, or coefficients from a model test--could have been produced under a certain null hypothesis--such as equal distribution of proportions, zero contingency between categories, or perfect fit to a model[2]. Power calculations for the χ^2 test family are provided in G*Power.

There are many possible effect sizes for these kinds of data (e.g., proportions, odds ratios, risk ratios, etc.); G*Power uses the effect size measure *w*, and supplies a tool for calculating *w* for any set of proportions, including multidimensional contingency tables. In a 2 × 2 contingency table, *w* is equal to the ϕ (phi) correlation (Cohen, 1988) and can be interpreted as a correlation. As *w* is often not reported in empirical manuscripts, reviewers can quickly calculate its value with $w = \sqrt{\frac{\chi^2}{n}}$.

Multiple Regression

Multiple regression is a technique using ordered, continuous variables that assesses the overall strength of association of a set of independent variables with a dependent variable (*R^2 model*), the increase in such strength as new independent variables are added (*R^2 change*), and the contribution of each variable to predicting the dependent variable adjusting for intercorrelation with the others (regression coefficients). This section covers G*Power approaches under the following options:

- Linear Multiple Regression: Fixed Model, R^2 deviation from zero (*R^2 model),*
- Linear Multiple Regression: Fixed Model, R^2 increase (*R^2 change),*
- Linear Multiple Regression: Fixed Model, single regression coefficient (coefficient power)[3].

[2] For simple proportions, a binomial *z*-test or Fisher’s exact test can also be conducted. Discussions of these tests can be found in Howell (2008); power for these tests can be computed in G*Power under the. z test family.

[3] G*Power also provides analyses focused on specification of slopes. To use this approach, setting both standard deviations at 1.0 provides an estimate of sensitivity for a standardized regression coefficient.

Additional topics include estimation of power for multiple coefficients simultaneously, and power to detect all effects in a models. Going beyond G*Power will be necessary for some of these questions.

Power analyses for R^2 *model* and R^2 *change* use the effect size estimate f^2. Typically, researchers present R^2 values for these tests, so converting the estimate is useful. For coefficients, the f^2 value can be converted to a squared semi-partial correlation (sr^2) for the predictor of interest. This statistic reflects the proportion of variance uniquely explained by the predictor (analogous to eta-squared in ANOVA). Researchers commonly report standardized regression coefficients (a.k.a., beta coefficients or slopes) in lieu of this effect size measure. Although they bear some relation to effect size, standardized regression coefficients do not correspond directly to proportions of explained variance.

To show the differences between each type of power, the following example using G*Power starts from three predictors, a sample size of 100, Power = .80, and alpha = .01. For the test of R^2 *model*, sensitivity analysis yields 80% power for a population with $f^2 > .241$ (equivalent to R^2 *model* > .194). For R^2 *change* (with two predictors entered in the final model), $f^2 > .217$ (equivalent to R^2 *change* > .178). The coefficient test can detect $f^2 > .184$ (equivalent to $sr^2 > .155$). Although at first blush it appears that tests of coefficients are the most powerful, being sensitive to smaller effects, this is generally not the case. Coefficients test how much variance the predictor explains over and above all of the other predictors, so these values will tend to be much smaller in relation to model and change values, because they exclude shared variance.

Special issue: Influence of multicollinearity and power to detect multiple effects. The structure of G*Power's protocols may not match research designs employing multiple regression to estimate multiple individual coefficients. For example, a study might have three predictors, with hypotheses indicating that each relates to the dependent measure. Often, the researcher's interest is detecting significant effects for all three predictors. Accurate power estimates require specifying a model with all three predictors simultaneously, including a full correlation matrix to appropriately account for multicollinearity (i.e., the correlations between predictors).

Because of issues with multicollinearity, sensitivity approaches for these designs may lead to uninformed conclusions regarding power. For example, imagine a sensitivity analysis indicates that a sample n = 184 in a two predictor design yields 80% power to detect effects as small as f^2 = .043 (sr^2 = .040). This target effect size could be produced in a number of ways. If both predictors are uncorrelated with each other (no collinearity) they only need to be correlated zero-order with the DV at r = .20. But if the correlation between predictors is large (e.g., r = .775; high collinearity) it would require zero-order correlations of r = .50 between each predictor and the DV to achieve effect sizes approaching f^2 = .043.

Extending this to designs with more than two predictors, stronger correlations among predictors (multicollinearity, i.e., the square root of 1-tolerance) as well as stronger correlations between all the remaining predictors and the dependent variable, both reduce the size of sr^2. That is, adding more valid predictors correlated with existing ones will usually reduce sr^2 substantially.

Another issue is detecting power for all coefficients in a single study. Power to detect all effects in the same study differs substantially from power to detect any single effect. A study might have 80% power to detect any one coefficient but power to detect all coefficients at once will be lower. The lack of attention to this form of power -- called Power(All) in this paper — is a likely source of underpowered research in the behavioral sciences (Maxwell, 2004). Power(All) is a function of the power for each individual test, correcting for multicollinearity. Under most conditions, more power for individual tests increases Power(All) whereas multicollinearity decreases it (for a detailed discussion see Aberson, 2019).

Given these complexities, researchers studying multiple predictors in regression should consider a priori approaches that account for multicollinearity. The R package pwr2ppl (https://github.com/chrisaberson/pwr2ppl) provides code that takes correlations and sample size as input and returns power. Two approaches are demonstrated, the first establishes power for each predictor in model (individual coefficient power). The second addresses power to detect all of the effects in the same model [Power(All)].

This example uses the following correlations, r_{y1} = .35, r_{y2} = .40, r_{y3} = .40, r_{12}=.50, r_{13}=.30, and r_{23} = .40. Correlations with y are predictor-outcome correlations (e.g., r_{y1}) and numeric subscripts note correlations of predictors with each other. Using the MRC function demonstrates that with a sample of 150 participants power for R^2 *Model* is substantial (1.0) however, power for coefficients ranges from .499 to .918. Although this study would likely find significant effects for R^2 *Model* and the third coefficient, power is relatively low for detecting the other coefficients.

MRC(r_{y1}=.35, r_{y2}=.40, r_{y3} = .40, r_{12}=.50, r_{13}=.30, r_{23} = .40, n=150, alpha=.05)

```
[1] “Power R2 =  1”
[1] “Power b1 =  0.499”
[1] “Power b2 =  0.675”
[1] “Power b3 =  0.918”
```

Although the lowest powered predictor provides 50% power to detect effects, this does not mean that we will find significant effects for all of the predictors 50% of the time given these correlations. The MRC_all function provides estimates of power to detect all three predictors within the same model. Note that Power(All) is .241, substantially lower than power for the lowest powered predictor (.499). If researchers want to find support for all predictions, then power for detecting all of the effects in the same model is the most relevant approach.

MRC_all(r_{y1}=.35, r_{y2}=.40, r_{y3} = .40, r_{12}=.50, r_{13}=.30, r_{23} = .40, n=150, alpha=.05)

```
[1] “Sample size is  150”
[1] “Proportion Rejecting None =  0.0014”
[1] “Proportion Rejecting One =  0.1443”
[1] “Proportion Rejecting Two =  0.6135”
[1] “Power ALL (Proportion Rejecting All) =  0.2408”
```

ANOVA

Analysis of variance, or ANOVA, assesses whether one or more categorical, nominal independent variables relate to differences in a continuous dependent variable. Factorial ANOVA tests main effects and interactions between multiple independent variables; repeated measures ANOVA tests within-person differences among dependent variable means. Power calculations for ANOVA present a number of challenges for researchers using G*Power as of this writing.

Biased effect size estimates. The most commonly reported effect sizes for ANOVAs are eta-squared (η^2) and partial eta-squared (η_p^2). For example, η_p^2 is the standard effect size output in SPSS for ANOVA and can be input to G*Power, which will convert it to the effect size f. Unfortunately, eta-squared is upwardly biased in small sample sizes (Okada, 2013), leading to sample size estimates that are too small to achieve the desired level of power. Due to this bias, Okada (2013) recommends the use of either omega squared (ω^2) or epsilon squared (ε^2) for ANOVAs; see Appendix for formulas. These effect sizes are conceptually similar to η^2 but show less bias. For factorial designs in which all variables are manipulated, the partial versions of these effect sizes are preferred (Olejnik & Algina, 2003), similar to the use of η_p^2 in factor ANOVAs. If researchers are attempting to power a between subjects design or one in which all variables are manipulated, based on an original study which included measured variables or within-subjects factors, generalized ω^2 is preferred over partial ω^2 because the $\omega_p{}^2$ will overestimate the effect, as would $\eta_p{}^2$ (Maxwell & Delaney, 2004; Olejnik & Algina, 2003) That is, generalized ω^2 can be used to compare effect sizes across within- and between-subjects designs (Bakeman, 2005).

For power analyses, ω^2 or ε^2 can be used directly in place of η^2 . These effect sizes and their confidence intervals can also be calculated using the MOTE R package (Buchanan, Gillenwaters, Scofield, & Valentine, 2019) or the accompanying Shiny App, https://doomlab.shinyapps.io/mote/ (Buchanan, Gillenwaters, Padfield, Van Nuland, & Wikowsky, 2019).

In some cases, ω calculations require information from the full F-table that is not generally reported in papers, so researchers may only have access to η^2 or $\eta_p{}^2$. In these cases, we suggest researchers adjust the observed effect size downward to counteract the upward bias of the effect size. To help determine how much they may need to downward adjust the effect size, researchers can investigate the table and accompanying R code from Lakens' blog post (Lakens, 2015).

Within-subjects G*Power inputs require special care. G*power as of this writing, as admitted in the online manual, has no documentation for its within-subjects and mixed-ANOVA calculation methods (G*Power 3.1 Manual, March 1, 2017). At a glance, there are two ways in which these techniques are often used incorrectly based on the menu interface. The first, covered extensively in Lakens (2013), is that the default η^2 or $\eta_p{}^2$ that G*Power expects for repeated subjects factors does not include the correlation between the repeated factors in the effect size calculation, but psychologists nearly always report an η^2 or $\eta_p{}^2$ from these designs that already takes this correlation into account. If researchers input this effect size into the default repeated measures ANOVA interface, and then also input a ‘correlation among repeated measures’, the correlation is double-counted, leading to a required sample size that is too small. When calculating power for repeated measures factors in G*power, researchers need to go into the Options menu and choose the ‘as in SPSS’ effect size specification.

The second issue is that for mixed and repeated measures ANOVAs, G*power assumes the designs only have one between subjects and/or one within subject factor, but this is not clear from the input fields. For a 2x2 within subjects ANOVA, given that each participant provides four data points, many researchers may assume that the ‘number of measurements’ in G*power should equal four. However, this input wrongly results in an output parameter numerator df of 3, which is appropriate for a one-way within subject design with four levels, but not a 2x2 design, which would have 1 as its numerator degrees of freedom (df). The wrong numerator df will lead to wrong sample size requirements. A similar problem occurs for the ‘number of groups’ input for the ‘Repeated measures, within-between interaction’ power option. However, this does not occur in the ‘Fixed effects, special, main effects and interactions’ because researchers can input the numerator df and the number of groups separately, which leads to correctly calculated sample size. Until the manual and/or G*Power is updated, for interactions exclusively among within-subjects factors andfor mixed interactions involving only one, two-level between-subjects factor, we recommend entering “number of measurements” equal to the number of numerator df in the desired analysis, plus one. But in calculating power for a mixed-ANOVA with multiple between- subjects factors, or with a single between-subjects factor that has more than 2 levels, researchers will need to use simulation-based methods or the alternative tools listed below.

Repeated measures designs have the added complexity of a correlation structure that must be specified. Some sample size calculation methods, such as the GLIMMPSE tool or simulation methods, allow for researchers to input different expected correlations between different cells in their design. For tools such as G*Power which require a single universal correlation between DVs to be specified, if there are more than two DVs, there is no clear standard on how to derive this from prior data or theory. One option would be to use the lowest expected correlation, as that would result in the most conservative power estimate/sample size estimate. A more generous option would use the median correlation. If correlations among DVs vary greatly, though, it is most advisable to use a method that allows direct input of all expected correlations.

Recently, there have also been calls to analyze some types of repeated-measures data as multilevel models to more accurately account for variation when subjects respond to multiple stimuli (Judd, Westfall, & Kenny, 2012; Judd, Westfall, & Kenny, 2017). For power analysis considerations for these types of models, researchers can look to the multilevel model section below, or investigate the Pangea Shiny App (Westfall, 2016a).

Interactions: Factorial designs are extremely common in psychology, but present some special problems in power analysis. To begin with, many researchers have learned to think of between-subjects factorial designs in terms of a desirable *n* per cell, so that larger designs need numbers that are multiples of smaller designs. They then may be surprised, when conducting power analysis, to find out that for a given effect size, the required sample size for a two-level, one-factor between subjects ANOVA, a 2x2 between subjects ANOVA, and a 2x2x2 between subjects ANOVA are all the same because the numerator degrees of freedom of all the tests are the same (Westfall, 2015a; Collins, Dziak, & Li, 2009). If you put these designs into G*power with a η_p^2 = .059, it will correctly report that all designs need a total sample of 128 subjects to achieve 80% power.

While this observation is technically correct, other considerations make it advisable to recruit more participants for testing interactions in larger designs. First, sample size for interactions is often calculated too generously because the intuition of what effect sizes will be is incorrect. In particular, adding factors tends to decrease the effect size of the interaction compared to the simple effects (Simonsohn, 2014; Westfall, 2015b). For example, in a between subjects study, if an initial study found a d = .5 (η_p^2=.059) difference between two conditions, and a follow-up study adds a second factor with a condition that is expected to totally attenuate the initial effect (d = 0 in these cells), the overall interaction effect size would be d = .25, or η_p^2= .015 (see Westfall, 2015b for formula). Simonsohn (2014) used similar arguments to conclude that in a full-attenuation interaction, the researcher would need to have twice as many samples **per cell** as the two-cell design to retain the same level of power. In our example above, d = .5 is likely a much more reasonable effect size guess for the single-factor design than for the 2x2 interaction, and the smaller effect size expected in the latter would lead to a higher sample size recommendation.

As this example shows, the expected effect size of an interaction depends critically on the expected pattern of simple effects. Partial attenuation effects, in which the added simple effect goes in the same direction but is weaker, will entail an even steeper decrease in the effect size; for example, the difference between a d = .5 and a d = .25 simple effect is an interaction with size d = .125. Conversely, for the interaction effect size to equal the simple effect sizes, it takes a full “crossover” effect in which one simple effect is just as strong as the other but going in the opposite direction.

For interaction effects in factorial ANOVA, we recommend that researchers carefully think through the pattern of cell means they expect to see, and the sizes of the simple effects. With these two pieces of information, reasonable effect size estimates for higher order effects can be calculated and input to power analysis. While it can be easy to calculate the effect size for between subjects interactions with 2 levels for each factor, this can quickly get complicated in designs with more levels, or with mixed or within-subjects factorial designs. For these cases, tools that allow researchers to input expected means, standard deviations, and correlations, such as GLIMMPSE or simulation methods, may be especially useful.

***Follow-up Comparisons*.** A final complication in power analysis of ANOVA is that researchers are rarely interested in only the results of the overall F test. For interaction designs, researchers may also be interested in the direction and significance of simple effects, to support claims about the shape of the interaction. And, if any factors have three or more levels, researchers may further need to interpret some set of pairwise or complex contrasts among them. Studies that are well-powered for the F test of their higher order interactions may not be well-powered to test simple effects or contrasts. We suggest that researchers check power for all the tests of interest in their ANOVA. Doing so may well lead to a similar conclusion as considering likely interaction patterns: more participants are needed for the focused analyses in a factorial design than a basic power analysis of the interaction effect would suggest (Giner-Sorolla, 2018).

There are many different options for calculating follow-up tests (e.g. simple effects can either use the overall pooled ANOVA SD or the SD from just the cells involved in the comparison; researchers may or may not want to correct for follow-up tests, and if they do correct there are many different possible correction procedures). Researchers need to make sure that they calculate the power for the tests they are actually going to perform. Different software tools may make this more or less difficult. For example, if a research study will not be using the overall ANOVA pooled SD and will either be making no adjustments for follow-up tests or using adjustment methods that alter the alpha level, then point-and-click software tools may be sufficient for calculating the power for both the overall ANOVA and the follow-up tests. However, if a researcher plans to adjust for multiple follow-up tests using a method that alters the test statistic criterion, then it may be necessary to calculate power using simulation methods.

Alternative Tools. More flexible alternatives to G*power can help researchers analyze power for factorial designs. One point and click option is the GLIMMPSE tool, available freely online at: https://glimmpse.samplesizeshop.org/ (Kriedler et al., 2013). The tool allows for either power or sample size calculations and can be used for between subject, within subjects, factorial designs, and ANCOVAs. It does require researchers to input individual cell means, standard deviations, and correlations, which they may not know a priori. Another option is to use statistical software to do a power analysis through simulations. Similar to the GLIMMPSE tool however, researchers need to specify means, standard deviations, and correlation structures for ANOVA simulations. A final option is the PANGEA shiny app, https://jakewestfall.shinyapps.io/pangea/ (Westfall, 2016a). The app can calculate power for a number of different ANOVA designs, but effect sizes are inputted in terms of Cohen's d, which may not be intuitive for designs with more than 2 levels. We suggest that interested readers read the working paper on the app (Westfall, 2016b).

For all the above tools that require input of expected means and SD, the task can be made easier by thinking in terms of standardized mean differences rather than raw mean differences. If all SDs are set to 1, then the differences between cell means are the same as Cohen's ds. For example, imagine researchers wanted to calculate power for a 2 x 2 interaction where they expected to have a medium effect size, *d* = 0.5, in one pair of cells and then a 70% attenuation, *d* = 0.15 in the other. Regardless of what the actual means and SDs are, inputting means of 0, 0.5, 0, 0.15 and SDs all equal to 1 would represent the size of the interaction in this design. In this way, researchers can use intuition about simple effects, which may be easier to work with, to build up to higher order effects.

Mediation Analysis

Mediation analysis is a technique that has become popular in some fields of psychological science (e.g., social psychology) as well as other social science fields. It is an analytical method used to estimate the degree to which an independent variable may influence an outcome through a third variable, mediator. Mediation analysis cannot provide evidence of cause, as cause relies on not just covariation but also evidence of temporal precedence and elimination of alternative explanations, which no statistical test can provide (Hayes, 2018; Spencer, Zanna, & Fong, 2005; Thoemmes, 2015).

Modern methods of testing mediation focus around estimating the indirect effect, which is the product of the effect of the independent variable on the mediator and the effect of the mediator on the outcome, controlling for the independent variable. Because the sampling distribution of the indirect effect is irregular in shape (i.e., not normally distributed), inferential methods for mediation typically rely on resampling approaches (e.g., bootstrapping or Monte Carlo confidence intervals). A particular issue in power analysis is that the shape of the distribution for the indirect effect depends on the size of the population indirect effect, unlike other sampling distributions (e.g., means, regression coefficients) where the center and spread of the distribution are independent.

Though G*Power does not calculate power for indirect effects, a variety of tools are available for estimating power in mediation analysis (See Table A1). Most of these tools rely on Monte Carlo simulations rather than analytic algorithms, due to the irregular shape of the sampling distribution. As they stand, these packages have a variety of limitations. In particular, all packages require some estimate of the paths or correlations among the variables involved in the analysis. As has been mentioned before, accurate estimates of these paths are likely difficult to find, a priori. As such we recommend researchers consider taking a “smallest effect of interest” approach when estimating power; however, we acknowledge that there are many effects of interest in mediation analysis, and it can often be difficult to define the smallest effect of interest for each path. Different tools require the input in different forms: most popularly, unstandardized coefficients, raw correlations, standardized coefficients, or partial correlations. Researchers should be cautious and read documentation to know exactly what type of input is needed for accurate estimation in power. Unstandardized coefficients may be very different from standardized coefficients, and correlations and partial correlations are not necessarily scaled the same way as standardized coefficients, particularly when there are multiple predictors in the model (as in mediation analysis).

It is particularly important to choose a power tool which uses the inferential method that will ultimately be used in analysis. Mediation analysis is in this way unique, as there is no single preferred method for inference, but rather a few different methods which perform very similarly (Hayes & Scharkow, 2013). For example, WebPower (Zhang & Yuan, 2018) uses the delta method / Sobel test, which assumes that the sampling distribution of the indirect effect is normal.

This is only true in very large samples; however, the same assumption is made in many structural equation modeling programs, so if this assumption will be made in data analysis, the same assumption should be made in power analysis. Alternatively, many researchers use the PROCESS macro (Hayes, 2018) for analysis which uses bootstrapping or Monte Carlo confidence intervals. In this case WebPower (Schoemann, Boulton, & Short, 2017) or bmem (Zhang & Wang, 2013) should be used.

Precision approaches are particularly difficult in mediation analysis, because the precision of the estimate will depend on the size of the population indirect effect. To add to the difficulty, typical approaches to inference for indirect effects involve bootstrapping or other resampling approaches, these result in unpredictable confidence interval widths, as the widths depend on the estimated sampling distribution.

Power analysis in mediation is complex and requires computationally intensive methods to respect the irregularly shaped distribution of the indirect effect. Extensions from the simple mediation case are still quite limited, though Table A1 notes a few tools which can accommodate multiple mediator models of two forms: parallel, when there is no causal order among mediators assumed, and serial, when there is a causal order assumed among mediators. Extensions of power analysis to conditional process analysis, models that integrate mediation and moderation, are currently lacking. However, it is worthwhile to note that given the complexity of the models, many parameter estimates would be needed, adding to the arbitrary nature of power analysis as currently conducted. Intuitively, combining power analysis for mediation and rules of thumb of moderation/interaction, researchers should consider the recommended sample size based on a simple mediation effect they are interested in and scale the sample size based on the expected pattern of moderation (i.e., 1 N for full reversal of effect, 2 N for complete attenuation; see

Factorial ANOVA section, above); however, the only way to have precise sample size estimates is to use a Monte Carlo simulation. As in ANOVA, researchers should consider the complete set of tests, including simple effects, they intend to estimate, and consider their power to detect all effects of interest.

Table A1: Summary of power analysis tools for mediation.

Tool Name	MCpowrMed	WebPower	Bmem	MedPower	pwr2ppl
Interface	Online/R	Online/R	R	Online-Shiny	R functions
Reference	Schoemann, Boulton, Short (2017)	Zhang & Yuan (2018)	Zhang & Wang (2013)	Kenny (2017)	Aberson
Inputs	Correlations, standard deviations or estimates of paths	Estimates of a and b, variances	Estimates of all paths, skew, kurtosis, N	Estimates of paths or partial correlations	Correlations, N
Outputs	Required N (Given Power), Estimated Power (Given N)	Required N (Given Power), Estimated Power (Given N)	Estimated Power (Given N)	Required N (Given Power), Estimated Power (Given N)	Estimated Power (Given N)

Standardized	No	No	Yes	Yes	No
Method of Estimation	Monte Carlo	Sobel	Sobel, Bootstrap	Distribution of the Product	Sobel
Model Complexity	Simple, Parallel, Serial	Simple	Simple, Parallel, Serial	Simple	Simple, Parallel
URL	http://marlab.org/power_mediation/	https://webpower.psychstat.org/models/med01/	https://cran.r-project.org/web/packages/bmem/bmem.pdf	https://davidakenny.shinyapps.io/MedPower/	https://github.com/chrisaberson/pwr2ppl/blob/master/R/med.R

Path Analysis and Structural Equation Modeling

Structural Equation Modeling (SEM) is a model-fitting technique that takes into account measurement error and latent variables, testing relationships among several continuous variables. The technique is known as path analysis when all variables are single measured indicators rather than latent multiple-indicator variables. SEM can also be used to carry out factor analysis, mediation, and other multivariate techniques.

In SEM and similar approaches, as with regression, power can be determined for different aspects of a model. These include power: (a) to reject the null hypothesis that a model shows a pre-specified degree of overall fit (H_0); (b) to detect a significant difference in fit between nested models; and (c) to reject the null hypothesis that a given path-coefficient is zero. Analytic approaches to power for path analysis and SEM exist, but Monte Carlo power analysis is usually the simpler solution. Tools for Monte Carlo power analysis in this area include Mplus software (Muthén & Muthén, 1998-2017) and simsem (an R package using either lavaan or OpenMx). Monte Carlo might be the best option for power in Confirmatory Factor Analysis

(CFA) models in SEM, depending on the hypothesis being tested. In many cases, it is the only option.

Fit-based null-hypothesis testing for one model. MacCallum, Browne, and Sugawara (1996) discuss using one particular fit index, the Root Mean Square Error of Approximation (RMSEA; symbolized as Epsilon "ε"), to determine power to reject a given fit-based null-hypothesis. Examples include testing: the null-hypothesis of exact-fit (ε_0 = .00) vs. the alternative hypothesis of a close-fit (ε_A = .05); the null-hypothesis of close-fit (ε_0 = .05) vs. non-close fit (ε_A = .08); and the null-hypothesis of not-close fit (ε_0 = .05) vs. ε_A = .01. MacCallum et al.(1996) provide power tables for various values of *N*, df, and type of H_0 test (close/not-close/exact) and SAS syntax for determining power and minimum sample for a given level of power.

Power for detecting differences in fit between nested models. SEM researchers frequently compare nested models, in which, for the same constellation of latent constructs and manifest indicators, a baseline model contains certain paths and an alternative model either adds paths or subtracts paths (but not both) from the baseline model. Alternative models that freely estimate a set of paths vs. constraining them to equality (e.g., in a multiple-group model comparing men and women) are also nested. MacCallum, Browne, and Cai (2006) note that, "Power analysis for a test of the difference between two models requires the investigator to define an effect size representing the degree of difference under the alternative hypothesis" (p. 19). Each model generates a measure of discrepancy (or fit) between the sample covariance matrix (similar to correlations) and that implied by the model. The effect size then represents the difference in population discrepancy values between nested models A and B, with the noncentrality parameter defined as sample size times effect size (MacCallum et al., 2006, Equation 2).

Because population values generally are unknown, researchers need an estimate of the noncentrality parameter. MacCallum et al. (2006) again recommend using the RMSEA (ε) in this context (Equation 4). Researchers must choose the respective RMSEA values for the two nested models being compared, for which MacCallum et al. (2006) provide guidelines. MacCallum and colleagues (2006) also provide supplementary SAS syntax for power computations. Finally, these authors discuss factors (e.g., sample size, difference in models' degrees of freedom, differences in ε) that affect power (which are closely linked to the noncentrality parameter) in nested-model comparisons. Preacher and Coffman (2006) provide an online R-based interactive power calculator for comparing nested models.

Power for detecting significance of path coefficients. Wolf, Harrington, Clark, and Miller (2013) demonstrated the use of Monte Carlo simulations in studying the effects of several model features (e.g, number of factors, number of indicators, and magnitude of factor loadings) on minimum sample size required to meet simultaneously all of three criteria for a desirable model. These criteria included: power of at least 80% for all loadings, correlations, and paths of interest; accuracy of parameter estimates relative to "population parameters" specified in generating the simulation samples; and "solution propriety" (i.e., proper model convergence, no out-of-range parameter values). Wolf et al.'s (2013) investigation was thus very much in the spirit of Maxwell et al.'s (2008) call for attention to both statistical power and accuracy/precision. As Wolf et al. (2013) noted, "attending only to statistical power is problematic when contemplating sample size requirements. In many models, statistical power was not the limiting factor driving sample size requirements, but rather, bias or errors (i.e., solution propriety) were the culprit" (p. 927). Wolf and colleagues (2013) also provided Mplus syntax for conducting Monte Carlo simulations.

Advanced Latent-Variable Modeling

Power-related recommendations in the use of SEM-inspired techniques such as latent growth modeling and latent class analysis have been published. These are summarized briefly. Hertzog, von Oertzen, Ghisletta, and Lindenberger (2008) used Monte Carlo simulation to examine power in latent growth modeling to detect individual differences in participants' longitudinal slopes (i.e., slope variability) based on sample size, number of assessment waves, and growth-curve reliability (low residual variance in the measured indicators of the growth trends). A greater number of waves was found to increase power, but study features must be considered in conjunction with each other. Collecting four waves of longitudinal data may be beyond the resources of many investigators, but even with this many assessments, heightened power was shown to be difficult to attain. Maxwell et al.'s (2008) discussion of latent growth curves focused on randomized trials and other group comparisons. Aspects of latent class analysis (also known as mixture modeling) can be enhanced through power considerations, as well. Maxwell et al. (2008) briefly addressed mixture modeling, concluding that larger sample size facilitated selection of the correct model, but also led to overestimation of the number of classes. Dziak, Lanza, and Tan (2014) report on power in using the bootstrap likelihood ratio test to decide between selecting K vs (K — 1) latent classes.

Multilevel Modeling

Multilevel modeling (MLM) is a method investigating data which are nested at two or more levels, such as individuals nested within different work groups or responses to different vignettes nested within individual. The complexity of the data structures in MLM provides additional challenges in power analysis and sample size determination. In this section we will focus on two-level hierarchically nested data, but the approaches we discussion will work for other, more complex nesting structures, e.g. three-level data, cross-classified data.

Unlike single-level models, in an MLM there is not a single sample size, *N*, rather there are, at least, two sample sizes to consider, the number of level 1 units, *N*, and the number of level 2 units, *J*. In addition, researchers must consider whether the level 1 units are balanced across level 2 units, because when there is heavy imbalance, power is overestimated (Konstantopoulos, 2010). Power in ML models is also affected by the intraclass correlation (ICC), the proportion of variance in the outcome at level 2, and the residual variances at both levels 1 and level 2. In addition, there are two different types of effects estimated in MLM, fixed effects: an intercept or slope which does not vary across higher level units, and random effects, the variability of intercepts and or slopes across higher level units. Further complicating power analysis in MLM, there are no widely agreed upon, well defined effect size metrics across models. Some specific designs, e.g. cluster randomized trials (Hedges, 2007), have established effect sizes, but for many models determining an effect size of interest is very difficult (Bakeman, 2005; Rights & Cole, 2018; Olejnik & Algina, 2003; Westfall, Kenny, & Judd, 2014).

Required sample sizes for appropriate power to detect effects in MLMs are also affected by the level of a predictor of interest. In general, increasing sample size at the level of the predictor of interest will increase power to a greater degree than increasing sample size at a different level. If the predictor of interest is measured at level 1, then increasing level 1 sample size (N) will have a larger effect on power than increasing level 2 sample size (J). If a predictor is measured at level 2 then increasing level 2 sample size (J) will have a larger effect on power than increasing level 1 sample size (N).

Given these complexities, traditional power analysis software, e.g., G*Power, does not compute power for MLMs. Power analysis software for MLM using an analytic approach have focused on specific constrained situations, e.g. cluster randomized trials. Monte Carlo power analyses provide a much more flexible framework for power analysis with MLMs and are available with specific software, e.g. mlPowSim or SIMR, or as part of general modeling software, e.g. Mplus. Excellent tutorials are available for SIMR (Arend & Schäfer, 2019) and Mplus (Lane & Hennes, 2018).

Table A2. Summary of power analysis tools for multilevel models.

Tool Name	**Optimal Design**	**WebPower**	**PinT**	**mlPowSim**	**SIMR**	**ML Power Tool**
Interface	Stand Alone (Windows)	Online/R	Stand Alone (Windows)	Stand Alone/R/ MlwiN	R	R/Web (Shiny)
Reference	Raudenbush et al. (2011)	Zhang & Yuan (2018)	Snijders & Bosker (1993)	Brown, Lahi, & Parker (2009)	Green & MacLeod (2016)	Mathieu, Aguinis, Culpepper, & Chen (2012)
Inputs	N, J, ICC, fixed effect of interest (*d*), R^2 for covariates	N, J, ICC or level 1 and level 2 variances, fixed effect of interest (*d*)	Means, variances, and covariances of predictors, level 1 and level 2 variances, cost function and budget or range of N and J	N, J, means, variances, and covariances of predictors, level 1 and level 2 variances	N, J, means, variances, and covariances of predictors, level 1 and level 2 variances, input must be formatted as an lmer model	N, J, ICC, fixed effects, random effects
Outputs	Power, MDES, J, N, ICC	Power, MDES, J, N, ICC	Standard errors	Power	Power	Power
Types of Parameters Considered	Fixed effects with a binary predictor	Fixed effects with a binary predictor, random effects for some models	Fixed effects, random effects	Fixed effects, random effects	Fixed effects, interactions of fixed effects, random effects	Fixed effect, cross-level interaction effect
Method	Analytic	Analytic	Analytic	Monte Carlo	Monte Carlo	Monte Carlo

(continued)

Models Fit	Two or three level models for randomized trials with continuous outcomes, (person or cluster randomized)	Two level models for randomized trials with continuous outcomes, with two or three arms (cluster randomized only)	Two level models with continuous outcomes, any combination of level 1 and level 2 variables and any number of random slopes	Two and three level, and cross classified models, with continuous, binary, and count outcomes and any combination of level 1 and level 2 variables and any number of random slopes	Multilevel (no known limit on levels) and cross classified models with continuous, binary, and count outcomes; any combination of level 1 and level 2 variables; any number of random slopes	Two level model estimating a cross-level interaction
Additional Features	Empirically derived effects sizes from education	Effect size calculator		MCMC estimation when using MlwiN	Any features available for lmer models	
URL	http://hlmsoft.net/od/	https://webpower.psychstat.org/wiki/	https://www.stats.ox.ac.uk/~snijders/multilevel.htm#progPINT	http://www.bristol.ac.uk/cmm/software/mlpowsim/	https://cran.r-project.org/web/packages/simr/index.html	http://www.hermanaguinis.com/crosslevel.html

Appendix: Calculations

Omega Squared Calculations

For one-way designs or 2-way between subject designs, ω^2 can be calculated based on the formulas in Maxwell & Delaney, 2004; Olejnik & Algina, 2000):

One-way designs: (F-1)/(F + (df2 + 1)/df1))

2-way between subjects design: df1(F – 1)/df1(F -1) + N

For factorial within subjects, mixed designs, and designs in which not all factors are manipulated, the calculations become more complicated. Interested readers should look to Maxwell and Delaney (2004) and Olejnik & Algina (2000, 2003) for the appropriate formulas.

Effect Size for Chi Squared Demonstration

The effect size *w* for chi-squared tests in G*Power is defined across *i* = 1 … *n* cells as:

$$w = \sqrt{\sum_{i=1}^{n} \frac{\left(p_{0i} - p_{1i}\right)^2}{p_{0i}}}$$

where p_{0i} refers to the proportion in a given cell under H_0 and p_{1i} refers to the same proportion under H_1 (in this case, the smallest departure from p_{0i} of interest). For example, if one wanted to whether reports of a coin toss differ from what is by chance (win: 50%, loss: 50%) with a minimum effect size of (win: 60%, loss: 40%), then *w* would be:

$$w = \sqrt{\frac{\left(.5 - .60\right)^2}{.5} + \frac{\left(.5 - .40\right)^2}{.5}} = .20$$

Multiple Regression Calculations

Examining the formulae for the regression coefficient and squared semipartial correlation demonstrates some challenges in determining power for multiple regression designs. Formula X and Y (relevant to a model with two predictors), demonstrates that two primary issues influence the size of *b** (the regression coefficient) and sr^2 (squared semipartial correlation). The numerator shows the correlation between the predictor of interest and the dependent variable (r_{y1}) on the left. On the right (i.e., being subtracted from r_{y1}) is the correlation between the second predictor and dv (r_{y2}) times the correlation between the two predictors (r_{12}). These formulae show the important role the correlation between predictors (multicollinearity) plays in estimation of the effect size. As correlations between predictors rise, effects tend to become smaller.

$$b^*_{y1.2} = \frac{\rho_{y1} - \rho_{y2}\rho_{12}}{1 - \rho_{12}^2} \qquad \text{X}$$

$$sr_1^2 = \left(\frac{r_{y1} - r_{y2}r_{12}}{\sqrt{1 - r_{12}^2}} \right)^2 \qquad \text{Y}$$

Non Centrality Parameter Calculation and Demonstration

The noncentrality parameter defined in Formula Z for an independent samples t-test is simply a measure of distance between the centers of the central and noncentral distribution.

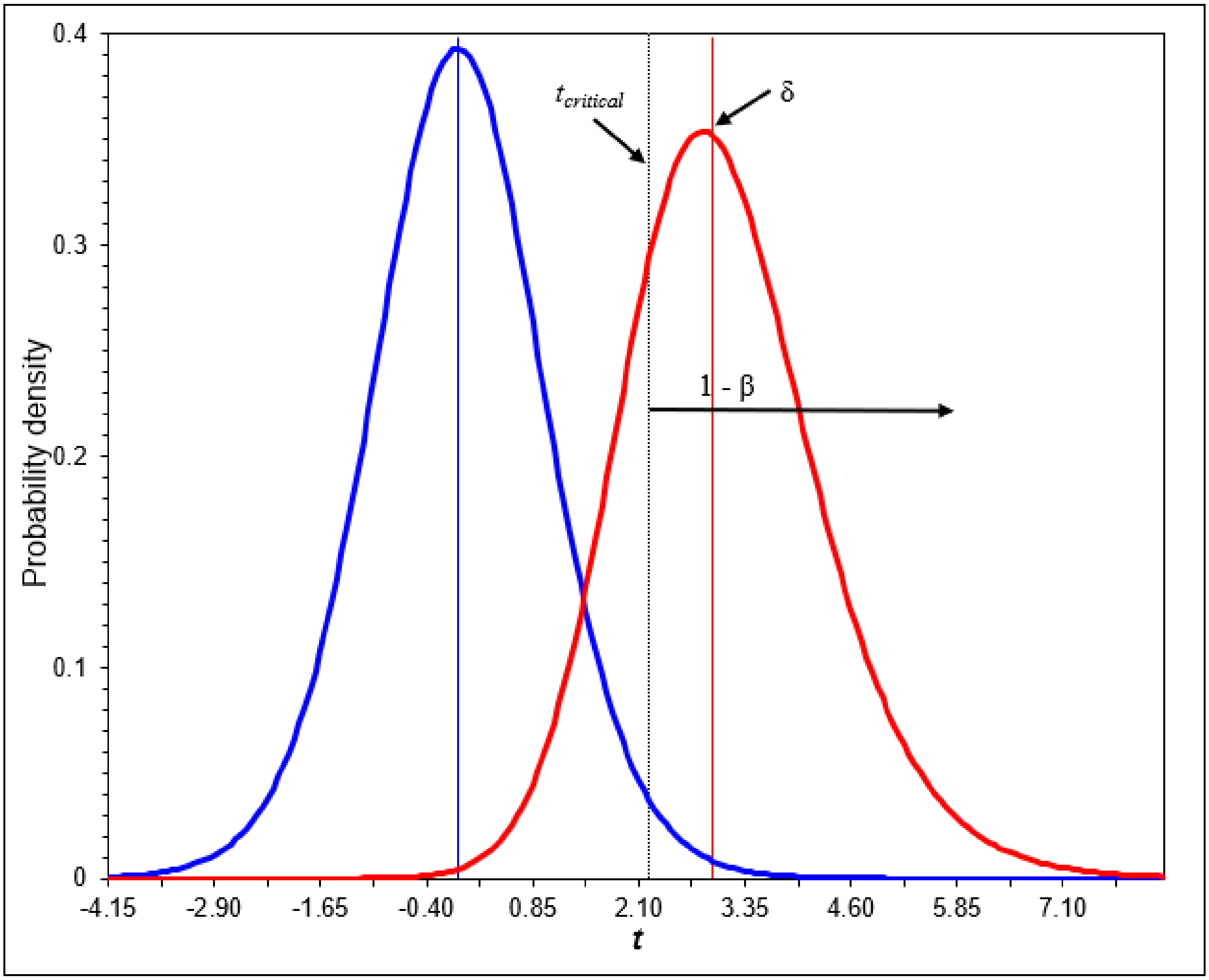
$$\delta = d\sqrt{\frac{n_j}{2}} \qquad \text{Z}$$

As an example, Figure B1 shows a situation with *df* = 18 (n_j = 10, reflecting 10 people per group), and a *d* = 1.34, yielding δ = 3.00.

$$\delta = 1.34\sqrt{\frac{10}{2}} = 3.0$$

The figure shows that the value of δ is simply the distance between the centers of the null (central, on the left) and alternative (non-central, on the right) distributions. The figure also presents $t_{critical}$, this is the t value corresponding to a two-tailed test using alpha = .05. Sample results that exceed that value, allow for rejection of the null hypothesis. The area under the alternative distribution that falls above that value reflects power (noted as 1-β).

Figure B1. Null (central) and alternative (noncentral) distributions and power.[4]



[4] This graph was generated using ESCI (https://thenewstatistics.com/itns/esci/)

References

Aberson, C. (2019). pwr2ppl: Power analysis for common designs. R package version 0.1. Retrieved from https://cran.r-project.org/web/packages/pwr2ppl/index.html.

Arend, M. G., & Schäfer, T. (2019). Statistical power in two-level models: A tutorial based on Monte Carlo simulation. *Psychological Methods*, *24*, 1-19.

Beaujean, A. A. (2014). Sample size determination for regression models using Monte Carlo Methods in R. *Practical Assessment, Research & Evaluation*, *19*(12). Available online: http://pareonline.net/getvn.asp?v=19&n=12

Behavioral Statistics, *32*, 41-370.

Botella, J., Ximenez, C., Revuelta, J., & Suero, M. (2006). Optimization of sample size in controlled experiments: The CLAST rule. *Behavior Research Methods, 38*, 65-76.

Browne, W. J., Lahi, M.G., & Parker, R. M. (2009). *A guide to sample size calculations for random effect models via simulation and the MLPowSim software package.* Retrieved from http://www.bristol.ac.uk/cmm/software/mlpowsim/mlpowsim-manual.pdf.

Buchanan, E. M., Gillenwaters, A. M., Padfield, W., Van Nuland, A., & Wikowsky, A. (2019). MOTE [Shiny App]. Retrieved from https://doomlab.shinyapps.io/mote/.

Buchanan, E. M., Gillenwaters, A. M., Scofield, J. E., & Valentine, K. D. (2019). MOTE. R package version 1.02. https://cran.r-project.org/web/packages/MOTE/MOTE.pdf.

Champely, S., Ekstrom, C., Dalgaard, P., Gill, J. … & De Rosario, H. (2018). pwr: Basic Functions for Power Analysis R package version 1.2-2. Retrieved from https://cran.r-project.org/web/packages/pwr/index.html.

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

Collins, L. M., Dziak, J. J., & Li, R. (2009). Design of experiments with multiple independent variables: A resource management perspective on complete and reduced factorial designs. *Psychological Methods, 14*, 202-224.

Dziak, J. J., Lanza, S. T., & Tan, X. (2014). Effect size, statistical power and sample size requirements for the bootstrap likelihood ratio test in latent class analysis. *Structural Equation Modeling, 21*, 534-552. Doi: 10.1080/10705511.2014.919819

Fitts, D. A. (2010a). Improving stopping rules for the design of efficient small-sample experiments in biomedical and biobehavioral research. *Behavior Research Methods, 42,* 3-22.

Fitts, D. A. (2010b). The variable-criterion sequential stopping rule: Generality to unequal sample sizes, unequal variances, or to large ANOVAs. *Behavior Research Methods, 42*, 918-929.

Giner-Sorolla, R., Aberson, C. L., Bostyn, D. H., Carpenter, T. Conrique, B. G., Lewis, Jr., N. A., Montoya, A. K., Ng, B. W., Reifman, A., Schoemann, A. M., & Soderberg, C. (2019). Powered to detect what? Considerations for planning and evaluating sample size. Preprint available online at https://osf.io/9bt5s/.

Green, P., & MacLeod, C. J. (2016). Simr: An R package for power analysis of generalised linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*, 493-498. Doi: 10.1111/2041-210X.12504

Hayes, A. F. (2018). *Introduction to mediation, moderation, and conditional process analysis*. (2nd Ed.). New York: The Guilford Press.

Hayes, A. F., & Scharkow, M. (2013). The relative trustworthiness of inferential tests of the indirect effect in statistical mediation analysis: Does method really matter?. *Psychological Science, 24*(10), 1918-1927.

Hayes, A. F., & Scharkow, M. (2013). The relative trustworthiness of inferential tests of the indirect effect in statistical mediation analysis: Does method really matter?. *Psychological Science, 24*(10), 1918-1927.

Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and*

Hertzog, C., von Oertzen, T., Ghisletta, P., Lindenberger, U. (2008). Evaluating the power of latent growth curve models to detect individual differences in change. *Structural Equation Modeling, 15*, 541–563. Doi: 10.1080/10705510802338983

Hertzog, C., von Oertzen, T., Ghisletta, P., Lindenberger, U. (2008). Evaluating the power of latent growth curve models to detect individual differences in change. *Structural Equation Modeling, 15*, 541–563. Doi: 10.1080/10705510802338983

Hughes, J. (2017). Paramtest: Run a function Iteratively while varying parameters. R package version 0.1.0. https://CRAN.R-project.org/package=paramtest

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, *103*, 54 – 69.

Judd, C. M., Westfall, J., & Kenny, D. A. (2017). Experiments with more than one random factor: Designs, analytic models, and statistical power. *Annual Review in Psychology, 68,* 1 – 25.

Kelley, K. (2007). Methods for the behavioral, educational, and social Science: An R package. *Behavior Research Methods, 39*, 979–984.

Kelley, K., & Maxwell S. E. (2003). Sample size for multiple regression: Obtaining regression coefficients that are accurate, not simply significant. *Psychological Methods, 8*, 305–321.

Kelley, K., & Rausch J. R. (2006). Sample size planning for the standardized mean difference: Accuracy in parameter estimation via narrow confidence intervals. *Psychological Methods, 11*, 363–385

Kenny, D. A. (2017, February). MedPower: An interactive tool for the estimation of power in tests of mediation [Computer software]. Available from https://davidakenny.shinyapps.io/MedPower/.

Kriedler, S. M., Muller, K. E., Grunwald, G. K., Ringham, B. M., Coker-Dukowitz, Z. T., Sahhadeo, U. R., … Glueck, D. H. (2013). GLIMPPSE: Online power computation for linear models with and without baseline covariate. *Journal of Statistical Software, 54*, i10.

Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology*, *44*, 701-710.

Lakens, D. (2016, December 3). Sequential analyses. Retrieved from osf.io/uygrs.

Lakens, D., & Caldwell, (2019). Simulation-based power-analysis for factorial ANOVA designs. Retrieved from https://psyarxiv.com/baxsf. (note: supports the ANOVAPower r package)

Lane, S. P., & Hennes, E. P. (2018). Power struggles: Estimating sample size for multilevel relationships research. *Journal of Social and Personal Relationships*, *35*(1), 7-31.

Leschinski, C. H. (2019). MonteCarlo: Automatic parallelized Monte Carlo simulations. R package version 1.0.6. https://CRAN.R-project.org/package=MonteCarlo.

Lin, H. (2019). hauselin/rshinyapp_effectsizeconverter: shiny effect size converter v0.0.1 (Version v0.0.1). Zenodo. https://doi.org/10.5281/zenodo.2563830

MacCallum, R. C., Browne, M. W., & Cai, L. (2006). Testing differences between nested covariance structure models: Power analysis and null hypotheses. *Psychological Methods*, *11*(1), 19-35. doi: 10.1037/1082-989X.11.1.19

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, *1*(2), 130-149. doi:http://dx.doi.org/10.1037/1082-989X.1.2.130.

Mathieu, J. E., Aguinis, H., Culpepper, S. A., & Chen. G. (2012). Understanding and estimating the power to detect cross-level interaction effects in multilevel modeling. *Journal of Applied Psychology, 97*(5), 951-966.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods, 9*, 147-163. doi: 10.1037/1082-989X.9.2.147

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data. A model comparison perspective*. Belmont, CA: Wadsworth.

Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus user's guide* (8th ed.). Los Angeles, CA: Muthén & Muthén. Retrieved from: https://www.statmodel.com/download/usersguide/MplusUserGuideVer_8.pdf

Olejnik, S., & Algina, J. (2000). Measures of effect size for comparative studies: Applications, interpretations and limitations. *Contemporary Educational Psychology*, *25*, 241-286.

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods, 8*(4), 434-447.

Preacher, K. J., & Coffman, D. L. (2006, May). Computing power and minimum sample size for RMSEA [Computer software]. Available from http://quantpsy.org/.

Raudenbush, S. W., Spybrook, J., Congdon, R., Liu, X. F., Martinez, A., Bloom, H., & Hill, C. (2011). Optimal design software for multi-level and longitudinal research (Version 3.01)[Software]. Available from *www.wtgrantfoundation.org*.

Reboussin, D. M., DeMets, D. L., Kim, K., & Lan, K. K. (2000). Computation for group sequential boundaries using the Lan-DeMets spending function method. *Controlled Clinical Trials, 21*(3), 190-207.

Rights, J. D. & Cole, D. A. (2018). Effect size measures for multilevel models in clinical child and adolescent research: New R-squared methods and recommendations. *Journal of Clinical Child & Adolescent Psychology*, *47*(6), 863-873.

Sagarin, B. J., Ambler, J. K., & Lee, E. M. (2014). An ethical approach to peeking at data. *Perspectives on Psychological Science*, *9*(3), 293-304.

Schoemann, A. M., Boulton, A. J., & Short, S. D. (2017). Determining power and sample size for simple and complex mediation models. *Social Psychological and Personality Science,* *8*(4), 379-386.

Sigal, M. J., & Chalmers, R. P. (2016). Play it again: Teaching statistics with Monte Carlo simulation, *Journal of Statistics Education*, *24*(3), 136-156.

Snijders, T. A. B. & Bosker, R. J. (1993). Standard errors and sample sizes for two-level research. *Journal of Educational Statistics*, *18*(3), 237-259.

Spencer, S. J., Zanna, M. P., & Fong, G. T. (2005). Establishing a causal chain: Why experiments are often more effective than mediational analyses in examining psychological processes. *Attitudes and Social Cognition*, *89*(6), 845 - 851. DOI: 10.1037/0022-3514.89.6.845

Thoemmes, F. (2015). Reversing arrows in mediation models does not distinguish plausible models. *Basic and Applied Social Psychology, 37*(4), 226-234. DOI: 10.1080/01973533.2015.1049351.

Westfall, J. (2016a). PANGEA (v0.2): Power analysis for general anova designs. [Shiny App]. Retrieved from https://jakewestfall.shinyapps.io/pangea/

Westfall, J., Kenny, D. A., & Judd, C. A. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General*, *143*(5), 2020-2045.

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement, 73*, 913-934. doi: 10.1177/0013164413495237

Ximenez, C. & Revuelta, J. (2007). Extending the CLAST sequential rule to one-way ANOVA under group sampling. *Behavior Research Methods, 39*(1), 86-100.

Zhang, Z., & Wang, L. (2013). Methods for mediation analysis with missing data. *Psychometrika*, *78*(1), 154-184.

Zhang, Z., & Yuan, K. H. (2018). *Practical Statistical Power Analysis Using Webpower and R* (Eds). Granger, IN: ISDSA Press.