INTERNATIONAL REVIEW OF SOCIAL PSYCHOLOGY

Perugini, M., et al. (2018). A Practical Primer To Power Analysis for Simple Experimental Designs. *International Review of Social Psychology*, 31(1): 20, 1–23, DOI: https://doi.org/10.5334/irsp.181

RESEARCH ARTICLE

A Practical Primer To Power Analysis for Simple Experimental Designs

Marco Perugini, Marcello Gallucci and Giulio Costantini

Power analysis is an important tool to use when planning studies. This contribution aims to remind readers what power analysis is, emphasize why it matters, and articulate when and how it should be used. The focus is on applications of power analysis for experimental designs often encountered in psychology, starting from simple two-group independent and paired groups and moving to one-way analysis of variance, factorial designs, contrast analysis, trend analysis, regression analysis, analysis of covariance, and mediation analysis. Special attention is given to the application of power analysis to moderation designs, considering both dichotomous and continuous predictors and moderators. Illustrative practical examples based on G*Power and R packages are provided throughout the article. Annotated code for the examples with R and dedicated computational tools are made freely available at a dedicated web page (https://github.com/mcfanda/primerPowerIRSP). Applications of power analysis for more complex designs are briefly mentioned, and some important general issues related to power analysis are discussed.

Keywords: power analysis; effect size; moderation; sensitivity analysis; uncertainty

> *"When I finally stumbled onto power analysis, and managed to overcome the handicap of a background with no working math beyond high school algebra (to say nothing of mathematical statistics), it was as if I had died and gone to heaven."*
>
> —Jacob Cohen

Power analysis is one of the most fundamental tools that researchers can use when planning studies. It was pioneered in psychology more than fifty years ago by Jacob Cohen (1962, 1990). Since then, many have recommended its use and suggested it as good research practice (Wilkinson & TSFI, 1999) within the Null-Hypothesis Significance Testing (NHST) framework.[1] However, interest in power analysis has increased considerably only during the last few years. One reason for this is the recent replicability crisis in psychology; one of the main culprits for the difficulty in replicating some results was that original studies were often underpowered to start with (Asendorpf et al., 2013; Bakker, van Dijk & Wicherts, 2012; Swiatkowski & Dompnier, 2017). In the presence of publication bias, systematically performing studies that lack the power to detect effect sizes of interest results in a prevalence of false-positive findings in the literature (Button et al., 2013; Maxwell, 2004). One of the main benefits of power analysis when planning studies is that researchers become aware of their chances of finding an effect of interest. If these chances are insufficient, they should consider changes that could increase the probability of observing a significant effect.

This article aims to remind readers what power analysis is, why it matters, and when and how it should be used. The focus is on simple experimental designs often encountered in social psychology, and we will provide illustrative examples throughout the article. We will focus mainly on between-subject designs, and we will limit our discussion of repeated-measures designs to the simplest case of two dependent groups. We will also discuss some important issues related to power analysis. The goal is to present a practical primer to power analysis that complements other reviews of power analysis (Cohen, 1992a, 1992b; Faul, Erdfelder, Buchner & Lang, 2009; Faul, Erdfelder, Lang & Buchner, 2007; Maxwell, Kelley & Rausch, 2008) as well as more comprehensive and advanced textbooks (Cohen, 1988; Liu, 2014).

What Power Analysis Is and Why It Matters

Within the NHST approach, the main goal is to ascertain whether the null hypothesis (H_0) can be rejected. There are two types of errors: rejecting the null hypothesis when it is true (False Positive, typically referred to as α or Type I error) and failing to reject it when it is false (False Negative, typically referred to as β or Type II error). Within this framework, the power of a statistical test is the probability of successfully rejecting the null hypothesis when it is false ($1 - \beta$). Power depends on sample size, effect size, and the decision criterion (α-level): given three of these elements, one can derive the fourth. In particular, power

Department of Psychology, University of Milan Bicocca, IT

Corresponding author: Marco Perugini (marco.perugini@unimib.it)

increases with increasing sample size, increasing effect size, and more lenient decision criteria (e.g., $\alpha = .10$ instead of $\alpha = .01$). The conventional value of power as .80 (and of β as .20) considers the cost of a Type I error four times more serious than the cost of a Type II error when α is also set to its conventional value of .05 (thus $\beta/\alpha = 4$). Different values of β can be appropriate, depending on the desired balance between Type I and Type II errors (Cohen, 1988). This latter relation might give a wrong impression that one can only choose between the two types of errors, that is by balancing false positives (α) against false negatives (β). However, given a certain α level, increasing statistical power by collecting larger samples increases the accuracy of any result that emerges, which means that inferences from data are in general more correct (Maxwell et al., 2008; see also Ioannidis, 2005; Sterne & Davey Smith, 2001). When results are more accurate and inferences from data more correct, everything else being equal, they are more likely to be replicable (Asendorpf et al., 2013; Maxwell, 2004). In brief, statistical power matters not only because it directly increases the likelihood of finding an effect if it exists, but also because it contributes indirectly to reducing the overall rate of data inference errors (O'Brien & Castelloe, 2007). Said otherwise, an appropriate use of power analysis when designing a study increases the chance of getting it right, which is a main motivating factor for a scientist.

How Power Analysis Works and When To Use It

If statistical power is so important, a key question becomes how to increase it. The answer is simple: for any given α-level, statistical power goes up with increasing samples sizes and effect sizes. What it means to increase sample size is straightforward. We shall briefly address some possible strategies for increasing effect sizes as well. Several indicators of effect sizes can be used depending on the specific study design and the level of measurement of the variable of interest. We restrict our attention to interval level (continuous) dependent variables and moderately simple study designs. Interested readers should consult dedicated literature for further details on effect size types and their corresponding equations (Ellis, 2010; Fritz, Morris & Richler, 2012; Lakens, 2013). We assume that most readers are familiar with Cohen's *d*, which expresses effect size as the standardized mean difference between two conditions $\left(d = \frac{M_1 - M_2}{SD_{pooled}}\right)$, and with its conventional values of 0.20, 0.50, and 0.80 used to indicate a small, medium, and large effect size, respectively. They might be less familiar, however, with the corresponding benchmark values when expressed in other metrics. For this reason, in **Table 1** we provide a simple conversion to other common effect size indicators, such as r, f, and η^2. The lesser-known effect size area under the receiver operating characteristics (AUC) requires a brief explanation. This index expresses effect size as the probability that a person picked at random from one group will have a higher score than a person picked at random from the other group (Ruscio, 2008; Ruscio & Mullen, 2012). An AUC value of 0.50 means that the effect size is null (e.g., no improvement from a random selection device, such as tossing a coin); whereas, values going towards 1 imply larger effect sizes, until every person from one group has a greater score than every person from the other group (i.e., the two distributions do not overlap). This index of effect size also applies to ordinal dependent variables and is robust to violations of normality and to outliers.

Table 1: Conversion between some effect sizes.

	small	medium	large
d	0.20	0.50	0.80
r	0.10	0.24	0.37
f	0.10	0.25	0.40
η^2	0.01	0.06	0.14
AUC	0.56	0.64	0.71

Power crucially depends on the population effect size, which is typically unknown. When performing power analysis, a researcher should always use the best available guess of the population effect size. If previous research is available, especially meta-analyses, one can estimate the population effect size using sample-based effect size indices. However, different sample estimates of effect size are often available for the same population quantity, each index having different degrees of bias. Many sample estimates of effect sizes are upwardly biased: using these indices, as compared to unbiased estimates, tends to affect power analysis towards suggesting smaller samples or larger power. It is evident that one should try to input the least-biased index available. Note, however, that for many indices the difference in bias tends to become increasingly small as the sample becomes larger. For instance, Cohen's *d* is defined for the population (Cohen, 1988) and using the same formula on sample data to estimate the population parameter leads to biased results. It is known that Hedges' *g* is a less-biased estimate[2] of Cohen's *d* (Hedges, 1981), and its bias tends to become negligible for sample sizes $N > 20$ (Hunter and Schmidt, 2004). Thus, it may be used in software that requires a standardized difference as the input effect size index.

When data are not available as the basis for the effect size estimation, the researcher needs to guess the population effect size. There are effect sizes that are easier to guess because they correspond better to what a researcher may anticipate about the expected data. In the following pages, we try to outline different methods and several effect size indices that are, in our opinion, relatively easy to anticipate given some general hypotheses about the expected results.

There are different ways to perform power analysis (cf. Faul et al., 2007). The most common is *a priori* (prospective) *power analysis* in which the goal is to achieve a given desirable power level (e.g., .80) given a certain α-level. This value is commonly fixed to .05, but one should consider also using .005, in line with recent calls to redefine the significance threshold for novel findings (Benjamin et al., 2018), or justifying an α-level before beginning a data collection (Lakens et al., 2018). Once power and α-level are fixed, it is required to estimate an expected effect size and then calculate how many participants are needed to achieve the desired power. The relative simplicity of this calculation masks an important problem: the expected effect size is one's best guess, and its inaccuracy has substantial implications for the actual sample size needed to achieve the desired level of power. Researchers should routinely consider different scenarios by varying the expected effect size and ascertaining what would be the implications regarding achieved levels of power given a certain effect size and needed sample sizes, given a certain desirable power level. They can also formally consider the uncertainty in the estimate, which is reflected in its confidence interval, and then settle on a sample size that takes into account the desired level of protection against overestimating the effect size and consequently running an underpowered study (Safeguard Power Analysis; Perugini, et al., 2014; see also Anderson, Kelley & Maxwell, 2017).

Sometimes, however, researchers do not have much leeway for increasing sample size and instead have a relatively fixed maximum sample size. Under these relatively common conditions, power analysis can still be useful for determining the strength of an effect that can be reliably detected. This approach is called *sensitivity analysis* (Faul et al., 2007) and requires fixing a certain α-level, the available sample size, and a desired level of power to identify the minimum size of the effect that can be reliably detected. By plotting power levels and effect sizes, one can inspect their interplay. An interesting variation of this scenario is to calculate the minimum detectable effect size (MDES; Bloom, 1995) or, similarly, the smallest effect size of interest (SESOI; Lakens, 2014). The basic idea is that, given a certain α-level, sample size, and desired level of power, there is a minimum effect size that can be significantly detected. Effect sizes smaller than that value will not be significant. Researchers could commit to collecting a sample whose size is sufficient to detect "the smallest effect size that is deemed worthwhile to study" (Albers & Lakens, 2018). The implied value can be calculated by transforming the probability distribution statistics (e.g., t-value) into the effect size estimate (e.g., Cohen's *d*; see for instance Lakens, 2013).

Finally, we wish to stress that power calculated based on the results of the study (*post hoc* or retrospective power analysis) is pointless and potentially misleading. It amounts to a trivial transformation of the obtained p-value and provides no valid information concerning the actual power of the study (cf. Zumbo & Hubley, 1998). Therefore, requiring a power analysis after a study has been conducted (e.g., for the revision of a manuscript) is of questionable utility. Instead, a sensitivity analysis, whereby the minimum effect size that could reliably yield (e.g., with power 0.80) a statistically significant result (e.g., setting $\alpha = .05$) given the sample size, could be more informative. In fact, in this way, a reader will have some elements to judge whether the minimum effect size is realistic given knowledge accumulated in the field. Another possible approach is to calculate the safeguard sample ratio, which reflects the strength of empirical evidence provided by a given study by comparing the required sample size, as estimated with safeguard power analysis, and the sample size of the original study (Perugini, Gallucci & Costantini, 2014). For example, suppose that in one study with that 100 participants one obtains $d = 0.50$. One can calculate the needed sample size to obtain a safeguard power at 80% is 232 participants. Hence, the safeguard sample ratio (required sample size divided by original sample size) is 2.32. If, instead, another study obtained the same effect size value with 200 participants, the needed sample size was 176 participants, resulting in a safeguard sample ratio of 0.88. Therefore, one can infer that this second study provides more robust evidence than the first one, everything else being equal.

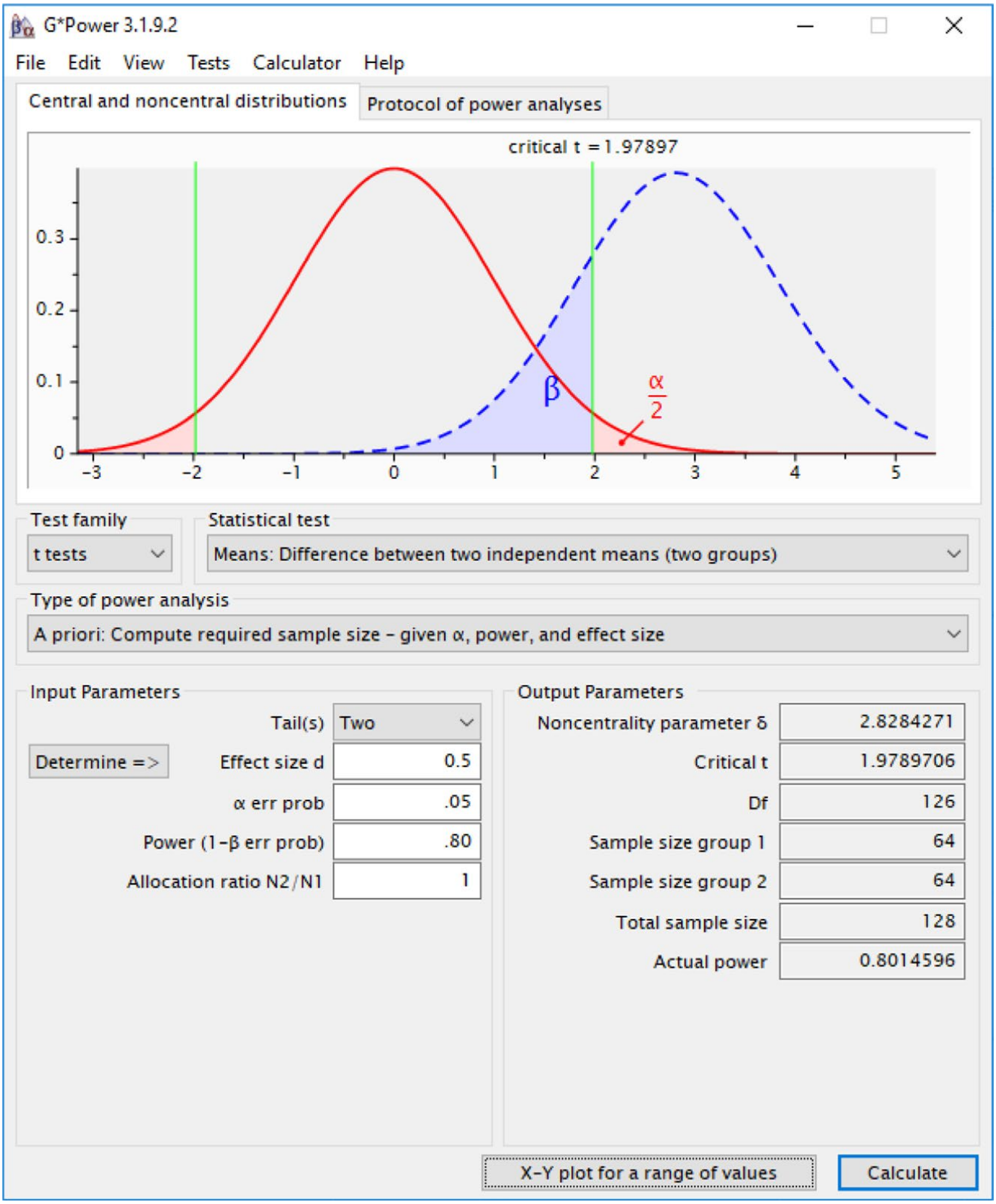
How To Do Power Analysis

Power analysis can be performed with a range of dedicated packages and routines. We shall focus on the most known and widely used free software for power analysis, G*Power (Faul et al., 2009), and on packages and routines available in the most known free statistical package, R (R Core Team, 2017). The software allows estimating power parameters for the same research design using different methods and different user interfaces. In the following examples, we present one way to resolve each of the discussed designs, with the implicit assumption that the methods we employ are not the only ones available. In general, we wish to use the same method for as many research designs as possible, with the aim of reducing the number of software commands and interfaces the user needs to learn.

*Main features of G*Power*

One of the simplest applications of power analysis is on a two independent groups design, where a dependent variable is measured in two groups of participants, with each participant belonging to one group only. Often, the aim of the study is to compare the means of the dependent variable between the two groups, employing a t-test, and the most commonly used effect size index for this design is Cohen's *d* (Cohen, 1988). In the simplest scenario, the researcher finds substantial evidence in the literature to estimate the expected effect size. For instance, suppose that a meta-analysis on the effect under investigation suggests a Cohen's *d* of 0.50, that the conventional statistical significance level is set to $\alpha = .05$, and that the desired power to $1-\beta = .80$. The aim of the power analysis in the following examples is prospective to estimate the minimum sample size *N* necessary to obtain a statistically significant test with a certain likelihood applied to the expected effect size index.

The first action needed in G*Power is to select the appropriate test in the Test Family menu by choosing *t-test* (cf. **Figure 1**). This action makes the list of applications of the t-test available in the Statistical Test menu. In this menu, select "Means: Difference between two independent means (two groups)", and then select "A priori […]" as the type of power analysis to be executed. The lower panel of the window presents the analysis input parameters that are required. The three actions taken so far are common to any power analysis run in G*Power: select a test family, a specific application of the test, and the type of power analysis. For this prospective power analysis, the



Test family: t tests

Statistical test: Means: Difference between two independent means (two groups)

Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

Input Parameters

Tail(s): Two
Effect size d: 0.5
α err prob: .05
Power (1−β err prob): .80
Allocation ratio N2/N1: 1

Output Parameters

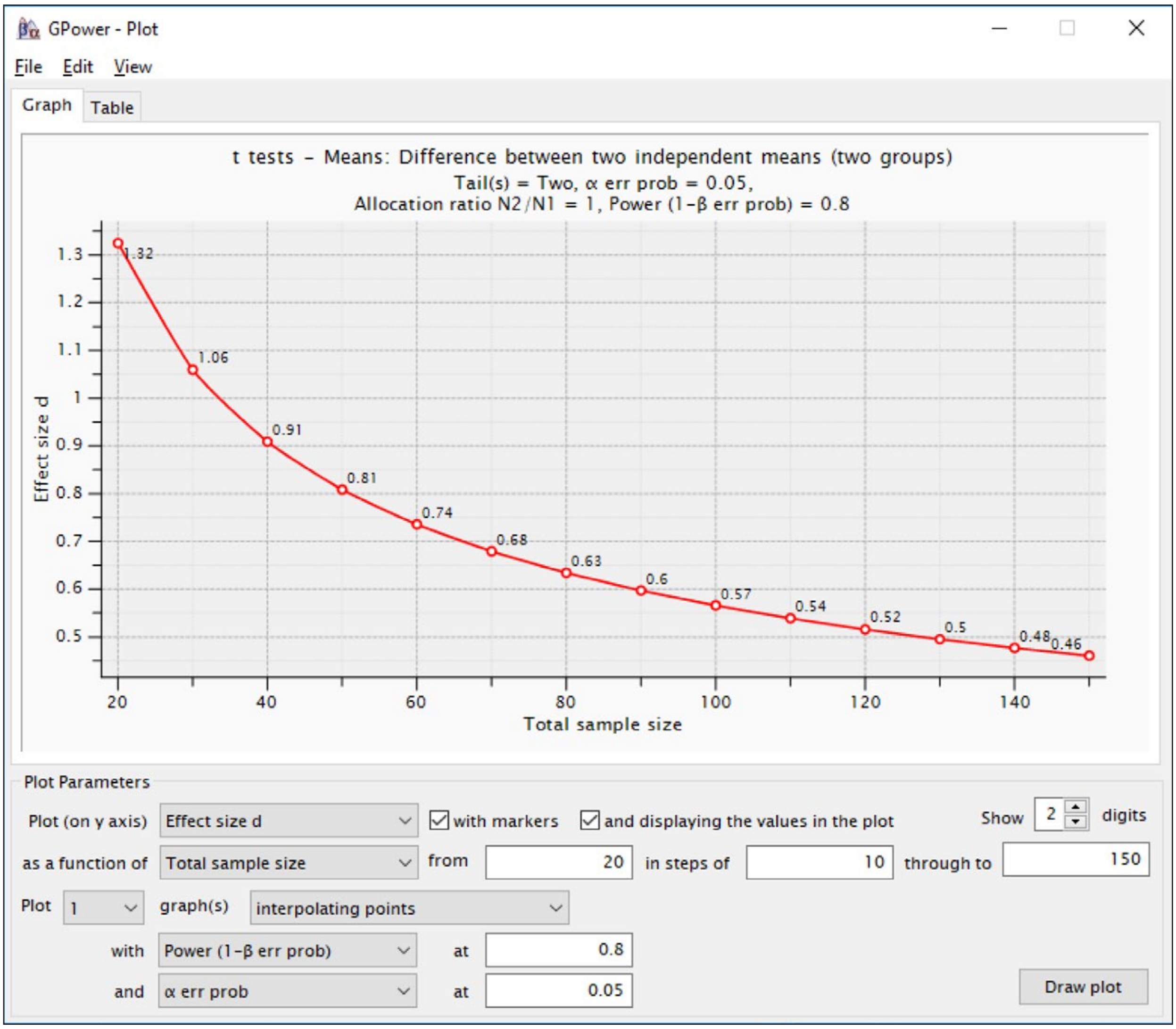
Noncentrality parameter δ: 2.8284271
Critical t: 1.9789706
Df: 126
Sample size group 1: 64
Sample size group 2: 64
Total sample size: 128
Actual power: 0.8014596

Figure 1: Main window of G*Power calculating the power of a two independent samples t-test.

input parameters are the effect size *d* (0.50), the direction of the test (two-tailed or one-tailed, in this example it is two-tailed), the α level (.05), and the expected power (.80). The last parameter, "Allocation ratio N2/N1" gives the possibility of analyzing unbalanced designs by specifying the ratio of the two planned group sizes. In the following sections, we assume that the planned designs feature equal-sized groups, but one can easily adjust many of the examples for unbalanced designs.

After setting the input parameters, hitting the calculate button results in the output filling in the output parameters. The required *N* in this case is 128, meaning that if one collected 128 cases divided into two groups of 64 participants each, drawing from a population where the exact Cohen's *d* is 0.50, in 80% of the cases one should expect the t-test to come out as statistically significant, fixing α = .05 (two-tailed). The top panel of **Figure 1** shows the t-distributions under the null-hypothesis (red curve, population d = 0) and the alternative hypothesis (blue curve, population d = 0.50). The shaded area under the blue curve indicated by β is the probability of obtaining a nonsignificant result (at α = .05 level); whereas, the nonshaded area under the blue curve is the probability of obtaining a significant result, the power of the test (1–β). The output reports some additional parameters: the noncentrality parameter delta, the critical t, and the df. The noncentrality parameter is (almost always, cf. Cohen, 1988) the standardized mean of the t-distribution of the estimates obtainable under the alternative hypothesis (d = 0.50), weighted by the size of the groups. Together with the df (degrees of freedom of the t-test) and the critical value, it may be useful in some advanced applications, such as computing the confidence limit of the effect size and conducting safeguard power analysis (Perugini et al, 2014). In simple applications, these indices are not usually of particular interest for the power analysis practitioner.

So far, we have obtained one estimation of the required *N*, assuming the effect size d = 0.50 is correct. To improve our ability to plan more powerful designs, we can explore more possibilities by conducting basic sensitivity analysis around the estimated required *N*. This can be accomplished by selecting "Sensitivity: […]" in "Type of analysis," plugging in the results we just obtained (α = .05, power = .80, sample size group 1 = 64, sample size group 2 = 64) and selecting "X-Y plot for a range of values." In the new window, one can plot different pairs of power analysis parameters and evaluate how each changes as a function of the other. An interesting pair (**Figure 2**) is the effect size change as a function of the total sample size (required *N*). In the example, the effect size d on the Y-axis indicates the lower bound of the set of sample effect sizes that would be statistically significant (with power .80) for each possible required *N* (total sample size). Here, one can appreciate how small increases in the sample size would not much change the minimum effect size that would result as significant, but decreasing the *N* becomes increasingly detrimental for the researcher's ability to detect an effect size significantly different from zero. For instance, dividing the required *N* by half (from 128 to 64) would result in sample effect sizes lower than (approximately) 0.63 being not significant, making the expected population effect size (d = 0.50) less likely to produce a significant result.



Plot Parameters

Plot (on y axis) Effect size d ☑ with markers ☑ and displaying the values in the plot Show 2 digits

as a function of Total sample size from 20 in steps of 10 through to 150

Plot 1 graph(s) interpolating points

with Power (1-β err prob) at 0.8

and α err prob at 0.05

Figure 2: Sensitivity Plot of G*Power calculating the power of a two independent samples t-test: Lowest detectable effect size as a function of required *N*.

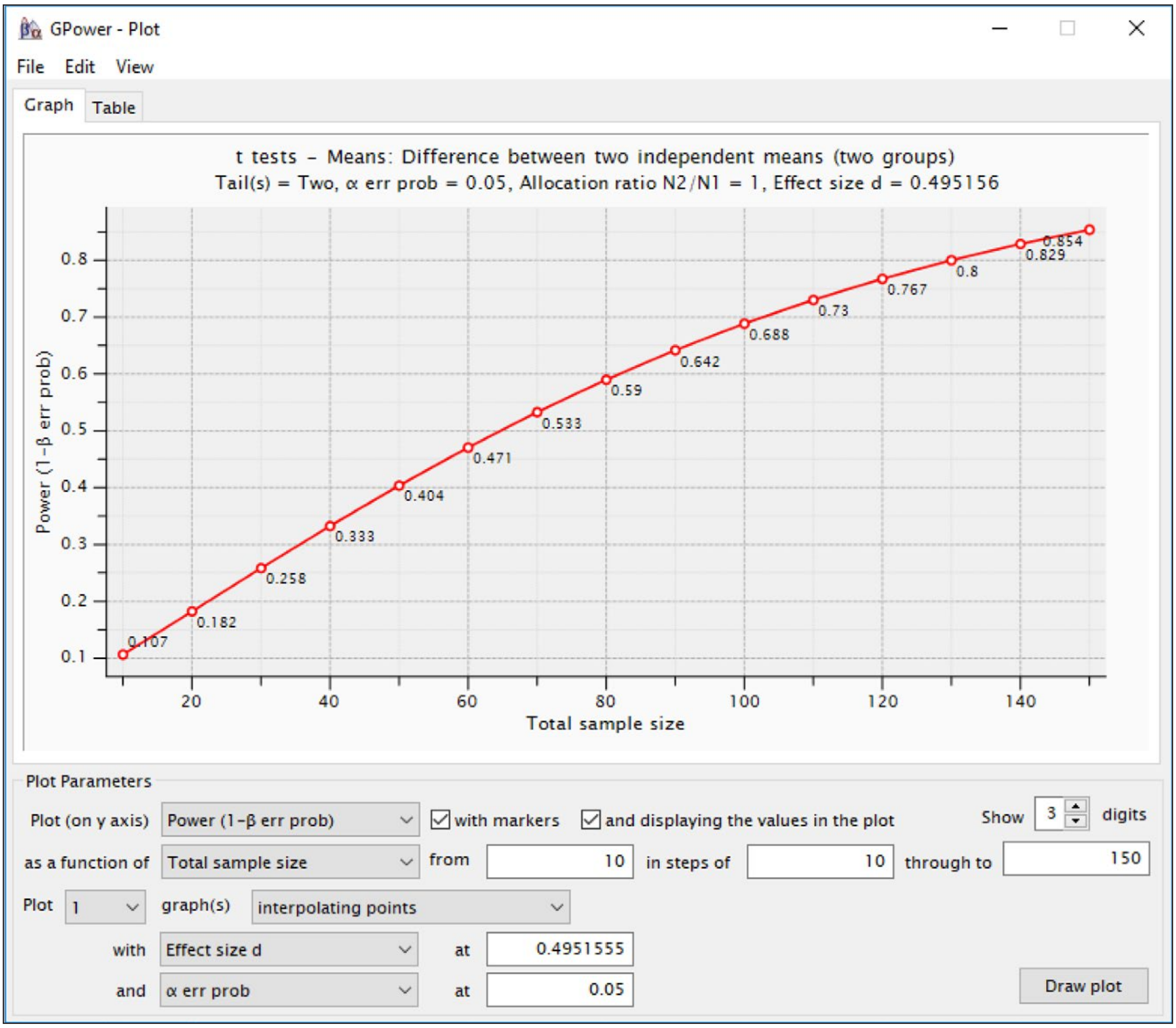
A similar insight can be derived by plotting the expected power on the Y-axis as a function of the sample size (**Figure 3**). In **Figure 3**, one can appreciate that given d = 0.50, the power of the test decreases to around .50 with a total sample size of $N = 64$ participants, thus giving the researcher only a 50% chance of finding a statistically significant result. In general, the risk of reducing the sample size, or the relative advantage of increasing it, can be evaluated using the sensitivity plots as in **Figures 2** and **3**.

G*Power also allows computing the effect size starting from the group means and their standard deviations (cf. the option "Determine" in **Figures 1** and **4**). In recent years, reporting an effect size index has become common practice in the published literature. Thus, the researcher seldom needs to input the raw means and standard deviation. Even when Cohen's d is not reported in the literature, a t-test is usually available. If t is the observed t-test value for a two-groups t-test on N participants, we can obtain the effect size as $d = \frac{2t}{\sqrt{N-2}}$.

A cautionary note is needed about the required total sample suggested in output by G*Power with designs involving groups and the actual required sample the researcher needs. It is often the case, as in many examples below, that one is planning a balanced design and G*Power prescribes a required total sample size that is not divisible by the number of planned groups. For instance, in a planned design with four groups, G*Power may yield a required $N = 34$, which is not evenly divisible by four. The solution is to round up the total sample size to the first whole number divisible by the number of groups, such as $N = 36$ in this example.

Basic power analysis with R

Equivalent results can be obtained using R. As is often the case in R, one can obtain the same result in different ways, so here we show some basic results that require commonly used R functions and minimal data transformation. In R *stats* package (installed on all R distribution by default) one can use the function `power.t.test()`, which allows specifying the four parameters of power computation: `n` (N for each group), `sig.level` (the α level), `power` (1–β), the effect size, and the `type` of test, whether a two-



Plot Parameters

Plot (on y axis) Power (1–β err prob) ☑ with markers ☑ and displaying the values in the plot Show 3 digits

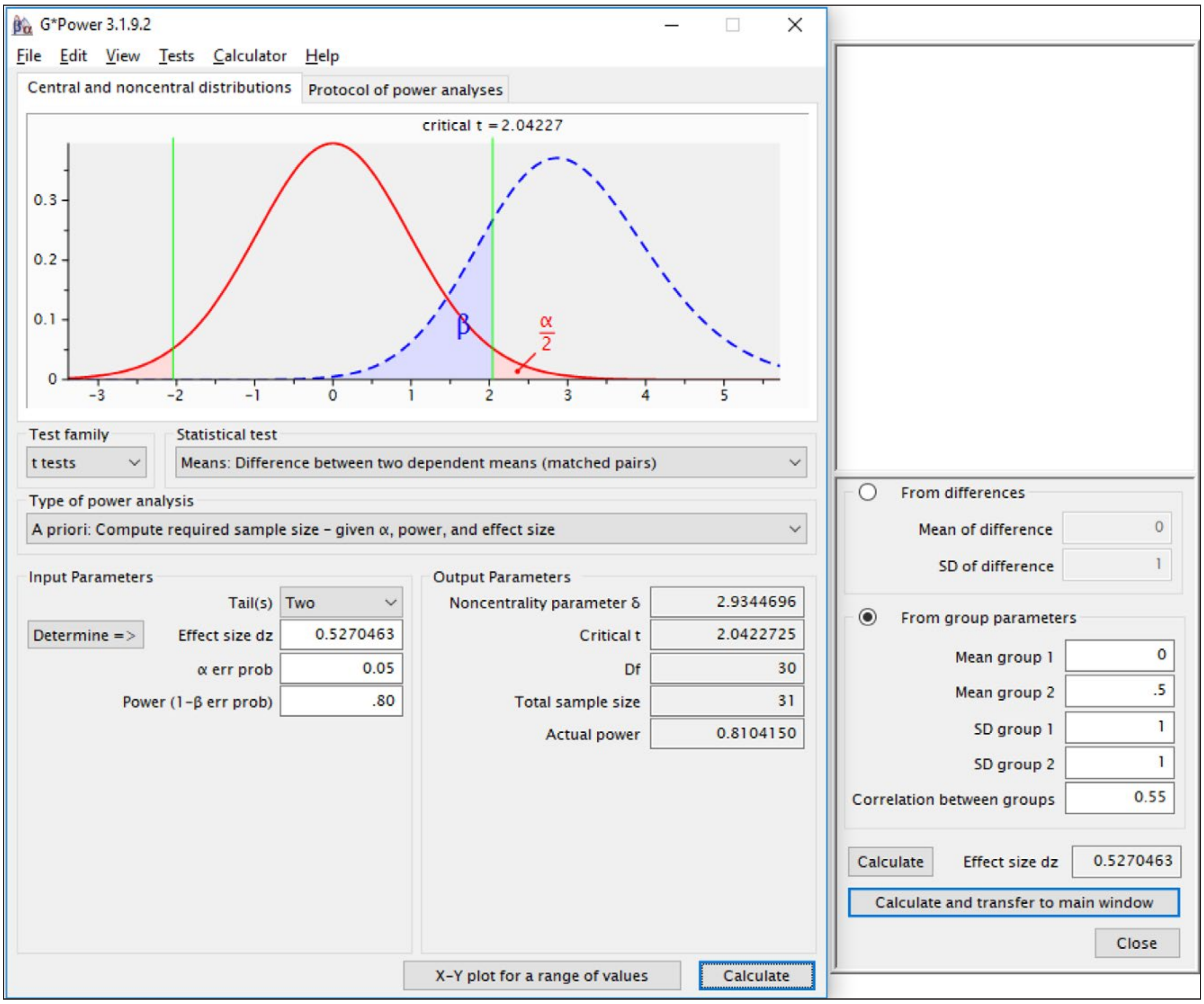
as a function of Total sample size from 10 in steps of 10 through to 150

Plot 1 graph(s) interpolating points

with Effect size d at 0.4951555

and α err prob at 0.05

Figure 3: Sensitivity Plot of G*Power calculating the power of a two independent samples t-test: Power as a function of required N for fixed effect size.


G*Power 3.1.9.2
File Edit View Tests Calculator Help
Central and noncentral distributions
Protocol of power analyses
critical t = 2.04227
0.3
0.2
0.1
0
−3
−2
−1
0
1
2
3
4
5
β
α/2
Determine =>
X-Y plot for a range of values
Calculate
Calculate
Calculate and transfer to main window
Close

Test family: t tests

Statistical test: Means: Difference between two dependent means (matched pairs)

Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

Input Parameters

- Tail(s): Two
- Effect size dz: 0.5270463
- α err prob: 0.05
- Power (1−β err prob): .80

Output Parameters

- Noncentrality parameter δ: 2.9344696
- Critical t: 2.0422725
- Df: 30
- Total sample size: 31
- Actual power: 0.8104150

○ From differences

- Mean of difference: 0
- SD of difference: 1

● From group parameters

- Mean group 1: 0
- Mean group 2: .5
- SD group 1: 1
- SD group 2: 1
- Correlation between groups: 0.55

Effect size dz: 0.5270463

Figure 4: Example of derivation of the effect size based on its constituent parameters: paired t-test case.

samples (the default), a one-sample, or a paired-samples t-test. One should specify the effect size by declaring the mean difference (`delta`) and the pooled standard deviation (`sd`). When Cohen's d is the effect size at hand, one should set the input parameter `delta` = d, and leave the `sd` parameter equal to 1. Optionally, one can specify the direction of the test with the `alternative` parameter.

Type of power analysis is simply decided by omitting the parameter that one desires to compute. a priori power analysis is achieved by omitting the parameter `n` and specifying α, power, and effect size. The line of R code:

```
power.t.test(power=.80,
             sig.level=.05,
             delta=.5, sd=1)
```

Produces the required N = 63.76, which can be rounded up to match G*Power results. Notice that the `power.t.test()` function returns the required participants per group, so the total sample size should be twice as large as the returned N.

Sensitivity analysis can be obtained by changing the obtained N and omitting the parameter we intend to study. For instance, repeatedly running:

```
power.t.test(n=n*,
             sig.level=.05,
             delta=.5, sd=1)
```

By changing n* in the vicinity of $N = 64$ can inform us of the change in expected `power` (notice that the parameter `power` is not set) as we change N. The same logic applies to retrospective power analysis. R offers several packages to run power parameters. All the worked examples in the paper are replicated in R in additional material available at https://github.com/mcfanda/primerPowerIRSP.

Also for R packages, a cautionary note is needed about the required total sample output by the software and the actual required sample the researcher needs. In R, the required sample is often not a whole number, so it may seems strange to require 63.76 participants per group. The solution is simply to round up the required N to the next integer.[3]

Paired groups (repeated measures)

Power analysis for two-cell repeated-measures designs is logically simple. A paired-sample t-test is simply a one-sample t-test on the difference score obtained by subtracting one repeated measure from the other. If we know the average of the difference score (Δ) and its standard deviation (*sd*), the effect size is given by $d_z = \Delta/sd$. This effect size is often called standardized difference score d_z (Cohen, 1988). In G*Power, one selects "Means: difference between two dependent means" in the Statistical Test field and plugs the numbers as in the two-sample t-test. The same logic applies for comparing one sample mean to a theoretical value (one-sample t-test), which yields the same results as the paired-sample t-test, provided that the effect size is the same.

More interesting is the case in which the expected effect size is not directly available, for instance, when previous studies have employed a between-subject design that we want to replicate in a within-subject design. When this is the case, recall that the standard deviation of the difference score depends on the variances of the repeated measures and their correlation. Thus, to transfer an independent groups Cohen's *d* into repeated measures d_z, the correlation ρ between measures has to be known or guessed. If ρ can be guessed, one can obtain the correct effect size by computing $d_z = \frac{d}{\sqrt{2\cdot(1-\rho)}}$. To illustrate, assume we observe the previous example $d = 0.50$ from a between-subject design, but we plan to employ a repeated measure design and expect the correlation between measures to be $\rho = .55$. The within-subject effect size will be $d_z = \frac{.50}{\sqrt{2(1-.55)}} = 0.527$, which G*Power associates with an expected $N = 31$ (less than one-fourth of the sample required for the corresponding between-subject design). G*Power also offers the possibility of running the effect size calculation with the option Determine (cf. **Figure 4**). The corresponding R code would be

```
power.t.test(delta = .527,
             sig.level = .05
             power = .80,
             type = "paired")
```

An important note of caution is in order about the effect size d_z in paired-samples t-tests. Not all authors use the standardized difference score effect size, yet they may refer to their effect sizes as Cohen's *d*. Different indices may yield dramatically different values; thus, it is important to be sure that d_z is used in power analysis software. To be sure, one can take the t-test reported in the article and check $d_z = \frac{t}{\sqrt{N}}$. If this is the case, *d* is the correct one. For more complex designs, such as factorial within-subjects designs and mixed designs, accessible introductions to power analysis can be found in Brysbaert and Stevens (2018) and Guo, Logan, Glueck and Muller (2013).

One-way analysis of variance

The ANOVA is a well-known strategy for analyzing data comparing more than two group means. Most power analysis software, including G*Power, use the *f* parameter as the measure of effect size (Cohen, 1988). The *f* effect size is the expected standard deviation of the group means divided by the pooled within-group standard deviation. However, the *f* parameter is neither intuitive, nor commonly used in published empirical research, so it may be convenient to use more popular effect sizes. A better choice is the eta-squared (η^2). The eta-squared is the proportion of the total variance explained by the means variance. The good news is that G*Power allows computing the *f* parameter starting from the η^2 (using the option Determine). The bad news is that G*Power, as any other power analysis software, requires the population η^2. This may not correspond to the sample eta-squared η_s^2, which is the effect size computed by several well-known statistical software programs, such as SPSS (G*Power 3.1 manual; Porter, 2017), and is the one commonly reported in published literature. The discrepancy is due to the fact that G*Power requires the ratio of population variances; whereas, SPSS eta-squared (η_s^2) is the sample-based estimation of η^2. The solution we suggest to estimate the population eta-squared (η^2) is to use epsilon-squared (Kelly, 1935), which has been shown to be less biased than both omega-squared (Hays, 1963) and eta-squared, the latter being the most biased estimator of the three (Okada, 2013).[4] Epsilon-squared can be easily computed starting from the sample eta-squared with a simple formula (cf. Eikeland, 1975):

$$\varepsilon^2 = 1 - \left(1 - \eta_s^2\right) \cdot \frac{N-1}{N-k} \tag{1}$$

Where *N* is the total sample size and *k* is the number of groups. Epsilon-squared can then be used as the value for the population eta-squared input in G*Power.[5]

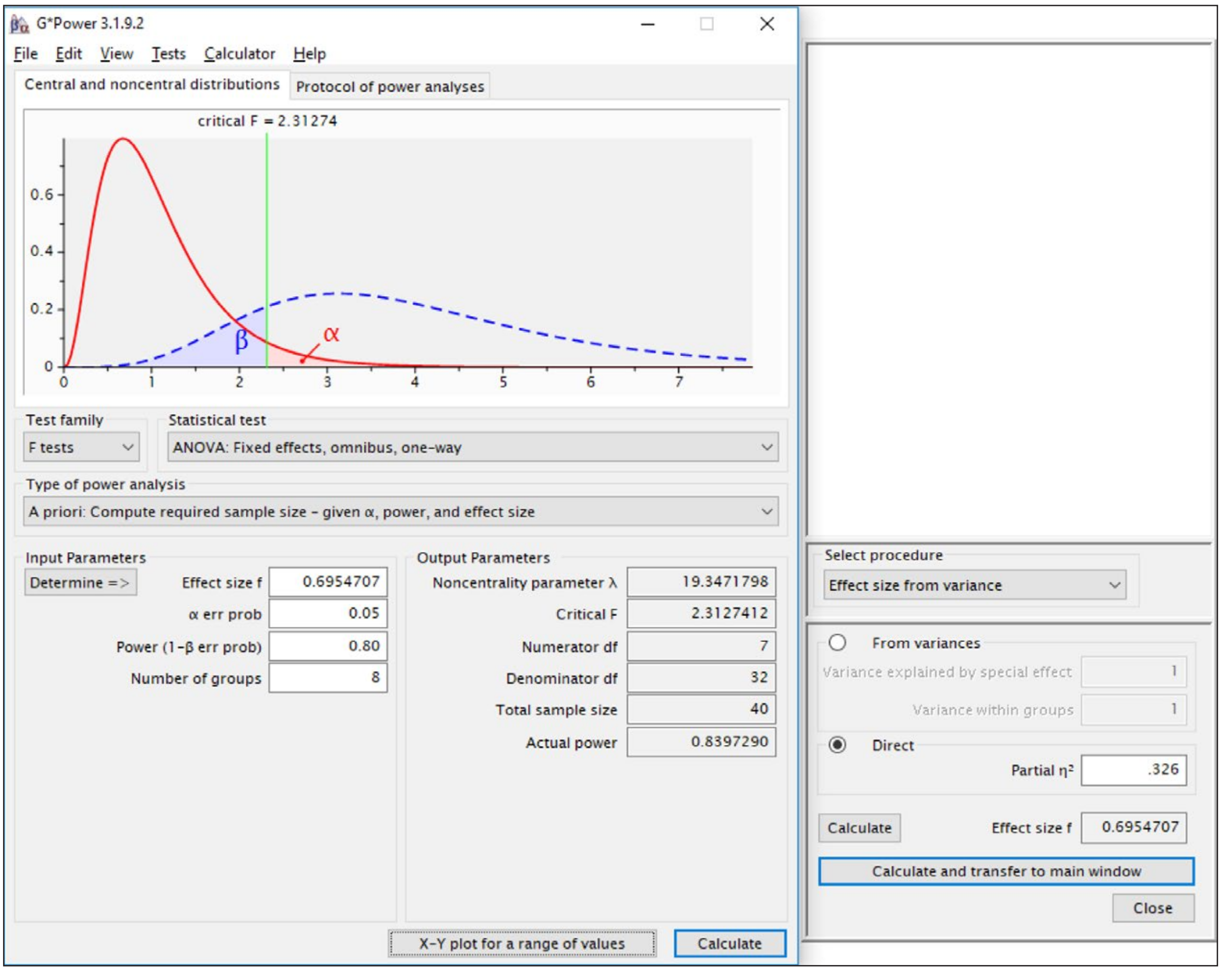
Assume that in the published literature we found research that obtained $\eta_s^2 = .35$ in a design with 200 participants divided in eight groups. We wish to compute the minimum required *N* (total sample size) for achieving a power of .80. Applying the formula in (1), we obtain $\varepsilon^2 = .326$. We can insert it into the direct panel in the Partial η^2 field (cf. **Figure 5**).

Note that in one-way designs the effect size is named η^2 but in G*Power, we find the field Partial η^2 because in factorial designs the partial eta-squared η_p^2 is used, and G*Power employs the more general term. In one-way designs, there is simply nothing to partial out (Richardson, 2011). We then ask the software to compute and transfer the computed *f* into the main window. The required *N* we obtain is 40, meaning that given the effect size, we require 5 participants per cell to attain a power of around .80.

For the previous example, we selected "F-test" in the Type of Test field and "ANOVA: Fixed effects, omnibus and one-way" in the Statistical Test field. For any other effect involved in the ANOVA, such as the effect estimated in factorial designs, we can select "ANOVA: Fixed effects – special, main effects and interactions."

Factorial designs

Power analysis for factorial designs can be obtained following the same steps as those for the one-way ANOVA, with some specifications. Because main effects and interactions are embedded into a larger design, one needs to specify the total number of groups in the design and the effect degrees of freedom to obtain the correct computation of the power parameters. Thus, in a 3 (A) × 2 (B) design, the computation of the required *N* to attain


G*Power 3.1.9.2
File Edit View Tests Calculator Help
Central and noncentral distributions
Protocol of power analyses
critical F = 2.31274
0.6
0.4
0.2
0
0 1 2 3 4 5 6 7
β
α
Determine =>
X-Y plot for a range of values
Calculate
Calculate
Calculate and transfer to main window
Close

Test family: F tests
Statistical test: ANOVA: Fixed effects, omnibus, one-way
Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

Input Parameters
Effect size f: 0.6954707
α err prob: 0.05
Power (1-β err prob): 0.80
Number of groups: 8

Output Parameters
Noncentrality parameter λ: 19.3471798
Critical F: 2.3127412
Numerator df: 7
Denominator df: 32
Total sample size: 40
Actual power: 0.8397290

Select procedure: Effect size from variance
○ From variances
Variance explained by special effect: 1
Variance within groups: 1
● Direct
Partial η²: .326
Effect size f: 0.6954707

Figure 5: One-way ANOVA computation of effect size indexes and power in G*Power.

.80 power for the main effect of A requires specifying $k - 1 = 2$ degrees of freedom for the effect (numerator df in G*Power), the total number of groups (3 × 2 = 6), and the effect size. In the same design, the power of the interaction A × B can be estimated by inserting (3 – 1)*(2 – 1) = 2 as the numerator *df* and 3 × 2 = 6 as the number of groups.

The effect size index that can be used is the partial eta-squared (η_p^2). The partial eta-squared is the variance explained by the effect (main effects or interactions) expressed as a proportion of the variance not explained by the other effects. Thus, if σ_f^2 is the population variance explained by the effect and σ^2 is the population residual variance, we have:

$$\eta_p^2 = \frac{\sigma_f^2}{\sigma_f^2 + \sigma^2} \tag{2}$$

As for the one-way design case, the expected effect size is the population effect size. Thus, the same considerations regarding its empirical estimates apply. In particular, one can adjust the sample eta-squared by computing the partial epsilon-squared as follows:

$$\epsilon_p^2 = 1 - \left(1 - \eta_p^2\right) \cdot \frac{N - K + df}{N - K} \tag{3}$$

Where *df* are the degrees of freedom of the effect and K is the total number of groups in the design. For instance, if one has a two by two design with a total sample of 20 participants and $\eta_p^2 = .20$, the formula yields:

$$\epsilon_p^2 = 1 - \left(1 - .20\right) \cdot \frac{20 - 4 + 1}{20 - 4} = 1 - .80 \cdot \frac{17}{16} = .15 \tag{4}$$

When the effect size can be computed from previous literature as partial epsilon-squared, it can be inserted into the Partial η^2 field. When the effect size is not available, it can be computed by guessing the variance explained and the residual variance, expressed as proportion, using the Determine option offered by G*Power. To do so, however, one also needs to guess the variance explained by the other factors in the design, because that variance influences the size of the residual variance. For example, assume that in a 2-factors design the researcher expects the interaction to explain around 10% of the variance. If the researcher expects no main effect, the proportion of residual variance is 1 – .10 = .90, so the $\eta_p^2 = .10$. If the researcher instead believes that a main effect will be found explaining 40% of the variance, the proportion of residual variance will be 1 – .10 – .40 = .50, and thus $\eta_p^2 = .166$. The error variance can also be deduced from previous research when the overall ANOVA R^2 is available. In fact, σ^2 is roughly equal to 1 – R^2. When the partial eta-squared is computed inserting in G*Power the expected proportions of variance, no transformation is needed and the software transfers the eta-squared estimates into the effect size *f* field.

The real difficulty with factorial designs and complex designs in general is obtaining a sensible guess of a reasonable effect size. Factorial designs, in fact, are often planned to expand and revise previous literature; thus, the effect size may not be available in the form required by the power computation. A researcher may observe, for instance, an effect in a one-way ANOVA and may wish to establish whether it replicates in two different conditions, thus aiming at a main effect in a factorial design. Another case can be that the observed one-way ANOVA effect is expected to be moderated by a factor that suppresses the effect in one condition and replicates it in another, thus aiming at an interaction in the factorial design (cf. Wahlsten, 1991). In all these conditions, estimating the appropriate effect size value may be challenging. One possible solution is to specify the expected pattern of means or take inspiration from observed results and decompose it into contrasts (Cohen et al., 2003).

Power analysis for contrasts

When prospective power analysis is applied to contrast analysis, one can legitimately speak of contrasts regarding planned comparisons. A planned comparison is specified by assigning weights to the expected means and testing the weighted sum of the means against the mean of the weights. Because proper contrasts have means equal to zero, the contrast is tested against zero. Assume one is expecting a pattern of means across four groups equal to {10, 0, 0, 0}, with the within-group standard deviation (σ) equal to 5. To test any hypothesis across these four means, one can specify a set of weights c_i, with sum equal to zero, that compares the desired set of means. One can, for instance, compare the first mean against the other three by specifying a contrast, such as c = {3, −1, −1, −1}, or the first two against the second two means by specifying a contrast c = {1, 1, −1−1}.

Contrasts can also be useful in complex designs because they allow testing main effects and interactions effects, or a subset of those, by guessing the pattern of means (and the within-group variability) or observing it in a different design published in the literature. The pattern of means {10, 0, 0, 0}, for instance, may be expected in a 2 × 2 design similar to the ones in **Table 2**.

In this design, the main effect of B can be tested with a contrast (going row-wise) c_B = {1, 1, −1, −1}, the main effect of A with c_A = {1,−1, 1, −1}, and the interaction with c_{AB} = {1, −1, −1, 1}.

Table 2: Example of 2 × 2 design.

	A1	A2
B1	10	0
B2	0	0

Note: Pooled standard deviation is equal to 5.

In between-subjects designs, contrasts are tested with the F-test with 1 and *N-k* degrees of freedom (*k* equal to the number of groups). Thus, one can employ ther G*Power "ANOVA: Fixed effects – special, main effects and interactions" statistical test to estimate the power parameters. If we name μ_i the expected means and c_i the corresponding contrast weights, the effect size *f* for a contrast is

$$f = \frac{\left|\sum c_i \cdot \mu_i\right|}{\sqrt{k \cdot \sum c_i^2 \cdot \sigma^2}} \tag{5}$$

where $\sum c_i \cdot \mu_i$ is the contrast value. In our example, main effects in **Table 2** yield the same $f = \frac{10}{\sqrt{4 \cdot 4 \cdot 25}} = .50$. Plugging the *f* in G*Power for a power = .80 suggests a required *N* = 34 (total sample), which can be approximated with 9 participants per cell, yielding a required total sample of *N* = 36. For the interaction contrast c_{AB}={1, −1, −1, 1}, the *f* parameter is the same; thus, one expects to achieve the same power for the interaction with the required 9 participants per cell computed for the main effects. R code and an Excel file to help with calculations for contrasts, as well as for interaction, moderation, and mediation effects (see below), are available online at https://github.com/mcfanda/primerPowerIRSP.

It is crucial to realize that the power associated with a contrast depends, given a fixed α level, exclusively on the expected effect size *f* and the degrees of freedom of the test. However, given a certain contrast value, the corresponding expected effect size can dramatically change depending on the design one is planning to analyze. These properties of contrast analysis make it easy to use power analysis software, when an effect size is correctly anticipated, because the software commands and the interpretation of the results will always be the same given a certain effect size. However, adapting the correct effect size of a contrast value to the planned research design might be challenging. This issue is particularly important when results from one design are used to compute the power parameters of different, larger designs. We now consider this issue in more detail.

Guessing the interaction effect size from one-way designs

A common case in experimental psychology is observing an effect of a factor in a one-way design and planning a larger design where a moderator variable is included. For example, consider Case 1 in **Table 3**, in which the pattern of means is taken from a one-way design where A1 and A2

Table 3: Example of 2 × 2 design expected results.

		Case 1		Case 2	
		A1	A2	A1	A2
B	replicated	5	2	5	2
	moderated	0	0	2	5

Note: Pooled standard deviation is equal to 1.

show means equal to 5 and 2, respectively, and the same within groups variability (say equal to 1). The researcher observes in the literature a one-way design with only A as a factor and wishes to test the moderating effect of B in a 2 × 2 design. The B factor has two levels, which, for simplicity, we name replicated and moderated. The problem is determining the power parameters of the expected interaction effect.

This problem has raised much interest in the methodology community (McClelland & Judd, 1993) and has recently caught the attention of several commentators (Gelman, 2018; Giner-Sorolla, 2018), although different solutions have been proposed. We suggest that in these situations a contrast approach can solve many difficulties that accompany power analysis. Here is a step-by-step example followed by a general simple solution.

The first step is to compute the contrast value for the observed design, $C_A = (1) \cdot 5 + (-1) \cdot 2 = 3$, and its effect size, $f = \frac{3}{\sqrt{2 \cdot 2 \cdot 1}} = 1.50$. The second step is to try to anticipate how the moderator will change the observed effect. A simple case is that the researcher expects the moderator to suppress the effect completely, as shown in **Table 3**, Case 1. In this case, the anticipated interaction contrast value is still $C_{AB} = (1) \cdot 5 + (-1) \cdot 2 + (-1) \cdot 0 + (1) \cdot 0 = 3$, but the effect size is now $f = \frac{3}{\sqrt{4 \cdot 4 \cdot 1}} = 0.75$; thus, it is half the original effect size, and it will require (almost) double the size of each design cell to achieve the same power of the one-way design. These results led commentators to suggest doubling the cell size in case of interactions (Simonsohn, 2014) or even increasing the sample size by higher multipliers (Gelman, 2018) when interactions are involved. However, a simple sample size multiplier would not work in the general case, because the expected effect size depends on the shape of the interaction one is expecting.

Consider Case 2 in **Table 3**, in which the researcher is expecting the moderator to reverse the effect, creating a crossover interaction. In this case, the contrast would result in $C_{AB} = (1) \cdot 5 + (-1) \cdot 2 + (-1) \cdot -2 + (1) \cdot 5 = 6$ and the effect size will be $f = \frac{6}{\sqrt{4 \cdot 4 \cdot 1}} = 1.50$, exactly as in the one-way design. To achieve the same power of the original one-way design, one would need the same cell size of the original study and no multiplier of the sample size would be required.

A simple and general way to anticipate an interaction effect size starting from a one-way design is to think in terms of percentage of moderation (p_m). The researcher has to anticipate the percentage of expected change of the original effect, with 100% of change indicating a complete suppression of the effect and 200% indicating a complete reverse of the effect. When this percentage of moderation can be anticipated, one computes the expected effect size using the following formula:

$$f_n = \frac{p_m}{100} \cdot f_o \cdot \sqrt{\frac{k_o}{k_n \cdot l}} \tag{6}$$

where f_n is the expected effect size of the interaction for the planned research, f_o is the observed effect size of the original effect, k_o and k_n are the number of cells in the original and the planned design, respectively, and l is the number of levels of the moderator.

In the 2 × 2 design of our examples, the calculation of the expected interaction effect size simplifies to:

$$f_n = \frac{p_m}{100} \cdot f_o \cdot \frac{1}{2} \tag{7}$$

Thus, if the researcher expects a complete suppression of the effect (**Table 3**, Case 1), $p_m = 100\%$ and $f_n = 1.50 \cdot \frac{1}{2} = 0.75$ whereas, if the researcher expects a complete reverse of the effect (**Table 3**, Case 2), the expected effect size will be $f_n = 2 \cdot \frac{3}{2} \cdot \frac{1}{2} = 1.50$, as we have shown before.

This approach makes it easy to evaluate scenarios where the expectations are not clear-cut. Assume the researcher is planning research where a moderator is expected to suppress the effect but not reduce it to zero. The researcher may, for instance, expect a reduction of 50% of the effect. Applying the logic of the proportion of moderation, it is easy to estimate that the expected interaction effect size will be $f_n = 0.50 \cdot 1.50 \cdot \frac{1}{2} = .375$, with no need to anticipate the exact expected means for the moderator levels or the standard deviations of the cells. Power parameters can then be computed using the estimated interaction effect size.

An example of a more complex design

Consider a researcher who wishes to design a moderation study based on a one-way design with four conditions implementing an increasing intensity of a stimulus, such that the observed pattern of means shows a linear trend. In particular, the observed linear trend contrast has an $\eta_p = .184$, corresponding to a $f = 0.475$ (cf. G*Power manual, p. 29). The observed pattern of means is shown in **Figure 6**, as the Replicated mean pattern.

Had the observed study been conducted with 10 participants per cell, it would have a power slightly higher than .80. Assume that the researcher expects the moderator, featuring two conditions, to replicate the observed linear trend in one condition (Replicated in **Figure 6**) and to reverse it in the other condition (Moderated in **Figure 6**). However, the reverse is not expected to be complete, only weak.

Employing the percentage of moderation approach, one needs to estimate only the percentage of change, keeping in mind that 100% means suppression of the effect and 200% means a complete reverse. Thus, under the scenario above, one can guess that the expected mild reverse would correspond to 125% of moderation. Because the original design has four cells and the planned design has eight, the expected interaction effect size is:

$$f_n = 1.25 \cdot .475 \cdot \sqrt{\frac{4}{8 \cdot 2}} = 0.296$$

Plugging the *f* into G*Power "ANOVA: Fixed effects – special, main effects and interactions," with input $\alpha = .05$, power = .80, numerator *df* = 1, and number of groups 8, G*Power suggests a required total sample of *N* = 92, which can be rounded to 12 participants per cell, yielding a required total sample of *N* = 96. Checking calculations of the *f* based

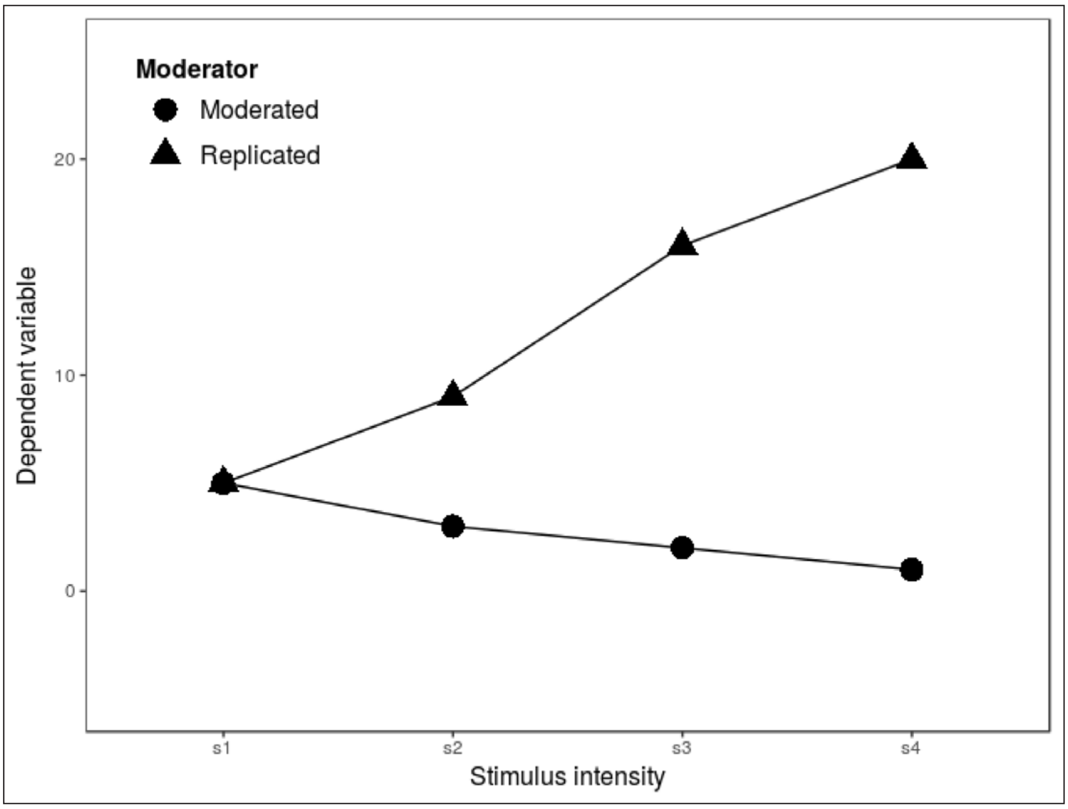


Figure 6: Example of a 4 × 2 expected interaction based on a one-way observed pattern of means.

on expected means and variances yields the same results as applying formula (6) on the original effect size.

Regression analysis

From a statistical point of view, power analysis for regression is based on the same logic and the same parameters as power analysis for ANOVA. Regression parameters are tested through an F-test, and the elective effect size index is f^2 (Cohen, 1988). In order to estimate the power parameters, one needs to calculate the effect size f^2 from the effect size indices commonly reported in published papers. The partial eta-squared is again the easiest index to employ, and it is defined exactly as in the ANOVA. This should not be surprising, because both regression analysis and ANOVA are applications of the general linear model; thus, the same inferential tests and the same effect size indices are available. Not every software package allows computing the partial eta-squared within the regression model. SPSS, for instance, computes the partial eta-squared for regression effects within the general linear model command, but in the regression command this option is not available. Fortunately, the partial correlation squared is equal to the partial eta-squared, only presented with a different name, so the partial correlation squared can be used as an estimate of eta-squared. In G*Power, the power parameters of any effect in multiple regression can be computed employing “F test: Multiple Regression – Fixed model, R^2 increase” command, letting the software compute f^2 based on the Partial R^2. Notice that in the interface of the command the eta-squared is named Partial R^2 because the partial R^2 generalizes the eta-squared to a set of variables. Nonetheless, the command can be used for one independent variable, and the expected eta-squared can be input in the Partial R^2 field.

When published data are not available, one should rely on guessing the proportion of variance explained by the effect and the residual variance, as we have seen in the ANOVA examples. Otherwise, one can use Cohen’s guidelines (Cohen, 1988) and guess whether the effect under investigation may be small ($f^2 = 0.02$), medium ($f^2 = 0.15$), or large ($f^2 = 0.35$).

Moderated regression

For moderated regression, namely a regression with an interaction involving at least one continuous variable, the effect size can be computed as for any other effect in the linear model. If the literature describes a similar regression, one can use the eta-squared of the observed interaction and follow the steps described for the multiple regression.[7] However, interactions are somehow special terms, because the variance of the effect depends on the variance of two predictors, rather than one predictor as for the linear terms (Jaccard & Turrisi, 2003). This extra variance (McClelland & Judd, 1993) impacts the power of the F-test associated with the interaction. Furthermore, although OLS regression assumes the predictors are error-free, they are typically measured with error, which decreases their reliability and negatively impacts the power of the test associated with the predictors’ interaction (Cohen et al., 2003). Despite these difficulties, reasonable approximations of the power function of interaction effects have been suggested (Shieh, 2009), but they are not readily available in G*Power.

Nonetheless, the general recommendation deduced from the relevant literature is that interaction effects tend to be less powerful than linear effects with the same effect size index. This general advice can be taken into account by setting a lower effect size than expected or by performing a thorough sensitivity analysis to understand the range of sample sizes that would guarantee sufficiently high power even if the expected effect size overestimates the variance explained by the interaction.

Even considering this advice, researchers planning to estimate a moderated regression face the difficulty of anticipating the interaction effect size. When the interaction effect size cannot be retrieved from the literature, the partial eta-squared is admittedly not the easiest effect size index to guess. Interactions often explain little variance, although they may reflect crucial effects for proving or disproving a theory. Our suggestion is to reason in terms of standardized regression coefficients (*beta*) and to obtain, under some constraining assumptions, a rough but reasonable estimate of the partial eta-squared based on the *beta's*.

Standardized coefficients are regression coefficients computed using standardized variables; thus, their interpretation can be based on the standard deviation scale. In simple regression, for instance, a standardized coefficient equal to β indicates that the expected value of the dependent variable increases β standard deviations as one increases the predictor of one standard deviation. They share the same scale as the Pearson correlation, although in multiple regression they are neither correlations nor partial correlations. The interesting fact in moderated regression is that the standardized coefficient associated with the interaction indicates the difference between the effect of one predictor computed for two consecutive units of the moderator. If the moderator is dichotomous, for instance, the interaction coefficient is the difference between the predictor coefficients computed in the two groups defined by the moderator (cf. **Figure 7a**). If the moderator is continuous, the interaction is the difference of the predictor coefficients computed for the moderator equal to its mean and the moderator equal to one standard deviation above (or below) the mean (cf. **Figure 7b**).

An exact estimation of the variance explained by the interaction is complex and requires several pieces of information only available in a published study. In the absence of this information, one can get a rough but reasonable estimation of the expected interaction coefficient and, in specific circumstances, infer the variance required to compute power parameters. We consider two cases below.

Interaction between continuous and dichotomous predictors

It is not uncommon that a researcher observes a relationship between two continuous variables and wishes to plan new research in which a dichotomous variable is supposed to moderate the original relationship. The original relationship could be tested in two different experimental conditions or for two classes of individuals representing two levels of a categorical variable.

The moderated regression applicable to this case is a regression featuring the linear effect of the predictor and the dichotomous moderator and their interaction. The standardized coefficient associated with the interaction is the difference in regression coefficients between the two levels of the moderator (i.e., the two groups defined by the moderator). The logic of the reduction of moderation can be applied here. Consider the following example: Imagine published research determines that Y and X are mildly related (r = .25) in the population because it is an average

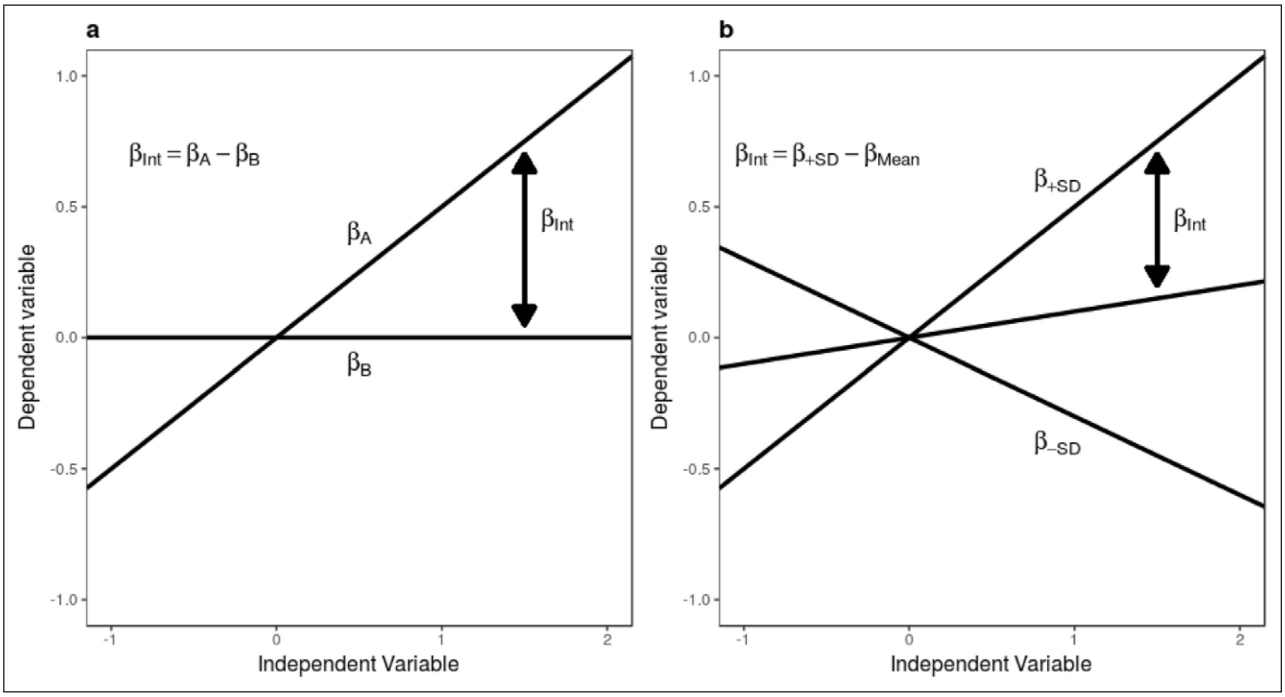
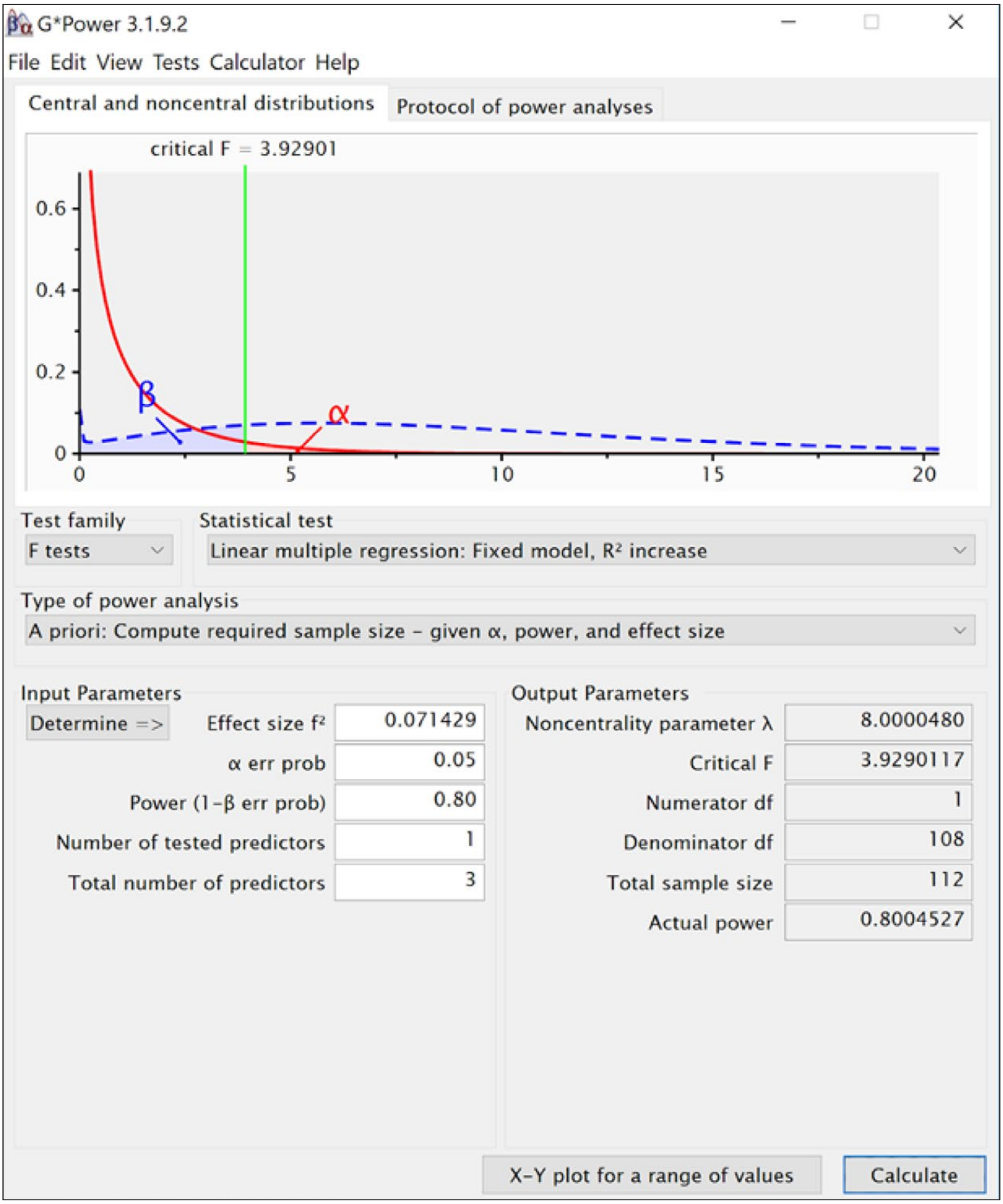


Figure 7: Geometrical interpretation of the interaction beta coefficient, with a dichotomous moderator (a) and a continuous moderator (b).

between a high positive value expected under the experimental condition (r_a = .50) and no correlation expected under the control condition (r_b = .00). The expected interaction coefficient in a planned experiment is simply β_{int} = $|r_a - r_b|$, that is $|.50 - .00|$ = $|.50|$. The question is now how to determine the variance explained by that effect. If one assumes that the two groups have equal size and they are not different in the X variable and in the Y variable, the effect size of the interaction can be approximated as follows:

$$f^2 \approx \frac{\beta_{int}^2}{2\cdot\left(2-r_a^2-r_b^2\right)} \tag{8}$$

Note that the formula is an approximation of the effect size based on the independence of the dichotomous moderator to the other variables in the model. If the moderator has a main effect on the dependent variable, the power analysis based on the f^2 will underestimate the power of the test; whereas, if the moderator is correlated with the


G*Power 3.1.9.2
File Edit View Tests Calculator Help
Central and noncentral distributions
Protocol of power analyses
critical F = 3.92901
0.6
0.4
0.2
0
0
5
10
15
20
β
α
Determine =>
X-Y plot for a range of values
Calculate

Test family: F tests

Statistical test: Linear multiple regression: Fixed model, R² increase

Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

Input Parameters

Effect size f² 0.071429

α err prob 0.05

Power (1-β err prob) 0.80

Number of tested predictors 1

Total number of predictors 3

Output Parameters

Noncentrality parameter λ 8.0000480

Critical F 3.9290117

Numerator df 1

Denominator df 108

Total sample size 112

Actual power 0.8004527

Figure 8: Example of power analysis for interaction in moderated regression.

predictor (i.e., the two groups defined by the moderator differ in the predictor means), the power will be overestimated.

In our example, we obtain $f^2 = 0.0714$.[6] One can now use the G*Power "F test: Multiple Regression – Fixed model, R^2 increase" command as shown in **Figure 8**.

The G*Power command requires the f^2; the α and the required power; the number of predictors (i.e., the number of coefficients tested), in our case 1; and the total number of predictors (i.e., the total number of coefficients in the model), in our case 3, for two main effects and one interaction. The resulting required total sample size is 112.

In G*Power, similar results can be achieved using t-tests—"Linear bivariate regression: Two groups, difference between slopes," although the underlying test employed to compare the slopes is different (Armitage et al., 2002) and the interface is peculiar. Thus, we suggest using the G*Power command described above, because it is the same command used for any other power analysis regarding multiple regression.[7]

Interaction between continuous predictors

As in the previous example, the power parameters for the test of the interaction effect with two continuous variables are easy to compute if one has an empirical estimation of the partial eta-squared inferred from the literature. This estimation can be plugged into the "F test: Multiple Regression – Fixed model, R^2 increase" command, and the software can compute the f^2 effect size, then input the α, the required power, and 1 for number of predictors and 3 for total number of predictors, and the power parameters will be obtained.

When the eta-squared is not available, one can obtain a reasonable estimate of the effect size by following a similar logic applied to the continuous by dichotomous interaction. In this case, however, it is necessary to anticipate the variance explained by the main effects. Furthermore, it is necessary to assume that the two independent variables are uncorrelated with each other or only mildly correlated. When the latter assumption cannot be met, the following procedure will overestimate the power of the interaction test, with overestimation being proportional to the correlation between predictors.

Under those assumptions, the interaction effect size can be computed following these steps. Assume *y*, *x*, and *m* are continuous variables, and we wish to compute the power of the test associated with the interaction *x*m*. Suppose *y* is time running on a treadmill, *x* is age, and *m* is hours of training per week (see a similar example in Cohen et al, 2003). The goal of the study is to ascertain how training moderates the relationship between age and endurance. The first step is to estimate the correlation between endurance and age, r_{yx}, and the correlation between endurance and training, r_{ym}. This can be inferred from research papers or other sources. Assume from previous research $r_{yx} = .35$ and $r_{ym} = .25$. Because we designated training as the moderator, we can think of r_{yx} as the average correlation between endurance and age, that is the correlation between endurance and age for the average level of training in the sample.

The second step is to apply a proportion of moderation logic similar to the continuous by dichotomous interaction. The relevant question is how much do we expect the correlation between endurance and age, r_{yx}, to increase if we compare participants with an average level of training and participants who train one standard deviation above average? More generally, how much do we expect the correlation between the predictor and the criterion to increase for a one standard deviation increase in the moderator?

Framed in this way, the question is easier to answer than estimating an eta-squared without other available information. One can say, for instance, that one standard deviation in training may increase the correlation about 50%, making the expected correlation for people that train above average equal to .525, with an increase equal to .175. The expected coefficient of the interaction will then be $\beta_{int} = .175$. Another way to anticipate the interaction beta is to guess, or to estimate from the literature, the expected correlation at high (or low) levels of the moderator and take the difference between the expected correlation for high levels and the expected correlation for average levels of the moderator. This difference will be the expected β_{int}.

The last step is to compute the effect size f^2. Under the described assumptions, the following formula gives an approximation of it:

$$f^2 \approx \frac{\beta_{int}^2}{1 - r_{yx}^2 - r_{ym}^2} \tag{9}$$

In our example, we obtain $f^2 = 0.0375$. Plugging this into G*Power "F test: Multiple Regression – special (increase of R^2), fixed model" yields a required total sample of $N = 212$ participants.

To recap, when previous research provides estimates of the eta-squared of the planned interaction, one can use it to compute the power parameters required. When the information is not available, the researcher needs to anticipate not only the correlations between the predictors and the criterion but also the difference in correlation of one predictor between the average effect and the effect expected at one standard deviation above the average of the moderator. Formula (9) gives a shortcut to compute the effect size required for the power analysis. Although the suggested method is more precise than a simplistic estimation of the effect size based on small, medium, or large categorization, it is nonetheless a rough estimation based on specific assumptions.[8]

Analysis of Covariance and Other Applications of the General Linear Model

Once one is capable of estimating the power parameters for regression and ANOVA, one can apply the same reasoning and follow the same practical steps for any other application of the general linear model. Analysis of covariance (ANCOVA), for instance, poses no particular challenge. When the researcher can guess the variance explained by the effect under investigation and the variance explained by the covariates, the eta-squared can be computed and the effect size f^2 can be estimated accordingly. From a practical point of view, G*Power offers the "ANCOVA: fixed effects, main effects and interaction," which employs the same statistical functions of the factorial ANOVA but allows specifying the number of covariates. This is useful for imputing the correct degrees of freedom of the F-test under investigation.

Mediation analysis

An analytic solution to power analysis has not been worked out for all possible models. A good example is provided by mediation analysis (Baron & Kenny, 1986), for which analytic solutions are available only under restrictive assumptions. In mediation analysis (**Figure 9**), a researcher tests whether the relationship between an independent variable X and a dependent variable Y can be explained by the effect of a third variable M, called the mediator. The total effect of X on Y is thus decomposed into two elements: the direct effect of X on Y and the indirect effect of X on Y through M. The direct effect is estimated by path *c* in **Figure 9**, the multiple regression coefficient of Y on X, controlling for M. The indirect effect is estimated by the product of *a***b*, where *a* is the simple regression coefficient of M on X and *b* is the multiple regression coefficient of Y on M, controlling for X (Baron & Kenny, 1986).

Under the assumption that the sampling distribution of the indirect *a***b* is normal, the standard errors, p-values, confidence intervals for the indirect effect can be computed using the Sobel test (Sobel, 1982). In this case, power analysis for testing the indirect effect can be performed analytically, using functions `ssMediation.Sobel()` and `powerMediation.Sobel()` in the R package *powerMediation* (Qiu, 2017).

Given a set of model parameters and an α level (parameter `alpha`, which is equal to .05 by default), the first function allows determining the sample size to achieve a certain power (argument `power` in `ssMediation.Sobel`), whereas, the second function allows determining the power achieved with a certain sample size (argument `n` in `powerMediation.Sobel`). For both functions, one has to specify model parameters. This step is analogous to specifying an effect size for the general linear model but is slightly more complex for mediation because of the larger number of parameters and because different programs require different specifications. In the case of the R package *powerMediation*, model parameters are specified through the following arguments:

- `theta.1a`, which is equivalent to *a* in **Figure 8**,
- `lambda.a`, which is equivalent to *b* in **Figure 8**,
- `sigma.x` (σ_x) and `sigma.m` (σ_m), which are respectively the standard deviations of X and M, and
- `sigma.epsilon` (σ_ε), which is the standard deviation of the error term in the multiple regression, in which Y is predicted by both X and M, $Y_i = \beta_1 X_i + \beta_2 M_i + \varepsilon_i$, $\varepsilon_i \sim N(0, \sigma_\varepsilon)$. Because $R^2 = 1 - \sigma_\varepsilon^2$, σ_ε can be simply computed as $\sqrt{1-R^2}$.

Typically, one has no specific idea of the standard deviations of X and Y or of σ_ε Thus, it is easier to think of parameters *a*, *b*, and *c* in terms of standardized regression coefficients. If all variables are standardized, parameter *a* (`theta.1a`) is simply the Pearson's correlation between X and M, whereas, *b* (`lambda.1a`) and *c* are the standardized multiple regression coefficients of Y on M and X, respectively. To specify that the coefficients refer to standardized variables, one must set `sigma.x = sigma.m = 1` and $\sigma_\varepsilon = \sqrt{1-\left(b^2+c^2+2abc\right)}$. This last equation can be derived from the definition of σ_ε by applying the properties of variance.

Preacher and Hayes (2004) report an example in which the standardized *a*, *b*, and *c* coefficients are *a* = .8186, *b* = .4039, and *c* = .4334. We will consider these values for our examples throughout this section. The following code computes the sample size for obtaining 80% power in a Sobel test assuming the same coefficients, which results in a sample of 57 participants. Notice that the value of σ_ε = .6020 can be easily computed from *a*, *b*, and *c* using the formula provided above.

```
ssMediation.Sobel(power = .80,
               theta.1a = .8186,
               lambda.a = .4039,
                sigma.x = 1,
                sigma.m = 1,
          sigma.epsilon = .6020)
```

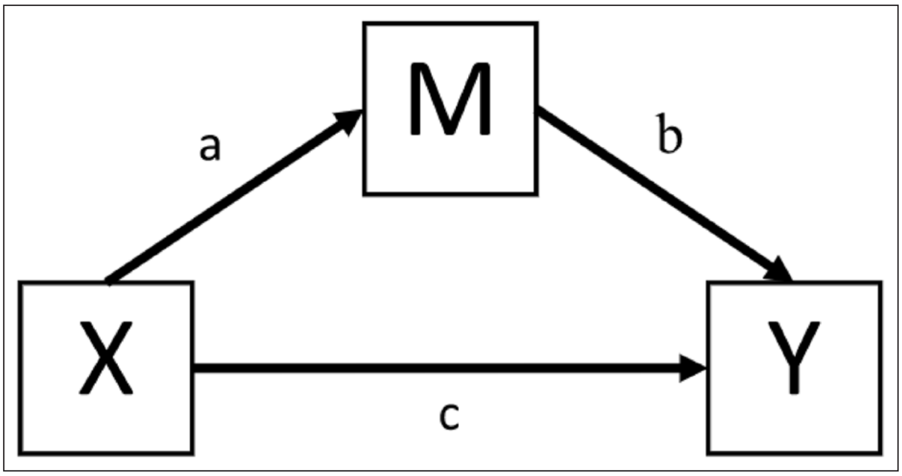


Figure 9: Mediation model.

The following code allows computing the power achieved from a sample size N = 100 for testing the same model, which results in power = .96.

```
powerMediation.Sobel(n = 100,
              theta.1a = .8186,
              lambda.a = .4039,
               sigma.x = 1,
               sigma.m = 1,
         sigma.epsilon = .6020)
```

The assumption that the indirect effect could normally be distributed has been criticized; therefore, the indirect effect $a*b$ is often tested not only through the Sobel test but also using bootstrap confidence intervals, which do not depend on the normality assumption (Hayes & Scharkow, 2013; Preacher & Hayes, 2004). If the significance of the indirect effect is assessed through bootstrap confidence intervals, analytic formulae are not available. In this case, power can be estimated using Monte Carlo methods (Schoemann, Boulton & Short, 2017; Thoemmes, MacKinnon & Reiser, 2010; Zhang, 2014). The general idea is simple: If power is the probability of rejecting H0 if H1 is true, one can determine power by (1) defining the expected values of the population parameters (e.g., *a*, *b*, and *c*) under H1, (2) generating a sample size N from the population parameters, (3) testing the significance of the target effect (e.g., the indirect effect) using the preferred method (e.g., bootstrap confidence intervals), (4) repeating steps 2 and 3 a large number of times, and (5) estimating power as the proportion of R simulated samples in which H0 is rejected (Zhang, 2014). This strategy has several advantages: it can be used for estimating power not only for the indirect effect but also for any model parameter; it can accommodate for specific data characteristics, such as nonnormality and missing values; and it yields more accurate results than other methods. However, Monte Carlo methods have some drawbacks: They can be computationally cumbersome, and power has to be estimated separately for each sample size (Thoemmes et al., 2010).

The R package *bmem* (Zhang, 2014) implements bootstrap power analysis for bootstrap confidence intervals. First, it is necessary to define all model parameters using the *lavaan* model syntax (Rosseel, 2012). With this syntax, a model is specified as a text string in which each new line can represent either a regression relationship (using symbol '~') or a variance (using symbol '~~'). In the following example, we specify the model discussed above, in which a = .8186, b = .4039, c = .4334, and all variables have unitary variance. We save the model to a variable called model.

```
model <-'
  M ~ a*X + start(.8186)*X
  Y ~ b*M + c*X + start(.4039)*
  M + start(.4334)*X
  X ~~ start(1)*X
  M ~~ start(1)*M
  Y ~~ start(1)*Y'
```

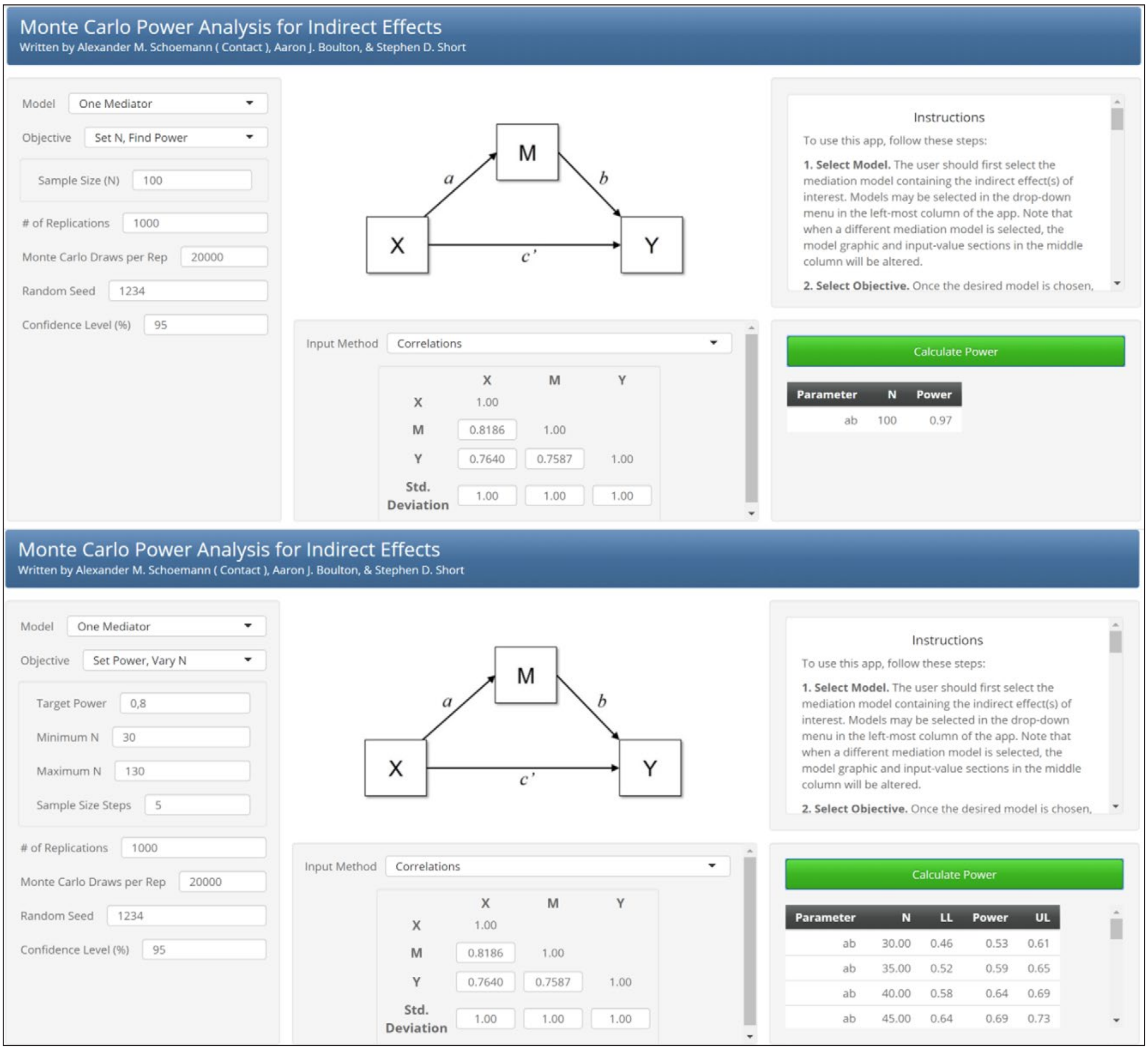
In this code, the regression of M on X is specified as $M = aX$. The syntax start (.8186)*X is used to specify that a = .8186. Similarly, the third row specifies the regression equation $bM + cX$, as well as the values of b and c. The remaining three rows specify that all variables have unitary variance. Once a model is specified, it can be used as the input of function `power.boot()` in package *bmem*, as shown in the following code.[9]

```
set.seed(1234)
power.result  <-  power.boot(model,
            indirect = 'ab := a*b',
            nobs = 100))
summary(power.result)
```

Because `power.boot()` is based on random resamples, command `set.seed()` can be used to ensure the exact reproducibility of results by forcing the R random number generator to draw the same sequence of random samples every time the code is executed. The first argument of function `power.boot` is the model specified above, the second argument is a text string that specifies that the power should be tested for the composite effect defined[10] as the product of the coefficients a and b. The third parameter specifies a sample size for which power should be computed. By default, the function tests the hypothesis of a significant indirect effect by computing 95% bootstrap confidence intervals using 1,000 bootstrap samples. Power is computed by using 1,000 Monte Carlo samples.[11] The code returns a summary of the main results of the Monte Carlo simulation, including the power for detecting each parameter. This method suggests that the power for testing the indirect effect is 97%.

Schoemann and colleagues (2017) recently proposed two strategies that allow a substantial reduction of computational time in power computations. First, instead of bootstrap confidence intervals, they consider Monte Carlo confidence intervals, which show good performance without being computationally intensive (Hayes & Scharkow, 2013; Preacher & Selig, 2012). Second, instead of estimating power separately for each sample size N, they propose a varying parameters approach in which sample size varies across replications of the Monte Carlo simulation and logistic regression is used to identify a sample size that yields the desired power. This approach has been implemented for a selected set of models in a Shiny application, which requires no programming experience. The app can be found online (https://schoemanna.shinyapps.io/mc_power_med/), or it can be installed and executed from R. This second option provides faster computations. The graphical app interface is shown in **Figure 10**. First, the app requires specifying the model for which power should be computed. One mediator indicates a model with a single mediator.[12] Second, one has to specify the model parameters in the form of correlations[13] among variables X, M, and Y.

These can be easily computed from standardized *a*, *b*, and *c* parameters in **Figure 9** as $r_{mx} = a$, $r_{yx} = c + ab$, and $r_{ym} = b + ac$. For our example, the correlations are r_{mx} = .8186, r_{yx} = .7640, and r_{ym} = .7587. Third,



Monte Carlo Power Analysis for Indirect Effects

Written by Alexander M. Schoemann (Contact), Aaron J. Boulton, & Stephen D. Short

Model: One Mediator

Objective: Set N, Find Power

Sample Size (N): 100

# of Replications: 1000

Monte Carlo Draws per Rep: 20000

Random Seed: 1234

Confidence Level (%): 95

Input Method: Correlations

	X	M	Y
X	1.00		
M	0.8186	1.00	
Y	0.7640	0.7587	1.00
Std. Deviation	1.00	1.00	1.00

Instructions

To use this app, follow these steps:

1. Select Model. The user should first select the mediation model containing the indirect effect(s) of interest. Models may be selected in the drop-down menu in the left-most column of the app. Note that when a different mediation model is selected, the model graphic and input-value sections in the middle column will be altered.

2. Select Objective. Once the desired model is chosen,

Parameter	N	Power
ab	100	0.97

Monte Carlo Power Analysis for Indirect Effects

Written by Alexander M. Schoemann (Contact), Aaron J. Boulton, & Stephen D. Short

Model: One Mediator

Objective: Set Power, Vary N

Target Power: 0,8

Minimum N: 30

Maximum N: 130

Sample Size Steps: 5

# of Replications: 1000

Monte Carlo Draws per Rep: 20000

Random Seed: 1234

Confidence Level (%): 95

Input Method: Correlations

	X	M	Y
X	1.00		
M	0.8186	1.00	
Y	0.7640	0.7587	1.00
Std. Deviation	1.00	1.00	1.00

Instructions

To use this app, follow these steps:

1. Select Model. The user should first select the mediation model containing the indirect effect(s) of interest. Models may be selected in the drop-down menu in the left-most column of the app. Note that when a different mediation model is selected, the model graphic and input-value sections in the middle column will be altered.

2. Select Objective. Once the desired model is chosen,

Parameter	N	LL	Power	UL
ab	30.00	0.46	0.53	0.61
ab	35.00	0.52	0.59	0.65
ab	40.00	0.58	0.64	0.69
ab	45.00	0.64	0.69	0.73

Figure 10: Shiny app for Monte Carlo power analysis (Schoemann et al., 2017).

the user can choose between two possible values of Objective: Set N, Find Power and SetPower, Vary N. The first option (top panel of **Figure 10**) estimates power for a specific sample size, specified by Sample Size (N) (in this case, we set N = 100). The second option implements the varying parameters approach (bottom panel of **Figure 10**). In this case, the user can specify the target power and a range of values of N (we explored samples between N = 30 and N = 130 in steps of 5). The remaining parameters allow setting the total number of replications (# of Replications, which defaults to 1,000), the number of draws for computing Monte Carlo confidence intervals (Monte Carlo Draws per Rep, which defaults to 20,000), a random seed to ensure the exact replicability of the results, and the confidence level (which defaults to 95%). The results of the first analysis show that a sample of N = 100 participants results in 97% power. The results of the second analysis show the power for each of the sample sizes considered. Sixty participants are sufficient to achieve 80% power.

Conclusions

We have presented a review of power analysis and several examples of applications in some common study contexts. In this last section, we focus on a few additional issues before drawing some final considerations.

The true effect size is unknown. The (relative) simplicity in the mechanics of power analysis masks an essential problem. After deciding *a priori* a level of statistical significance and desired power, the needed sample size can be easily determined, at least for simple designs, given an expected effect size. However, the problem is precisely this: the expected effect size is only an estimation based on an educated guess. This problem has far-reaching implications. First, even seemingly minor errors in estimation can lead to unwanted consequences. Suppose that we best guess an expected effect size to be d = 0.40 with a clear directional hypothesis. For a simple independent two-group study, this would imply a sample size of N = 156 to achieve a power of .80 with α = .05 (one-tailed). However, if the true effect size turned out to be d = 0.30, we would have actually needed N = 276. Said otherwise, the study will turn out to be substantially underpowered (with power equal to .59). Second, the impact of offset estimates is asymmetric. Overestimating the effect size has a stronger impact than underestimating it. Continuing the previous example, if the true effect turns out $d = 0.50$, a sample size of $N = 100$ would have been sufficient, meaning that if we had collected 156 participants, the actual power would have been 0.93. Compare the consequences of the two incorrect estimations. Both are offset by 0.10 from the true effect size. However, whereas underestimating the true effect size implies collecting 56 more participants than strictly needed and an increase in actual power of .13, overestimating it by the same amount implies 176 participants less than needed and a reduction of power of .21. If we consider the ubiquitous optimistic biases and superiority illusions, it is easy to predict that underestimations are more common than overestimations, with the implication that most studies are actually underpowered.

There are ways to counteract this problem. Researchers should routinely engage in a sensitivity analysis, meaning they should explore different scenarios with a range of plausible effect sizes, rather than focusing on a unique value; optionally, they could also consider the uncertainty in the estimate by using safeguard power analysis (Perugini, et al., 2014). Researchers can also consider planning for a higher level of power (e.g., .90) for their focal prediction as this might allow running ancillary analyses (e.g., a potential moderation effect suggested by a reviewer) without the test being hopelessly underpowered.[14] What is not a solution is running a small pilot study before the main study to have a better-expected effect size estimate. The problem with this seemingly sensible approach is that the estimate will be highly uncertain (i.e., with a wide confidence interval) given that it is based on a small sample; hence, it will be of little use and potentially misleading (Albers & Lakens, 2018).

Power analysis for complex designs and multiple outcomes is complex. We have focused on and provided examples of how to calculate power in a number of common but relatively simple designs. Power calculations get increasingly demanding with complex designs and analyses (e.g., multi-level designs, structural equation models, longitudinal studies). A good source for the application of power analysis for more advanced statistical models and techniques with exemplary R codes is Liu (2014; see also Maxwell et al., 2008 for a brief review of power analysis in advanced statistical models). It is worth noting that in recent years Monte Carlo simulations have started to be used especially for power calculations in complex designs (e.g., Arnold, Hogan, Colford & Hubbard, 2011; Gelman & Hill, 2006; Lane & Hennes, 2017). Bear in mind, however, that complex designs can often be broken down into key predictions that can be simple if the underlying theoretical framework is well developed and focused analyses with appropriate coding are performed (Rosenthal & Rosnow, 1985; Judd, McClelland & Ryan, 2017; see also the previous section on contrast coding). Power analysis is equally, if not more, complex when multiple outcomes are involved, meaning that the researcher aims to have adequate power when testing two or more parameters in the same study. In these cases, there is no single definition of power: it depends on the researcher's aims and theoretical expectations (e.g., one of the outcomes is significant versus all outcomes are significant). Power also depends on how many outcomes are considered, their expected correlations, and their expected combined effect (e.g., R^2). Not adjusting power calculations when testing multiple outcomes usually leads to a decrease in individual power for a single outcome, but the effect can range from minor to substantial depending on the combinations of the other features (Porter, 2017). Moreover, there is no single ideal type of adjustment for multiple outcomes, although a generally well-performing adjustment is the false discovery rate (Benjamini, & Hochberg, 1995). The simplest adjustment is to use a Bonferroni correction that considers $\frac{\alpha}{k}$, where k refers to the number of multiple outcomes: if there are 5 outcomes of interest in a study, keeping $\alpha = .05$ for each means to calculate power considering $\alpha = \frac{.05}{5} = .01$. This correction tends to be conservative and might suggest a larger sample size that is strictly needed to reach the desired level of power given an expected effect size. However, if one considers the tendency and the risk of overestimating expected effect sizes, being conservative on the side of multiple testing might be a wise approach. Other solutions can also be adopted. For example, if the multiple outcomes are expected to be correlated and no *a priori* distinction is made between primary and ancillary hypotheses, power could be calculated with a MANOVA approach on the set of multiple outcomes.[15]

Suggestions for increasing power. The single most obvious and effective way to increase power is by increasing sample size. This strategy has the additional benefit of increasing accuracy in parameter estimates, which should be an important goal on its own (Maxwell et al., 2008). However, everything being equal, power also depends on the effect size. Larger effect sizes are easier to detect, requiring a smaller sample given a fixed power level. In their generic form, effect sizes reflect the proportion of the amount of variability in the data due to the specific effect of interest (signal) relative to the variability due to other sources (noise). Therefore, one can achieve more power with larger effect sizes that in turn can be obtained by increasing the signal relative to the noise. This can be achieved in different ways, both by trying to increase the signal and trying to reduce the noise. One should aim at using reliable measures, given that reliability positively influences power (LeBel & Paunonen, 2011; but see De Schryver, Hughes, Rosseel & De Houwer, 2016). Stronger experimental treatments also increase the signal and, hence, the effect size. The noise can be reduced by using highly standardized procedures and keeping the design as simple as possible, as this should reduce generic between-person variation (i.e., not due to the effect of interest). For the same reason, the use of within-subject designs can greatly increase power, given a certain sample size, or substantially reduce the required sample size to achieve the desired level of power, and this effect is stronger with increasing correlation between the measures. More generally, bearing in mind that the effect size reflects the ratio between signal and noise can be useful as it helps to focus on methodological improvements that increase this ratio when planning the study.

Final Considerations. We aimed to offer a practical guide to power analysis with an emphasis on the logic behind it and its concrete applications. We have focused on power analysis for a single study. However, we wish to stress the importance of thinking meta-analytically (Cumming, 2013). Science is cumulative, and empirical evidence is more convincing when accumulated across studies. Adopting a meta-analytical mindset means to focus on overall evidence across studies, even a few studies or one main and a replication study, rather than a single study. Inferences from data are more robust under these conditions. If an effect is actually there, an overall analysis across studies (i.e., a small meta-analysis) will be more likely to detect it (i.e., the overall power will increase), and as an important additional benefit, its estimate will be more accurate (i.e., the confidence interval will be smaller). In closing, we wish to emphasize again that power analysis is a friend and not a foe. It helps to plan a successful study by estimating how many data points (i.e., participants) are needed to have a reasonable chance to find what one is looking for. Would you embark on a trip to the desert searching for a remote oasis without making sure that you have enough fuel in your tank to make it there? Why would you want to run a study with fewer participants than needed to have a reasonable chance of finding the hypothesized effect? Even if you do not find the effect, having enough power means it is unlikely that it is there or that it is smaller than what you expected or that it is practically useful. After all, as Cohen wisely noted (1994), all effects exist given an infinite sample size; the real question is whether the magnitude of the effect is nontrivial.

Notes

[1] In this paper, we will focus on power from an NHST perspective. However, other valuable approaches to sample size determination have been developed, such as the Bayes factor design analysis (e.g., Schönbrodt & Wagenmakers, 2017), the sequential Bayes factors (Schönbrodt, Wagenmakers, Zehetleitner & Perugini, 2017), the sequential data analysis designs (Lakens, 2014), and the accuracy in parameter estimation (AIPE; Maxwell, Kelley & Rausch, 2008).

[2] Note that Hedges' *g* estimates Cohen's *d* using the weighted mean of the within group variances, weighted by group size, to estimate the pooled standard deviation. Many authors refer to this estimator simply as Cohen's *d*.

[3] For group-based designs, some R packages output the required total sample N, others the N-per-group. Researchers should check which N is output by the specific package they are using. Finally, some packages may report the required degrees of freedom of the test, and some simple calculation is needed to obtain the required N (see https://github.com/mcfanda/primerPowerIRSP for examples).

[4] Note that when the sample size is larger than minimal (e.g., $N > 50$), the bias tends to be minor (e.g., .01 or below) under a range of conditions (Okada, 2013).

[5] The G*Power manual suggests a different formula to adjust the sample eta-squared. We suggest instead the epsilon-squared adjustment, because it is well documented in the literature (Okada, 2013).

[6] We advise against rounding up *f*-squared values to two decimals and suggest instead using three or, better, four decimals. Rounding might introduce a bias in the calculation, especially with small *f*-squared values. For instance, following the example in the main text, suppose we have either $f^2 = 0.0749$ or $f^2 = 0.0751$. This will produce an estimated N = 107 in both cases. However, if they were rounded up as $f^2 = 0.07$ and $f^2 = 0.08$, the estimated *N* will be 115 and 101, respectively.

[7] If one runs the current example in G*Power—"Linear bivariate regression: Two groups, difference between slopes"—one obtains exactly the same required N. In other applications, the two commands may give slightly different results.

[8] We have noticed that formula (9) tends to overestimate large effect sizes (interaction $f^2 > 0.15$), yielding a required N that is smaller than the one needed to achieve the desired power. This is due to the instability of correlations for small sample sizes (Schönbrodt & Perugini, 2013). Consider, however, that such large effect sizes yield small required N ($N < 50$); thus, the researcher may simply increase the sample size when dealing with such large effects.

[9] Running the following code can take a long time, especially on older machines. The function implements parallel processing to speed up computation on multicore machines (for details, see Zhang, 2014).

[10] In the *lavaan* model syntax, the operator ":=" is used to define a parameter as the combination of other parameters. For more details on the *lavaan* model syntax, see Rosseel (2012).

[11] These default options can be overridden and several other parameters can be specified that allow, for instance, considering deviations from normality in variables X, M, and Y. Furthermore, although we limited our example to a simple mediation, package *bmem* can be used to estimate power for more complex models, including multiple mediators and latent variables (see Zhang, 2014 for a complete introduction to this package).

[12] For more details on this application and on other available models implemented, see Schoemann and colleagues (2017).

[13] One can also specify covariances by setting the standard deviations of each variable.

[14] We wish to thank Vincent Yzerbyt for suggesting this scenario.

[15] We wish to thank Roger Giner-Sorolla for suggesting this scenario.

Competing Interests

The authors have no competing interests to declare.

References

Albers, C., & **Lakens, D.** (2018). When power analyses based on pilot data are biased: Inaccurate effect size estimators and follow-up bias. *Journal of Experimental Social Psychology*, *74*, 187–195. DOI: https://doi.org/10.1016/j.jesp.2017.09.004

Anderson, S. F., Kelley, K., & **Maxwell, S. E.** (2017). Sample-size planning for more accurate statistical power: A method adjusting sample effect sizes for publication bias and uncertainty. *Psychological Science*, *28*(11), 1547–1562. DOI: https://doi.org/10.1177/0956797617723724

Armitage, P., Berry, G., & **Matthews, J.** (2002). *Statistical Methods in Medical Research* (4th ed.). Blackwell Science Ltd. DOI: https://doi.org/10.1002/9780470773666

Arnold, B. F., Hogan, D. R., Colford, J. M., & **Hubbard, A. E.** (2011). Simulation methods to estimate design power: an overview for applied research. *BMC Medical Research Methodology*, *11*(1), 94. DOI: https://doi.org/10.1186/1471-2288-11-94

Asendorpf, J. B., Conner, M., De Fruyt, F., et al. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, *27*(2), 108–119. DOI: https://doi.org/10.1002/per.1919

Bakker, M., van Dijk, A., & **Wicherts, J. M.** (2012). The rules of the game called psychological science. *Perspectives on Psychological Science, 7*(6), 543–554. DOI: https://doi.org/10.1177/1745691612459060

Baron, R. M., & **Kenny, D. A.** (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*(6), 1173–1182. DOI: https://doi.org/10.1037/0022-3514.51.6.1173

Benjamin, D. J., Berger, J. O., Johannesson, M., et al. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6–10. DOI: https://doi.org/10.1038/s41562-017-0189-z

Benjamini, Y., & **Hochberg, Y.** (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodological), 57*, 289–300.

Bloom, H. S. (1995). Minimum detectable effects: A simple way to report the statistical power of experimental designs. *Evaluation Review, 19*(5), 547–556. DOI: https://doi.org/10.1177/0193841X9501900504

Brysbaert, M., & **Stevens, M.** (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition*, *1*(1), 1–20. DOI: https://doi.org/10.5334/joc.10

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., et al. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. DOI: https://doi.org/10.1038/nrn3475

Cohen, J. (1962). The statistical power of abnormal—social psychological research: A review. *Journal of Abnormal and Social Psychology, 65*(3), 145–153. DOI: https://doi.org/10.1037/h0045186

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

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*(12), 1304–12. DOI: https://doi.org/10.1037/0003-066X.45.12.1304

Cohen, J. (1992a). Statistical power analysis. *Current Directions in Psychological Science*, *1*(3), 98–101. DOI: https://doi.org/10.1111/1467-8721.ep10768783

Cohen, J. (1992b). A power primer. *Psychological Bulletin, 112*(1), 155–159. DOI: https://doi.org/10.1037/0033-2909.112.1.155

Cohen, J. (1994). The earth is round (*p* < .05). *American Psychologist, 49*(12), 997–1003. DOI: https://doi.org/10.1037/0003-066X.49.12.997

Cohen, J., Cohen, P., West, S. G., & **Aiken, L. S.** (2003). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Mahwah, N.J.: L. Erlbaum Associates.

Cumming, G. (2013). *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-analysis*. New York: Routledge.

De Schryver, M., Hughes, S., Rosseel, Y., & **De Houwer, J.** (2016). Unreliable yet still replicable: A comment on LeBel and Paunonen (2011). *Frontiers in Psychology*, *6*, 2039. DOI: https://doi.org/10.3389/fpsyg.2015.02039

Eikeland, H. M. (1975). *Epsilon-squared Should Be Preferred To Eta-squared.* Technical Report, University of Oslo.

Ellis. (2010). *The Essential Guide To Effect Sizes.* Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511761676

Faul, F., Erdfelder, E., Buchner, A., & **Lang, A. G.** (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. DOI: https://doi.org/10.3758/BRM.41.4.1149

Faul, F., Erdfelder, E., Lang, A.-G., & **Buchner, A.** (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. DOI: https://doi.org/10.3758/BF03193146

Fritz, C. O., Morris, P. E., & **Richler, J. J.** (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General*, *141*(1), 2–18. DOI: https://doi.org/10.1037/a0024338

Gelman, A. (2018, March 15). You need 16 times the sample size to estimate an interaction than to estimate a main effect [blog post]. Retrieved from: http://andrewgelman.com/2018/03/15/need-16-times-sample-size-estimate-interaction-estimate-main-effect/.

Gelman, A., & **Hill, J.** (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge, England: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511790942

Giner-Sorolla, R. (2018, January 24). Powering your interaction [blog post]. Retrieved from: https://approachingblog.wordpress.com/2018/01/24/powering-your-interaction-2/.

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. DOI: https://doi.org/10.1177/0956797613480187

Hunter, J. E., & **Schmidt, F. L.** (2004). *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings.* (2nd ed.). Thousand Oaks, CA: Sage. DOI: https://doi.org/10.4135/9781412985031

Hays, W. L. (1963). *Statistics for Psychologists.* New York: Holt, Rinehart, and Winston.

Hedges, L. V. (1981). Distribution theory for Glass' estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*(2), 107–128. DOI: https://doi.org/10.3102/10769986006002107

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124. DOI: https://doi.org/10.1371/journal.pmed.0020124

Jaccard, J., & **Turrisi, R.** (2003). *Interaction Effects in Multiple Regression.* (2nd ed.). Thousand Oaks: Sage. DOI: https://doi.org/10.4135/9781412984522

Judd, C. M., McClelland, G. H., & **Ryan, C. S.** (2017). *Data Analysis: A Model Comparison Approach To Regression.* (3rd ed.). New York: Routledge.

Kelly, T. L. (1935). An unbiased correlation ratio measure. *Proceedings of the National Academy of Sciences*, *21*, 554–559. DOI: https://doi.org/10.1073/pnas.21.9.554

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*, 863. DOI: https://doi.org/10.3389/fpsyg.2013.00863

Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology*, *44*(7), 701–710. DOI: https://doi.org/10.1002/ejsp.2023

Lakens, D., Adolfi, F. G., Albers, C. J., et al. (2018). Justify your alpha. *Nature Human Behaviour*, *2,* 168–171. DOI: https://doi.org/10.1038/s41562-018-0311-x

Lane, S. P., & **Hennes, E. P.** (2017). Power struggles: Estimating sample size for multilevel relationships research. *Journal of Social and Personal Relationships*, *35*(1), 7–31. DOI: https://doi.org/10.1177/0265407517710342

LeBel, E. P., & **Paunonen, S. V.** (2011). Sexy but often unreliable: The impact of unreliability on the replicability of experimental findings with implicit measures. *Personality and Social Psychology Bulletin*, *37*(4), 570–583. DOI: https://doi.org/10.1177/0146167211400619

Liu, X. S. (2014). *Statistical Power Analysis for the Social and Behavioral Sciences: Basic and Advanced Techniques.* New York: Routledge.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: causes, consequences, and remedies. *Psychological Methods*, *9*(2), 147–163. DOI: https://doi.org/10.1037/1082-989X.9.2.147

Maxwell, S. E., Kelley, K., & **Rausch, J. R.** (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology, 59*, 537–563. DOI: https://doi.org/10.1146/annurev.psych.59.103006.093735

McClelland, G. H., & **Judd, C. M.** (1993). Statistical difficulties of detecting interactions and moderator effects. *Psychological Bulletin, 114*(2), 376–390. DOI: https://doi.org/10.1037/0033-2909.114.2.376

O'Brien, R. G., & **Castelloe, J. M.** (2007). Sample size analysis for traditional hypothesis testing: concepts and issues. In: Dmitrienko, A., Chuang-Stein, C., D'Agostino, R. (eds.), *Pharmaceutical Statistics Using SAS: A Practical Guide*, 237–71. Cary, NC: SAS.

Okada, K. (2013). Is omega squared less biased? A comparison of three major effect size indices in one-way ANOVA. *Behaviormetrika*, *40*(2), 129–147. DOI: https://doi.org/10.2333/bhmk.40.129

Perugini, M., Gallucci, M., & **Costantini, G.** (2014). Safeguard power as a protection against imprecise power estimates. *Perspectives on Psychological Science*, *9*(3), 319–332. DOI: https://doi.org/10.1177/1745691614528519

Porter, K. E. (2017). Statistical power in evaluations that investigate effects on multiple outcomes: A guide for researchers. *Journal of Research on Educational Effectiveness*, *11*(2), 267–295. DOI: https://doi.org/10.1080/19345747.2017.1342887

Preacher, K. J., & **Hayes, A. F.** (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, & Computers*, *36*(4), 717–731. DOI: https://doi.org/10.3758/BF03206553

Preacher, K. J., & **Selig, J. P.** (2012). Advantages of Monte Carlo confidence intervals for indirect effects. *Communication Methods and Measures*, *6*(2), 77–98. DOI: https://doi.org/10.1080/19312458.2012.679848

Qiu, W. (2017). *powerMediation: Power/Sample Size Calculation for Mediation Analysis.* R package version 0.2.7.

R Core Team. (2017). *R: A Language and Environment for Statistical Computing.* Vienna: R Foundation for Statistical Computing.

Richardson, J. T. (2011). Eta squared and partial eta squared as measures of effect size in educational research. *Educational Research Review, 6*(2), 135–147. DOI: https://doi.org/10.1016/j.edurev.2010.12.001

Rosenthal, R., & **Rosnow, R. L.** (1985). *Contrast Analysis: Focused Comparisons in the Analysis of Variance.* Cambridge University Press.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36. DOI: https://doi.org/10.18637/jss.v048.i02

Ruscio, J. (2008). A probability-based measure of effect size: robustness to base rates and other factors.

Psychological Methods, *13*(1), 19–30. DOI: https://doi.org/10.1037/1082-989X.13.1.19

Ruscio, J., & **Mullen, T.** (2012). Confidence intervals for the probability of superiority effect size measure and the area under a receiver operating characteristic curve. *Multivariate Behavioral Research*, *47*(2), 201–223. DOI: https://doi.org/10.1080/00273171.2012.658329

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. DOI: https://doi.org/10.1177/1948550617715068

Schönbrodt, F. D., & **Perugini, M.** (2013). At what sample size do correlations stabilize? *Journal of Research in Personality*, *47*(5), 609–612. DOI: https://doi.org/10.1016/j.jrp.2013.05.009

Schönbrodt, F. D., & **Wagenmakers, E.** (2017). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, (2014), 1–13. DOI: https://doi.org/10.3758/s13423-017-1230-y

Schönbrodt, F. D., Wagenmakers, E., Zehetleitner, M., & **Perugini, M.** (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods*, *22*(2), 322–339. DOI: https://doi.org/10.1037/met0000061

Shieh, G. (2009). Detecting interaction effects in moderated multiple regression with continuous variables power and sample size considerations. *Organizational Research Methods, 12*(3), 510–528. DOI: https://doi.org/10.1177/1094428108320370

Simonsohn, U. (2014). [17] No-way Interactions. *The Winnower,* 5, e142559.90552. DOI: https://doi.org/10.15200/winn.142559.90552

Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. *Sociological Methodology*, *13*, 290–312. DOI: https://doi.org/10.2307/270723

Sterne, J. A., & **Smith, D. G.** (2001). Sifting the evidence—What's wrong with significance tests? *Physical Therapy*, *81*(8), 1464–1469. DOI: https://doi.org/10.1093/ptj/81.8.1464

Swiatkowski, W., & **Dompnier, B.** (2017). Replicability Crisis in Social Psychology: Looking at the Past to Find New Pathways for the Future. *International Review of Social Psychology*, *30(1)*, 111–124. DOI: https://doi.org/10.5334/irsp.66

Thoemmes, F., MacKinnon, D. P., & **Reiser, M. R.** (2010). Power analysis for complex mediational designs using Monte Carlo methods. *Structural Equation Modeling*, *17*(3), 510–534. DOI: https://doi.org/10.1080/10705511.2010.489379

Wahlsten, D. (1991). Sample size to detect a planned contrast and a one degree-of-freedom interaction effect. *Psychological Bulletin, 110*(3), 587–595. DOI: https://doi.org/10.1037/0033-2909.110.3.587

Wilkinson, L., & **Task Force Statistical Inference.** (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*(8), 594–604. DOI: https://doi.org/10.1037/0003-066X.54.8.594

Zhang, Z. (2014). Monte Carlo based statistical power analysis for mediation models: Methods and software. *Behavior Research Methods*, *46*(4), 1184–1198. DOI: https://doi.org/10.3758/s13428-013-0424-0

Zumbo, B. D., & **Hubley, A. M.** (1998). A note on misconceptions concerning prospective and retrospective power. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *47*(2), 385–388. DOI: https://doi.org/10.1111/1467-9884.00139

How to cite this article: Perugini, M., Gallucci, M., and Costantini, G. (2018). A Practical Primer To Power Analysis for Simple Experimental Designs. *International Review of Social Psychology*, 31(1): 20, 1–23, DOI: https://doi.org/10.5334/irsp.181

Published: 09 July 2018

Copyright: © 2018 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See http://creativecommons.org/licenses/by/4.0/.

]u[*International Review of Social Psychology* is a peer-reviewed open access journal published by Ubiquity Press. OPEN ACCESS