Sample Size Justification

Daniël Lakens[1]

[1] Eindhoven University of Technology

This unpublished manuscript is submitted for peer review

An important step when designing an empirical study is to justify the sample size that will be collected. The key aim of a sample size justification for such studies is to explain how the collected data is expected to provide valuable information given the inferential goals of the researcher. In this overview article six approaches are discussed to justify the sample size in a quantitative empirical study: 1) collecting data from (an)almost) the entire population, 2) choosing a sample size based on resource constraints, 3) performing an a-priori power analysis, 4) planning for a desired accuracy, 5) using heuristics, or 6) explicitly acknowledging the absence of a justification. An important question to consider when justifying sample sizes is which effect sizes are deemed interesting, and the extent to which the data that is collected informs inferences about these effect sizes. Depending on the sample size justification chosen, researchers could consider 1) what the smallest effect size of interest is, 2) which minimal effect size will be statistically significant, 3) which effect sizes they expect (and what they base these expectations on), 4) which effect sizes would be rejected based on a confidence interval around the effect size, 5) which ranges of effects a study has sufficient power to detect based on a sensitivity power analysis, and 6) which effect sizes are plausible in a specific research area. Researchers can use the guidelines presented in this article to improve their sample size justification, and hopefully, align the informational value of a study with their inferential goals.

Keywords: sample size justification, study design, power analysis, value of information
Word count: 16642

Scientists perform empirical studies to collect data that helps to answer a research question. The more data that is collected, the more informative the study will be with respect to its inferential goals. A sample size justification should consider how informative the data will be given an inferential goal, such as an estimating an effect size, or testing a hypothesis. Even though a sample size justification is sometimes requested in manuscript submission guidelines, when submitting a grant to a funder, or submitting a proposal to an ethical review board, the number of observations is often simply *stated*, but not *justified*. This makes it difficult to evaluate how informative a study will be. To prevent such concerns from emerging when it is too late (e.g., after a non-significant hypothesis test has been observed), researchers should carefully justify their sample size before data is collected.

This work was funded by VIDI Grant 452-17-013 from the Netherlands Organisation for Scientific Research. I would like to thank Shilaan Alzahawi, José Biurrun, Aaron Caldwell, Gordon Feld, Yaov Kessler, Robin Kok, Maximilian Maier, Matan Mazor, Toni Saari, Andy Siddall, and Jesper Wulff for feedback on an earlier draft. A computationally reproducible version of this manuscript is available at https://github.com/Lakens/sample_size_justification.

Correspondence concerning this article should be addressed to Daniël Lakens, Den Dolech 1, 5600MB Eindhoven, The Netherlands. E-mail: D.Lakens@tue.nl

Six Approaches to Justify Sample Sizes

Researchers often find it difficult to justify their sample size (i.e., a number of participants, observations, or any combination thereof). In this review article six possible approaches are discussed that can be used to justify the sample size in a quantitative study (see Table 1). This is not an exhaustive overview, but it includes the most common and applicable approaches for single studies.[1] The first justification is that data from (almost) the entire population has been collected. The second justification centers on resource constraints, which are almost always present, but rarely explicitly evaluated. The third and fourth justifications are based on a desired statistical power or a desired accuracy. The fifth justification

[1] The topic of power analysis for meta-analyses is outside the scope of this manuscript, but see Hedges and Pigott (2001) and Valentine, Pigott, and Rothstein (2010).

Table 1
Overview of possible justifications for the sample size in a study.

Type of justification	When is this justification applicable?
Measure entire population	A researcher can specify the entire population, it is finite, and it is possible to measure (almost) every entity in the population.
Resource constraints	Limited resources are the primary reason for the choice of the sample size a researcher can collect.
Accuracy	The research question focusses on the size of a parameter, and a researcher collects sufficient data to have an estimate with a desired level of accuracy.
A-priori power analysis	The research question has the aim to test whether certain effect sizes can be statistically rejected with a desired statistical power.
Heuristics	A researcher decides upon the sample size based on a heuristic, general rule or norm that is described in the literature, or communicated orally.
No justification	A researcher has no reason to choose a specific sample size, or does not have a clearly specified inferential goal and wants to communicate this honestly.

relies on heuristics, and finally, researchers can choose a sample size without any justification. Each of these justifications can be stronger or weaker depending on which conclusions researchers want to draw from the data they plan to collect.

All of these approaches to the justification of sample sizes, even the 'no justification' approach, are valid justifications in the sense that they give others insight into the reasons that led to the decision for a sample size in a study. It should not be surprising that the 'heuristics' and 'no justification' approaches are often unlikely to impress peers. However, it is important to note that the value of the information that is collected depends on the extent to which the final sample size allows a researcher to achieve their inferential goals, and not on the sample size justification that is chosen.

The extent to which these approaches make other researchers judge the data that is collected as *informative* depends on the details of the question a researcher aimed to answer and the parameters they chose when determining the sample size for their study. For example, a badly performed a-priori power analysis can quickly lead to a study with very low informational value. These six justifications are not mutually exclusive, and multiple approaches can be considered when designing a study.

Six Ways to Evaluate Which Effect Sizes are Interesting

The informativeness of the data that is collected depends on the inferential goals a researcher has, or in some cases, the inferential goals scientific peers will have. A shared feature of the different inferential goals considered in this review article is the question which effect sizes a researcher considers meaningful to distinguish. This implies that researchers need to evaluate which effect sizes they consider interesting. These evaluations rely on a combination of statistical properties and domain knowledge. In Table 2 six possibly useful considerations are provided. This is not intended to be an exhaustive overview, but it presents common and useful approaches that can be applied in practice. Not all evaluations are equally relevant for all types of sample size justifications. These considerations often rely on the same information (e.g., effect sizes, the number of observations, the standard deviation, etc.) so these six considerations should be seen as a set of complementary approaches that can be used to evaluate which effect sizes are of interest.

To start, researchers should consider what their smallest effect size of interest is. Second, although only relevant when performing a hypothesis test, researchers should consider which effect sizes could be statistically significant given a choice of an alpha level and sample size. Third, it is important to consider the (range of) effect sizes that are expected. This requires a careful consideration of the source of this expectation and the presence of possible biases in these expectations. Fourth, it is useful to consider the width of the confidence interval around possible values of the effect size in the population, and whether we can expect this confidence interval to reject effects we considered a-priori plausible. Fifth, it is worth evaluating the power of the test across a wide range of possible effect sizes in a sensitivity power analysis. Sixth, a researcher can

consider the effect size distribution of related studies in the literature.

The Value of Information

Since all scientists are faced with resource limitations, they need to balance the cost of collecting each additional datapoint against the increase in information that datapoint provides. This is referred to as the *value of information* (Eckermann, Karnon, & Willan, 2010). Calculating the value of information is notoriously difficult (Detsky, 1990). Researchers need to specify the cost of collecting data, and weigh the costs of data collection against the increase in utility that having access to the data provides. From a value of information perspective not every data point that can be collected is equally valuable (Halpern, Brown Jr, & Hornberger, 2001; Wilson, 2015). Whenever additional observations do not change inferences in a meaningful way, the costs of data collection can outweigh the benefits.

The value of additional information can be a non-monotonic function when it depends on multiple inferential goals (see Figure 1). A researcher might be interested in comparing an effect against a previously observed large effect in the literature, a theoretically predicted medium effect, and the smallest effect that would be practically relevant. In such a situation the expected value of sampling information will lead to different optimal sample sizes for each inferential goal. It could be valuable to collect informative data about a large effect, with additional data having less marginal utility, up to a point where the data becomes increasingly informative about a medium effect size, with the value of sampling additional information decreasing once more until the study becomes increasingly informative about the presence or absence of a smallest effect of interest.

Because of the difficulty of quantifying the value of information, scientists typically use less formal approaches to justify the amount of data they set out to collect in a study. Even though the cost-benefit analysis is not always made explicit in reported sample size justifications, the value of information perspective is almost always implicitly the underlying framework that sample size justifications are based on. Throughout the subsequent discussion of sample size justifications, the importance of considering the value of information given inferential goals, will repeatedly be highlighted.

Measuring (Almost) the Entire Population

In some instances, it might be possible to collect data from (almost) the entire population under investigation. For example, researchers might use census data, are able to collect data from all employees at a firm or study a small population of top athletes. Whenever it is possible to measure the entire population, the sample size justification becomes straightforward: the researcher used all the data that is available.

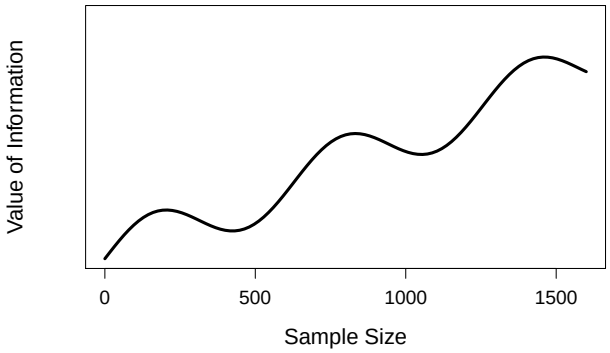

Figure 1. Example of a non-monotonically increasing value of information as a function of the sample size.

When the entire population is measured there is no need to perform a hypothesis test. After all, there is no population to generalize to.[2] When data from the entire population has been collected the population effect size is known, and there is no confidence interval to compute. If the total population size is known, but not measured completely, then the confidence interval width should shrink to zero the closer a study gets to measuring the entire population. This is known as the finite population correction factor for the variance of the estimator (Kish, 1965). The variance of a sample mean is σ^2/n, which for finite populations is multiplied by the finite population correction factor of the standard error:

$$FPC = \sqrt{\frac{(N-n)}{(N-1)}}$$

where N is the size of the population, and n is the size of the sample. When N is much larger than n, the correction factor will be close to 1 (and therefore this correction is typically ignored when populations are very large, even when populations are finite), and will not have a noticeable effect on the variance. When the total population is measured the correction factor is 0, such that the variance becomes 0 as well. For example, when the total population consists of 100 top athletes,

[2] It is possible to argue we are still making an inference, even when the entire population is observed, because we have observed a *metaphorical population* from one of many possible worlds, see Spiegelhalter (2019).

Table 2
Overview of possible ways to evaluate which effect sizes are interesting.

Type of evaluation	Which question should a researcher ask?
Smallest effect size of interest	What is the smallest effect size that that is considered theoretically or practically interesting?
The minimal statistically detectable effect	Given the test and sample size, what is the critical effect size that can be statistically significant?
Expected effect size	Which effect size is expected based on theoretical predictions or previous research?
Width of confidence interval	Which effect sizes are excluded based on the expected width of the confidence interval around the effect size?
Sensitivity power analysis	Across a range of possible effect sizes, which effects does a design have sufficient power to detect when performing a hypothesis test?
Distribution of effect sizes in a research area	What is the empirical range of effect sizes in a specific research area, and which effects are a priori unlikely to be observed?

and data is collected from a sample of 35 athletes, the finite population correction is $\sqrt{(100-35)/(100-1)} = 0.81$. The `superb` R package can compute population corrected confidence intervals (Cousineau & Chiasson, 2019).

Resource Constraints

A common reason for the number of observations in a study is that resource constraints limit the amount of data that can be collected at a reasonable cost (Lenth, 2001). In practice, sample size are always limited by the resources that are available. Researchers practically always have resource limitations, and therefore even when resource constraints are not the primary justification for the sample size in a study, it is always a secondary justification.

Despite the omnipresence of resource limitations, the topic often receives little attention in texts on experimental design. This might make it feel like acknowledging resource constraints is not appropriate, but the opposite is true: Because resource limitations always play a role, a responsible scientist carefully evaluates resource constraints when designing a study.

Resource constraint justifications are based on a trade-off between the costs of data collection, and the value of having access to the information the data provides. Even if researchers do not explicitly quantify this trade-off, it is revealed in their actions. For example, researchers rarely spend all the resources they have on a single study. Given resource constraints, researchers are confronted with an optimization problem of how to spend resources across multiple research questions. Time and money are two resource limitations all scientists face. A PhD student has a certain time to complete a PhD thesis, and is typically expected to complete multiple research lines in this time. In addition to time limitations, researchers have limited financial resources that often direclty influence how much data can be collected. A third limitation in some research lines is that there might simply be a very small number of individuals from whom data can be collected, such as when studying patients with a rare disease. A resource constraint justification puts limited resources at the center of the justification for the sample size that will be collected, and *starts* with the resources a scientist has available. These resources are translated into an expected number of observations (N) that a researcher expects they will be able to collect with an amount of money in a given time. The challenge is to evaluate whether collecting N observations is worthwhile. How do we decide if a study will be informative, and when should we conclude that data collection is *not* worthwhile?

When evaluating whether resource constraints make data collection uninformative, researchers need to explicitly consider which inferential goals they have when collecting data (Parker & Berman, 2003). Having data always provides more knowledge about the research question than not having data, so in an absolute sense, all data that is collected has value. However, it is possible that the benefits of collecting the data are outweighed by the costs of data collection.

It is most straightforward to evaluate whether data collection has value when we know for certain that someone will make a decision, with or without data. In such

situations any additional data will reduce the error rates of a well-calibrated decision process, even if only ever so slightly. For example, without data we will not perform better than a coin flip if we guess which of two conditions has a higher true mean score on a measure. With some data, we can perform better than a coin flip by picking the condition that has the highest mean. With a small amount of data we would still very likely make a mistake, but the error rate is smaller than without any data. In these cases, the value of information might be positive, as long as the reduction in error rates is more beneficial than the cost of data collection.

Another way in which a small dataset can be valuable is if its existence eventually makes it possible to perform a meta-analysis (Maxwell & Kelley, 2011). This argument in favor of collecting a small dataset requires 1) that researchers share the data in a way that a future meta-analyst can find it, and 2) that there is a decent probability that someone will perform a high-quality meta-analysis that will include this data in the future (Halpern, Karlawish, & Berlin, 2002). The uncertainty about whether there will ever be such a meta-analysis should be weighed against the costs of data collection.

One way to increase the probability of a future meta-analysis is if researchers commit to performing this meta-analysis themselves, by combining several studies they have performed into a small-scale meta-analysis (Cumming, 2014). For example, a researcher might plan to repeat a study for the next 12 years in a class they teach, with the expectation that after 12 years a meta-analysis of 12 studies would be sufficient to draw informative inferences (but see ter Schure and Grünwald (2019)). If it is not plausible that a researcher will collect all the required data by themselves, they can attempt to set up a collaboration where fellow researchers in their field commit to collecting similar data with identical measures. If it is not likely that sufficient data will emerge over time to reach the inferential goals, there might be no value in collecting the data.

Even if a researcher believes it is worth collecting data because a future meta-analysis will be performed, they will most likely perform a statistical test on the data. To make sure their expectations about the results of such a test are well-calibrated, it is important to consider which effect sizes are of interest. From the six ways to evaluate which effect sizes are interesting that will be discussed in the second part of this review, it is useful to consider the smallest effect size that can be statistically significant, the expected width of the confidence interval around the effect size, and to perform a sensitivity power analysis. If a decision or claim is made, a compromise power analysis is worthwhile to consider when deciding upon the error rates while planning the study. When reporting a resource constraints sample size justification it is recommended to address the five considerations in Table 3. Addressing these points explicitly facilitates evaluating if the data is worthwhile to collect.

A-priori Power Analysis

When designing a study where the goal is to test whether a statistically significant effect is present, researchers often want to make sure their sample size is large enough to prevent erroneous conclusions for a range of effect sizes they care about. In this approach to justifying a sample size, the value of information is to collect observations up to the point that the probability of an erroneous inference is, in the long run, not larger than a desired value. If a researcher performs a hypothesis test, there are four possible outcomes:

1. A false positive (or Type I error), determined by the α level. A test yields a significant result, even though the null hypothesis is true.
2. A false negative (or Type II error), determined by β, or 1 - power. A test yields a non-significant result, even though the alternative hypothesis is true.
3. A true negative, determined by 1-α. A test yields a non-significant result when the null hypothesis is true.
4. A true positive, determined by 1-β. A test yields a significant result when the alternative hypothesis is true.

Given a specified effect size, alpha level, and power, an a-priori power analysis can be used to calculate the number of observations required to achieve the desired error rates, given the effect size.[3] Figure 2 illustrates how the statistical power increases as the number of observations (per group) increases in an independent t test with a two-sided alpha level of 0.05. If we are interested in detecting an effect of $d = 0.5$ a sample size of 90 per condition would give us more than 90% power. Statistical power can be computed to determine the number of participants, or the number of items (Westfall, Kenny, & Judd, 2014) but can also be performed for single case studies (Ferron & Onghena, 1996; McIntosh & Rittmo, 2020)

Although it is common to set the Type I error rate to 5%

[3] Power analyses can be performed based on standardized effect sizes or effect sizes expressed on the original scale. It is important to know the standard deviation of the effect (see the 'Know Your Measure' section) but I find it slightly more convenient to talk about standardized effects in the context of sample size justifications.

Table 3
Overview of recommendations when reporting a sample size justification based on resource constraints.

What to address	How to address it?
Will a future meta-analysis be performed?	Consider the plausibility that sufficient highly similar studies will be performed in the future to make a meta-analysis possible.
Will a decision or claim be made regardless of the amount of data that is available?	If a decision is made then any data that is collected will reduce error rates. Consider using a compromise power analysis to determine Type I and Type II error rates. Are the costs worth the reduction in errors?
What is the critical effect size?	Report and interpret the critical effect size, with a focus on whether a hypothesis test would be significant for expected effect sizes. If not, indicate the interpretation of the data will not be based on p values.
What is the width of the confidence interval?	Report and interpret the width of the confidence interval. What will an estimate with this much uncertainty be useful for? If the null hypothesis is true, would rejecting effects outside of the confidence interval be worthwhile (ignoring how a design might have low power to actually test against these values)?
Which effect sizes will a design have decent power to detect?	Report a sensitivity power analysis, and report the effect sizes that can be detected across a range of desired power levels (e.g., 80%, 90%, and 95%) or plot a sensitivity analysis.

and aim for 80% power, error rates should be justified (Lakens, Adolfi, et al., 2018). As explained in the section on compromise power analysis, the default recommendation to aim for 80% power lacks a solid justification. In general, the lower the error rates (and thus the higher the power), the more informative a study will be, but the more resources are required. Researchers should carefully weigh the costs of increasing the sample size against the benefits of lower error rates, which would probably make studies designed to achieve a power of 90% or 95% more common for articles reporting a single study. An additional consideration is whether the researcher plans to publish an article consisting of a set of replication and extension studies, in which case the probability of observing multiple Type I errors will be very low, but the probability of observing mixed results even when there is a true effect increases (Lakens & Etz, 2017), which would also be a reason to aim for studies with low Type II error rates, perhaps even by slightly increasing the alpha level for each individual study.

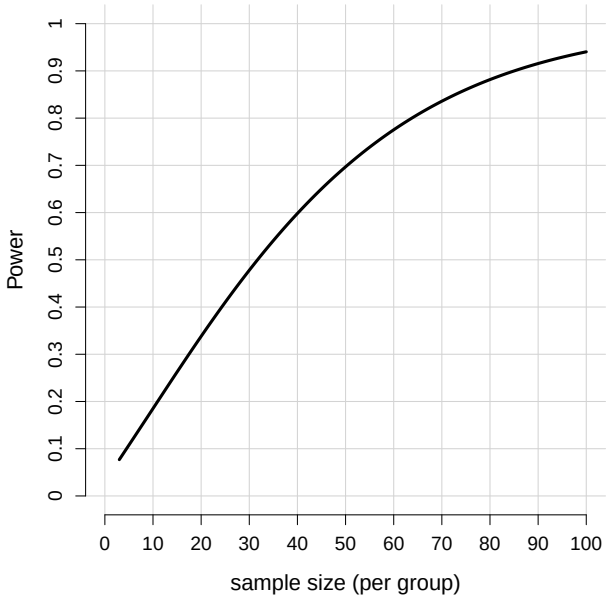

Figure 2. Power curve for an independent t test with a true effect size of $\delta = 0.5$ and an alpha of 0.05 as a function of the sample size.

Figure 3 visualizes two distributions. The left distribution (dashed line) is centered at 0. This is a model for the null hypothesis. If the null hypothesis is true a statistically significant result will be observed if the effect size is extreme enough (in a two-sided test either in the positive or negative direction), but any significant result would be a Type I error (the dark grey areas under the curve). If there is no true effect, formally statistical power for a null hypothesis significance test is undefined. Any significant effects observed if the null hypothesis is true are Type I errors, or false positives, which occur at the chosen alpha level. The right distribution (solid line) is centered on an effect of $d = 0.5$. This

is the specified model for the alternative hypothesis in this study, illustrating the expectation of an effect of d = 0.5 if the alternative hypothesis is true. Even though there is a true effect, studies will not always find a statistically significant result. This happens when, due to random variation, the observed effect size is too close to 0 to be statistically significant. Such results are false negatives (the light grey area under the curve on the right). To increase power, we can collect a larger sample size. As the sample size increases, the distributions become more narrow, reducing the probability of a Type II error.[4]

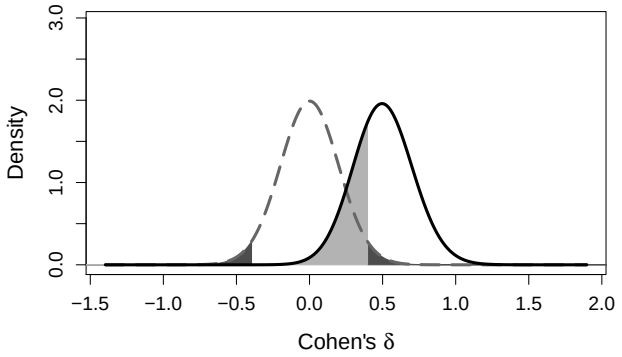

Figure 3. Null (δ = 0, grey dashed line) and alternative (δ = 0.5, solid black line) hypothesis, with α = 0.05 and n = 80 per group.

It is important to highlight that the goal of an a-priori power analysis is *not* to achieve sufficient power for the true effect size. The true effect size is unknown. The goal of an a-priori power analysis is to achieve sufficient power, given a specific *assumption* of the effect size a researcher wants to detect. Just like a Type I error rate is the maximum probability of making a Type I error conditional on the assumption that the null hypothesis is true, an a-priori power analysis is computed under the assumption of a specific effect size. It is unknown if this assumption is correct. All a researcher can do is to make sure their assumptions are well justified. Statistical inferences based on a test where the Type II error is controlled are conditional on the assumption of a specific effect size. They allow the inference that, assuming the true effect size is at least as large as that used in the a-priori power analysis, the maximum Type II error rate in a study is not larger than a desired value.

This point is perhaps best illustrated if we consider a study where an a-priori power analysis is performed both for a test of the *presence* of an effect, as for a test of the *absence* of an effect. When designing a study, it essential to consider the possibility that there is no effect (e.g., a mean difference of zero). An a-priori power analysis can be performed both for a null hypothesis significance test, as for a test of the absence of a meaningful effect, such as an equivalence test that can statistically provide support for the null hypothesis by reject the presence of effects that are large enough to matter (see Meyners, 2012; Lakens, 2017; Rogers, Howard, & Vessey, 1993). When multiple primary tests will be performed based on the same sample, each analysis requires a dedicated sample size justification. If possible, a sample size is collected that guarantees that all tests are informative, which means that the collected sample size is based on the largest sample size returned by any of the a-priori power analyses.

For example, if the goal of a study is to detect or reject an effect size of d = 0.4 with 90% power, and the alpha level is set to 0.05 for a two-sided independent t test, a researcher would need to collect 133 participants in each condition for an informative null hypothesis test, and 136 participants in each condition for an informative equivalence test. Therefore, the researcher should aim to collect 272 participants for an informative result for both tests that are planned. This does not guarantee a study has sufficient power for the true effect size (which can never be known), but it guarantees the study has sufficient power given an assumption of the effect a researcher is interested in detecting or rejecting. Therefore, an a-priori power analysis is useful, as long as a researcher can justify the effect sizes they are interested in.

If researchers are willing to reject a single hypothesis if any of multiple tests yield a significant result, then the alpha level in the a-priori power analysis should be corrected for multiple comparisons. For example, if four tests are performed, an overall Type I error rate of 5% is desired, and a Bonferroni correction is used. the a-priori power analysis should be based on a corrected alpha level of .0125.

An a-priori power analysis can be performed analytically, or by performing computer simulations. Analytic solutions are faster but less flexible. A common challenge researchers face when attempting to perform power analyses for more complex or uncommon tests is that available software does not offer analytic solutions. In these cases simulations can provide a flexible solution to perform power analyses for any test (Morris, White, & Crowther, 2019). The following code is an example of a power analysis in R based on 10000

[4] These figures can be reproduced and adapted in an online shiny app: http://shiny.ieis.tue.nl/d_p_power/.

simulations for a one-sample t test against zero for a sample size of 20, assuming a true effect of $d = 0.5$. All simulations consist of first randomly generating data based on assumptions of the data generating mechanism (e.g., a normal distribution with a mean of 0.5 and a standard deviation of 1), followed by a test performed on the data. By computing the percentage of significant results, power can be computed for any design.

```
p <- numeric(10000)
for (i in 1:10000) {
  x <- rnorm(n = 20, mean = 0.5, sd = 1)
  p[i] <- t.test(x)$p.value
}
sum(p < 0.05) / 10000
```

There is a wide range of tools available to perform power analyses. Whichever tool a researcher decides to use, it will take time to learn how to use the software correctly to perform a meaningful a-priori power analysis. Resources to educate psychologists about power analysis consist of book-length treatments (Aberson, 2019; Cohen, 1988; Julious, 2004; Murphy, Myors, & Wolach, 2014), general introductions (Baguley, 2004; Brysbaert, 2019; Faul, Erdfelder, Lang, & Buchner, 2007; Maxwell, Kelley, & Rausch, 2008; Perugini, Gallucci, & Costantini, 2018), and an increasing number of applied tutorials for specific tests (for example, DeBruine & Barr, 2019; Brysbaert & Stevens, 2018; Green & MacLeod, 2016; Kruschke, 2013; Lakens & Caldwell, 2019; Schoemann, Boulton, & Short, 2017; Westfall et al., 2014). It is important to be trained in the basics of power analysis, and it can be extremely beneficial to learn how to perform simulation-based power analyses. At the same time, it is often recommended to enlist the help of an expert, especially when a researcher lacks experience with a power analysis for a specific test.

When reporting an a-priori power analysis, make sure that the power analysis is completely reproducible. If power analyses are performed in R it is possible to share the analysis script and information about the version of the package. In many software packages it is possible to export the power analysis that is performed as a PDF file. For example, in G*Power analyses can be exported under the 'protocol of power analysis' tab. If the software package provides no way to export the analysis, add a screenshot of the power analysis to the supplementary files.

The reproducible report needs to be accompanied by justifications for the choices that were made with respect to the values used in the power analysis. If the effect size used in the power analysis is based on previous research the factors presented in Table 5 (if the effect size is based on a meta-analysis) or Table 6 (if the effect size is based on a single study) should be discussed. If an effect size estimate is based on the existing literature, provide a full citation, and preferably a direct quote from the article where the effect size estimate is reported. If the effect size is based on a smallest effect size of interest, this value should not just be stated, but justified (e.g., based on theoretical predictions or practical implications, see Lakens et al. (2018)). For an overview of all aspects that should be reported when describing an a-priori power analysis, see Table 4.

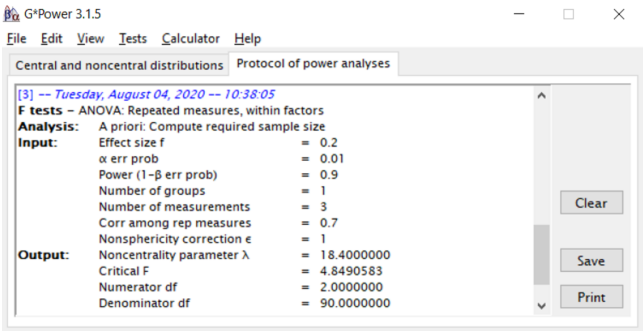

Figure 4. All details about the power analysis that is performed can be exported in G*Power.

Planning for Precision

Some researchers have suggested to justify sample sizes based on a desired level of precision of the estimate (Cumming & Calin-Jageman, 2016; Kruschke, 2018; Maxwell et al., 2008). The goal when justifying a sample size based on precision is to collect data to achieve a desired width of the confidence interval around a parameter estimate. The width of the confidence interval around the parameter estimate depends on the standard deviation and the number of observations. The only aspect a researcher needs to justify for a sample size justification based on accuracy is the desired width of the confidence interval with respect to their inferential goal, and their assumption about the population standard deviation of the measure.

If a researcher has determined the desired accuracy, and has a good estimate of the true standard deviation of the measure, it is straightforward to calculate the sample size needed for a desired level of accuracy. For example, when measuring the IQ of a group of individuals a researcher might desire to estimate the IQ score within an error range of 2 IQ points for 95% of the observed means, in the long run. The required sample size to achieve this desired level of accuracy (assuming normally distributed data) can be computed by:

Table 4
Overview of recommendations when reporting an a-priori power analysis.

What to take into account?	How to take it into account?
List all primary analyses that are planned.	Specify all planned primary analyses that test hypotheses for which Type I and Type II error rates should be controlled.
Specify the alpha level for each analysis	List and justify the Type I error rate for each analysis. Make sure to correct for multiple comparisons where needed.
What is the desired power?	List and justify the desired power (or Type II error rate) for each analysis.
For each power analysis, specify the effect size metric, the effect size, and the justification for powering for this effect size.	Report the effect size metric (e.g., Cohen's d, Cohen's f), the effect size (e.g., 0.3). and the justification for the effect size, and whether it is based is based on a smallest effect size of interest, a meta-analytic effect size estimate, the estimate of a single previous study, or some other source.
Consider the possibility that the null hypothesis is true.	Perform a power analysis for the test that is planned to examine the absence of a meaningful effect (e.g., power for an equivalence test).
Make sure the power analysis is reproducible.	Include the code used to run the power analysis, or print a report containing the details about the power analyses that has been performed.

$$N = \left(\frac{z \cdot sd}{error}\right)^2$$

where N is the number of observations, z is the critical value related to the desired confidence interval, sd is the standard deviation of IQ scores in the population, and *error* is the width of the confidence interval within which the mean should fall, with the desired error rate. In this example, (1.96 × 15 / 2)^2 = 216.1 observations. If a researcher desires 95% of the means to fall within a 2 IQ point range around the true population mean, 217 observations should be collected. If a desired accuracy for a non-zero mean difference is computed, accuracy is based on a non-central t-distribution. For these calculations an expected effect size estimate needs to be provided, but it has relatively little influence on the required sample size (Maxwell et al., 2008). It is also possible to incorporate uncertainty about the observed effect size in the sample size calculation, known as *assurance* (Kelley & Rausch, 2006). The MBESS package in R provides functions to compute sample sizes for a wide range of tests (Kelley, 2007).

What is less straightforward is to justify how a desired level of accuracy is related to inferential goals. There is no literature that helps researchers to choose a desired width of the confidence interval. Morey (2020) convincingly argues that most practical use-cases of planning for precision involve an inferential goal of distinguishing an observed effect from other effect sizes (for a Bayesian perspective, see Kruschke (2018)). For example, a researcher might expect an effect size of r = 0.4 and would treat observed correlations that differ more than 0.2 (i.e., $0.2 < r < 0.6$) differently, in that effects of r = 0.6 or larger are considered too large to be caused by the assumed underlying mechanism (Hilgard, 2021), while effects smaller than r = 0.2 are considered too small to support the theoretical prediction. If the goal is indeed to get an effect size estimate that is precise enough so that two effects can be differentiated with high probability, the inferential goal is actually a hypothesis test, which requires designing a study with sufficient power to reject effects (e.g., testing a range prediction of correlations between 0.2 and 0.6).

If researchers do not want to test a hypothesis, for example because they prefer an estimation approach over a testing approach, then in the absence of clear guidelines that help researchers to justify a desired level of precision, one solution might be to rely on a generally accepted norm of precision to aim for. Just as researchers normatively use an alpha level of 0.05, they could plan studies to achieve a desired confidence interval width around the observed effect that is determined by a norm. Future work is needed to help researchers choose a confidence interval width when planning for accuracy.

Heuristics

When a researcher uses a heuristic, they are not able to justify their sample size themselves, but they trust in a sample size recommended by some authority. When I started as a PhD student in 2005 it was common to collect 15 participants in each between subject condition. When asked why this was a common practice, no one

was really sure, but people trusted there was a justification somewhere in the literature. Now, I realize there was no justification for the heuristics we used. As Berkeley (1735) already observed: "Men learn the elements of science from others: And every learner hath a deference more or less to authority, especially the young learners, few of that kind caring to dwell long upon principles, but inclining rather to take them upon trust: And things early admitted by repetition become familiar: And this familiarity at length passeth for evidence."

Some papers provide researchers with simple rules of thumb about the sample size that should be collected. Such papers clearly fill a need, and are cited a lot, even when the advice in these articles is flawed. For example, Wilson VanVoorhis and Morgan (2007) translate an absolute *minimum* of 50+8 observations for regression analyses suggested by a rule of thumb examined in Green (1991) into the recommendation to collect ~50 observations. Green actually concludes in his article that "In summary, no specific minimum number of subjects or minimum ratio of subjects-to-predictors was supported". He does discuss how a general rule of thumb of N = 50 + 8 provided an accurate minimum number of observations for the 'typical' study in the social sciences because these have a 'medium' effect size, as Green claims by citing Cohen (1988). Cohen actually didn't claim that the typical study in the social sciences has a 'medium' effect size, and instead said (1988, p. 13): "Many effects sought in personality, social, and clinical-psychological research are likely to be small effects as here defined". We see how a string of mis-citations eventually leads to a misleading rule of thumb.

Rules of thumb seem to primarily emerge due to mis-citations and/or overly simplistic recommendations. Simonsohn, Nelson, and Simmons (2011) recommended that "Authors must collect at least 20 observations per cell". A later recommendation by the same authors presented at a conference suggested to use n > 50, unless you study large effects (Simmons, Nelson, & Simonsohn, 2013). Regrettably, this advice is now often miscited as a justification to collect no more than 50 observations per condition without considering the expected effect size. Schönbrodt and Perugini (2013) examined at which sample size correlations stabilize and suggest sample sizes ranging from 20 to 470 depending on expectations about the true effect size, the desired maximum deviation from the true effect, and the desired stability of the estimate. And yet, many researchers simply follow the summary recommendation in the abstract that "Results indicate that in typical scenarios the sample size should approach 250 for stable estimates". Viechtbauer et al. (2015) proposed an approach to compute the sample size for pilot studies where one is interested in identifying any issues that exist in a small percentage of participants. In their abstract, they mentioned as an example the sample size of 59 to detect a problem that occurs with a 5% prevalence with 95% confidence. Most researchers directly use this number, instead of computing a sample size based on the assumed prevalence and desired confidence level. If authors justify a specific sample size (e.g., n = 50) based on a general recommendation in another paper, either they are mis-citing the paper, or the paper they are citing is flawed.

Another common heuristic is to collect the same number of observations as were collected in a previous study. This strategy is not recommended in scientific disciplines with widespread publication bias, and/or where novel and surprising findings from largely exploratory single studies are published. Using the same sample size as a previous study is only a valid approach if the sample size justification in the previous study also applies to the current study. Instead of stating that you intend to collect the same sample size as an earlier study, repeat the sample size justification, and update it in light of any new information (such as the effect size in the earlier study, see Table 6).

Peer reviewers and editors should be extremely skeptical of rules of thumb sample size justifications. Whenever one encounters a sample size justification based on a heuristic, ask yourself: 'Why is this heuristic used?' It is important to know what the logic behind a heuristic is to determine whether the heuristic is valid for a specific situation. In most cases, heuristics are based on weak logic, and not widely applicable. It might be possible that fields develop valid heuristics for sample size justifications. For example, it is possible that a research area reaches widespread agreement that effects smaller than $d = 0.3$ are too small to be of interest, and all studies in a field use sequential designs (see below) that have 90% power to detect a $d = 0.3$. Alternatively, it is possible that a field agrees that data should be collected with a desired level of accuracy, irrespective of the true effect size. In these cases, valid heuristics would exist based on generally agreed goals of data collection. For example, Simonsohn (2015) suggests to design replication studies that have 2.5 times as large sample sizes as the original study, as this provides 80% power for an equivalence test against an equivalence bound set to the effect the original study had 33% power to detect, assuming the true effect size is 0. As original authors typically do not specify which effect size would falsify their hypothesis, the heuristic underlying this 'small

telescopes' approach is a good starting point for a replication study with the inferential goal to reject the presence of an effect as large as was described in an earlier publication. It is the responsibility of researchers to gain the knowledge to distinguish valid heuristics from mindless heuristics.

No Justification

It might sound like a *contradictio in terminis*, but it is useful to distinguish a final category where researchers explicitly state they do not have a justification for their sample size. Perhaps the resources were available to collect more data, but they were not used. A researcher could have performed a power analysis, or planned for precision, but they did not. In those cases, instead of pretending there was a justification for the sample size, honesty requires you to state there is no sample size justification. This is not necessarily bad. It is still possible to discuss the smallest effect size of interest, the minimal statistically detectable effect, the width of the confidence interval around the effect size, and to plot a sensitivity power analysis, in relation to the sample size that was collected. If a researcher truly had no specific inferential goals when collecting the data, such an evaluation can perhaps be performed based on reasonable inferential goals peers would have when they learn about the existence of the collected data.

Do not try to spin a story where it looks like a study was highly informative when it was not. Instead, transparently evaluate how informative the study was given effect sizes that were of interest, and make sure that the conclusions follow from the data. The lack of a sample size justification might not be problematic, but it might mean that a study was not informative for most effect sizes of interest, which makes it especially difficult to interpret non-significant effects, or estimates with large uncertainty.

What is Your Inferential Goal?

The inferential goal of data collection is often in some way related to the size of an effect. Therefore, to design an informative study, researchers will want to think about which effect sizes are interesting. First, it is useful to consider three effect sizes when determining the sample size. The first is the smallest effect size a researcher is interested in, the second is the smallest effect size that can be statistically significant (only in studies where a significance test will be performed), and the third is the effect size that is expected. Beyond considering these three effect sizes, it can be useful to evaluate ranges of effect sizes. This can be done by computing the width of the expected confidence interval around an effect size of interest (for example, an effect size of zero), and examine which effects could be rejected. Similarly, it can be useful to plot a sensitivity curve and evaluate the range of effect sizes the design has decent power to detect, as well as to consider the range of effects for which the design has low power. Finally, there are situations where it is useful to consider range of effects that are likely to be observed in a specific research area.

What is the Smallest Effect Size of Interest?

The strongest possible sample size justification is based on an explicit statement of the smallest effect size that is considered interesting. A smallest effect size of interest can be based on theoretical predictions or practical considerations. For a review of approaches that can be used to determine a smallest effect size of interest in randomized controlled trials, see Cook et al. (2014) and Keefe et al. (2013), for reviews of different methods to determine a smallest effect size of interest, see King (2011) and Copay, Subach, Glassman, Polly, and Schuler (2007), and for a discussion focused on psychological research, see Lakens et al. (2018).

It can be challenging to determine the smallest effect size of interest whenever theories are not very developed, or when the research question is far removed from practical applications, but it is still worth thinking about which effects would be too small to matter. A first step forward is to discuss which effect sizes are considered meaningful in a specific research line with your peers. Researchers will differ in the effect sizes they consider large enough to be worthwhile (Murphy et al., 2014). Just as not every scientist will find every research question interesting enough to study, not every scientist will consider the same effect sizes interesting enough to study, and different stakeholders will differ in which effect sizes are considered meaningful (Kelley & Preacher, 2012).

Even though it might be challenging, there are important benefits of being able to specify a smallest effect size of interest. The population effect size is always uncertain (indeed, estimating this is typically one of the goals of the study), and therefore whenever a study is powered for an expected effect size, there is considerable uncertainty about whether the statistical power is high enough to detect the true effect in the population. However, if the smallest effect size of interest is can be specified and agreed upon after careful deliberation, it becomes possible to design a study that has sufficient power (given the inferential goal to detect of reject the smallest effect size of interest with a certain error rate). A smallest effect of interest may be subjective (one researcher might find effect sizes smaller than d

= 0.3 meaningless, while another researcher might still be interested in effects larger than $d = 0.1$), and there might be uncertainty about the parameters required to specify the smallest effect size of interest (e.g., when performing a cost-benefit analysis), but after a smallest effect size of interest has been determined, a study can be designed with a known Type 2 error rate to detect or reject this value. For this reason an a-priori power based on a smallest effect size of interest is generally preferred, whenever researchers are able to specify one (Aberson, 2019; Albers & Lakens, 2018; Brown, 1983; Cascio & Zedeck, 1983; Dienes, 2014; Lenth, 2001).

The Minimal Statistically Detectable Effect

The minimal statistically detectable effect, or the critical effect size, provides information about the smallest effect size that, if observed, would be statistically significant given a specified alpha level and sample size (Cook et al., 2014). For any critical t value (e.g., $t = 1.96$ for $\alpha = 0.05$, for large sample sizes) we can compute a critical mean difference (Phillips et al., 2001), or a critical standardized effect size. For a two-sided independent t test the critical mean difference is:

$$M_{crit} = t_{crit} \sqrt{\frac{sd_1^2}{n_1} + \frac{sd_2^2}{n_2}}$$

and the critical standardized mean difference is:

$$d_{crit} = t_{crit} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

In Figure 5 the distribution of Cohen's d is plotted for 15 participants per group when the true effect size is either $d = 0$ or $d = 0.5$. This figure is similar to Figure 3, with the addition that the critical d is indicated. We see that with such a small number of observations in each group only observed effects larger than $d = 0.75$ will be statistically significant. Whether such effect sizes are interesting, and can realistically be expected, should be carefully considered and justified.

G*Power provides the critical test statistic (such as the critical t value) when performing a power analysis. For example, Figure 6 shows that for a correlation based on a two-sided test, with $\alpha = 0.05$, and $N = 30$, only effects larger than $r = 0.361$ or smaller than $r = -0.361$ can be statistically significant. This reveals that when the sample size is relatively small, the observed effect needs to be quite substantial to be statistically significant.

It is important to realize that due to random variation each study has a probability to yield effects larger than

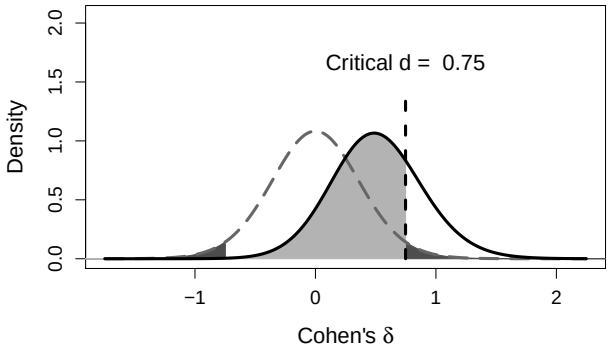

Figure 5. Critical effect size for an independent t test with n = 15 per group and $\alpha = 0.05$.

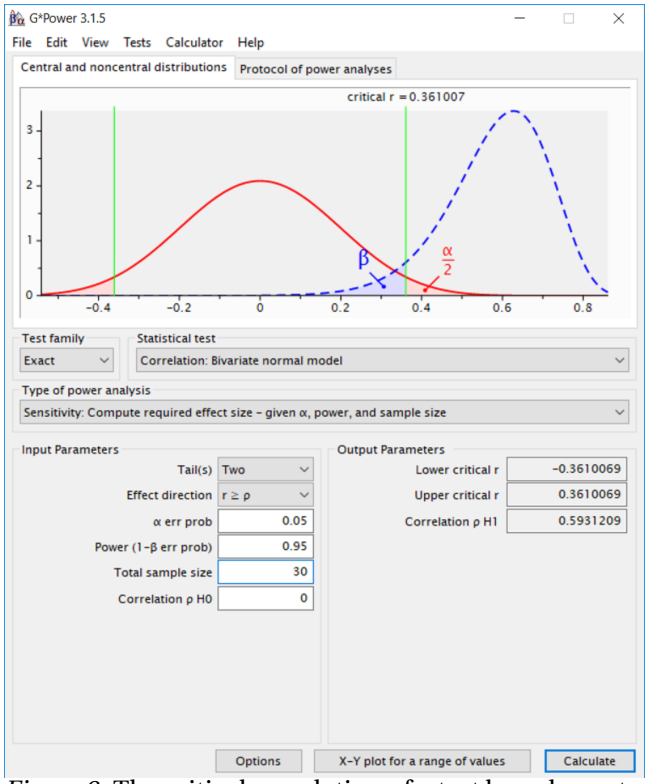

Figure 6. The critical correlation of a test based on a total sample size of 30 and $\alpha = 0.05$ calculated in G*Power.

the critical effect size, even if the true effect size is small (or even when the true effect size is 0, in which case each significant effect is a Type I error). Computing a minimal statistically detectable effect is useful for a study where no a-priori power analysis is performed, both for studies in the published literature that do not report a sample size justification (Lakens et al., 2018), as for researchers who rely on heuristics for their sample

size justification.

It can be informative to ask yourself whether the critical effect size for a study design is within the range of effect sizes that can realistically be expected. If not, then whenever a significant effect is observed in a published study, either the effect size is surprisingly larger than expected, or more likely, it is an upwardly biased effect size estimate. In the latter case, given publication bias, published studies will lead to biased effect size estimates. If it is still possible to increase the sample size, for example by ignoring rules of thumb and instead performing an a-priori power analysis, then do so. If it is not possible to increase the sample size, for example due to resource constraints, then reflecting on the minimal statistically detectable effect should make it clear that an analysis of the data should not focus on p values, but on the effect size and the confidence interval (see Table 3).

It is also useful to compute the minimal statistically detectable effect if an 'optimistic' power analysis is performed. For example, if you believe a best case scenario for the true effect size is $d = 0.57$ and use this optimistic expectation in an a-priori power analysis, effects smaller than $d = 0.4$ will not be statistically significant when you collect 50 observations in a two independent group design. If your worst case scenario for the alternative hypothesis is a true effect size of $d = 0.35$ your design would not allow you to declare a significant effect if effect size estimates close to the worst case scenario are observed. Taking into account the minimal statistically detectable effect size should make you reflect on whether a hypothesis test will yield an informative answer, and whether your current approach to sample size justification (e.g., the use of rules of thumb, or letting resource constraints determine the sample size) leads to an informative study, or not.

What is the Expected Effect Size?

Although the true population effect size is always unknown, there are situations where researchers have a reasonable expectation of the effect size in a study, and want to use this expected effect size in an a-priori power analysis. Even if expectations for the observed effect size are largely a guess, it is always useful to explicitly consider which effect size are expected. A researcher can justify a sample size based on the effect size they expect, even if such a study would not be very informative with respect to the smallest effect size of interest. In such cases a study is informative for one inferential goal (testing whether the expected effect size is present or absent), but not highly informative for the second goal (testing whether the smallest effect size of interest is present or absent).

There are typically three sources for expectations about the population effect size: a meta-analysis, a previous study, or a theoretical model. It is tempting for researchers to be overly optimistic about the expected effect size in an a-priori power analysis, as higher effect size estimates yield lower sample sizes, but being too optimistic increases the probability of observing a false negative result. When reviewing a sample size justification based on an a-priori power analysis, it is important to critically evaluate the justification for the expected effect size used in power analyses.

Using an Estimate from a Meta-Analysis

In a perfect world effect size estimates from a meta-analysis would provide researchers with the most accurate information about which effect size they could expect. Due to widespread publication bias in science, effect size estimates from meta-analyses are regrettably not always accurate. They can be biased, sometimes substantially so. Furthermore, meta-analyses typically have considerable heterogeneity, which means that the meta-analytic effect size estimate differs for subsets of studies that make up the meta-analysis. So, although it might seem useful to use a meta-analytic effect size estimate of the effect you are studying in your power analysis, you need to take great care before doing so.

If a researcher wants to enter a meta-analytic effect size estimate in an a-priori power analysis, they need to consider three things (see Table 5). First, the studies included in the meta-analysis should be similar enough to the study they are performing that it is reasonable to expect a similar effect size. In essence, this requires evaluating the generalizability of the effect size estimate to the new study. It is important to carefully consider differences between the meta-analyzed studies and the planned study, with respect to the manipulation, the measure, the population, and any other relevant variables.

Second, researchers should check whether the effect sizes reported in the meta-analysis are homogeneous. If not, and there is considerable heterogeneity in the meta-analysis, it means not all included studies can be expected to have the same true effect size estimate. A meta-analytic estimate should be used based on the subset of studies that most closely represent the planned study. Note that heterogeneity remains a possibility (even direct replication studies can show heterogeneity when unmeasured variables moderate the effect size in each sample (Kenny & Judd, 2019; Olsson-Collentine, Wicherts, & van Assen, 2020)), so the main

goal of selecting similar studies is to use existing data to increase the probability that your expectation is accurate, without guaranteeing it will be.

Third, the meta-analytic effect size estimate should not be biased. Check if the bias detection tests that are reported in the meta-analysis are state-of-the-art, or perform multiple bias detection tests yourself (Carter, Schönbrodt, Gervais, & Hilgard, 2019), and consider bias corrected effect size estimates (even though these estimates might still be biased, and do not necessarily reflect the true population effect size).

Using an Estimate from a Previous Study

If a meta-analysis is not available, researchers often rely on an effect size from a previous study in an a-priori power analysis. The first issue that requires careful attention is whether the two studies are sufficiently similar. Just as when using an effect size estimate from a meta-analysis, researchers should consider if there are differences between the studies in terms of the population, the design, the manipulations, the measures, or other factors that should lead one to expect a different effect size. For example, intra-individual reaction time variability increases with age, and therefore a study performed on older participants should expect a smaller standardized effect size than a study performed on younger participants. If an earlier study used a very strong manipulation, and you plan to use a more subtle manipulation, a smaller effect size should be expected. Finally, effect sizes do not generalize to studies with different designs. For example, the effect size for a comparison between two groups is most often not similar to the effect size for an interaction in a follow-up study where a second factor is added to the original design (Lakens & Caldwell, 2019).

Even if a study is sufficiently similar, statisticians have warned against using effect size estimates from small pilot studies in power analyses. Leon, Davis, and Kraemer (2011) write:

> Contrary to tradition, a pilot study does not provide a meaningful effect size estimate for planning subsequent studies due to the imprecision inherent in data from small samples.

The two main reasons researchers should be careful when using effect sizes from studies in the published literature in power analyses is that effect size estimates from studies can differ from the true population effect size due to random variation, and that publication bias inflates effect sizes. Figure 7 shows the distribution for η_p^2 for a study with three conditions with 25 participants in each condition when the null hypothesis is true and when there is a 'medium' true effect of $\eta_p^2 = 0.0588$ (Richardson, 2011). As in Figure 5 the critical effect size is indicated, which shows observed effects smaller than $\eta_p^2 = 0.08$ will not be significant with the given sample size. If the null hypothesis is true effects larger than $\eta_p^2 = 0.08$ will be a Type I error (the dark grey area), and when the alternative hypothesis is true effects smaller than $\eta_p^2 = 0.08$ will be a Type II error (light grey area). It is clear all significant effects are larger than the true effect size ($\eta_p^2 = 0.0588$), so power analyses based on a significant finding (e.g., because only significant results are published in the literature) will be based on an overestimate of the true effect size, introducing bias.

But even if we had access to all effect sizes (e.g., from pilot studies you have performed yourself) due to random variation the observed effect size will sometimes be quite small. Figure 7 shows it is quite likely to observe an effect of $\eta_p^2 = 0.01$ in a small pilot study, even when the true effect size is 0.0588. Entering an effect size estimate of $\eta_p^2 = 0.01$ in an a-priori power analysis would suggest a total sample size of 957 observations to achieve 80% power in a follow-up study. If researchers only follow up on pilot studies when they observe an effect size in the pilot study that, when entered into a power analysis, yields a sample size that is feasible to collect for the follow-up study, these effect size estimates will be upwardly biased, and power in the follow-up study will be systematically lower than desired (Albers & Lakens, 2018).

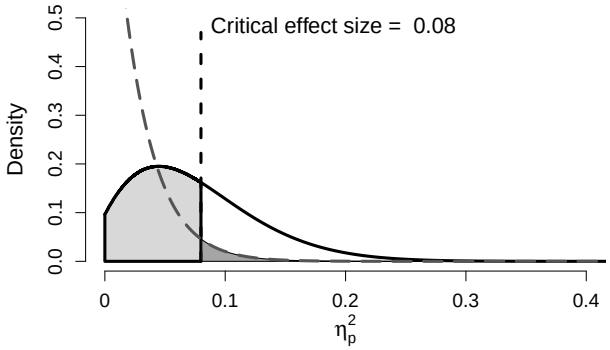

Figure 7. Distribution of partial eta squared under the null hypothesis (dotted grey curve) and a medium true effect of 0.0588 (solid black curve) for 3 groups with 25 observations.

In essence, the problem with using small studies to estimate the effect size that will be entered into an a-priori power analysis is that due to publication bias or follow-

Table 5
Overview of recommendations when justifying the use of a meta-analytic effect size estimate for a power analysis.

What to take into account	How to take it into account?
Are the studies in the meta-analysis similar?	Are the studies in the meta-analyses very similar in design, measures, and the population to the study you are planning? Evaluate the generalizability of the effect size estimate to your study.
Are the studies in the meta-analysis homogeneous?	Is there heterogeneity in the meta-analysis? If so, use the meta-analytic effect size estimate of the most relevant homogenous subsample.
Is the effect size estimate unbiased?	Did the original study report bias detection tests, and was there bias? If so, it might be wise to use a more conservative effect size estimate, based on bias correction techniques while acknowledging these corrected effect size estimates might not represent the true meta-analytic effect size estimate.

up bias the effect sizes researchers end up using for their power analysis do not come from a full F distribution, but from what is known as a *truncated F* distribution (Taylor & Muller, 1996). For example, imagine if there is be extreme publication bias in the situation illustrated in Figure 7. The only studies that would be accessible to researchers would come from the part of the distribution where $\eta_p^2 > 0.08$, and the test result would be statistically significant. It is possible to compute an effect size estimate that, based on certain assumptions, corrects for bias. For example, imagine we observe a result in the literature for a One-Way ANOVA with 3 conditions, reported as $F(2, 42) = 0.017$, $\eta_p^2 = 0.176$. If we would take this effect size at face value and enter it as our effect size estimate in an a-priori power analysis, the suggested sample size to achieve would suggest we need to collect 17 observations in each condition.

However, if we assume bias is present, we can use the `BUCSS` R package (Anderson, Kelley, & Maxwell, 2017) to perform a power analysis that attempts to correct for bias. A power analysis that takes bias into account (under a specific model of publication bias, based on a truncated F distribution where only significant results are published) suggests collecting 73 participants in each condition. It is possible that the bias corrected estimate of the non-centrality parameter used to compute power is zero, in which case it is not possible to correct for bias using this method. As an alternative to formally modeling a correction for publication bias whenever researchers assume an effect size estimate is biased, researchers can simply use a more conservative effect size estimate, for example by computing power based on the lower limit of 60% two-sided confidence interval around the effect size estimate, which Perugini, Gallucci, and Costantini (2014) refer to as *safeguard power*. Both these approaches lead to a more conservative power analysis, but not necessarily a more accurate power analysis. It is simply not possible to perform an accurate power analysis on the basis of an effect size estimate from a study that might be biased and/or had a small sample size (Teare et al., 2014). If it is not possible to specify a smallest effect size of interest, and there is great uncertainty about which effect size to expect, it might be more efficient to perform a study with a sequential design (discussed below).

To summarize, an effect size from a previous study in an a-priori power analysis can be used if three conditions are met (see Table 6). First, the previous study is sufficiently similar to the planned study. Second, there was a low risk of bias (e.g., the effect size estimate comes from a Registered Report, or from an analysis for which results would not have impacted the likelihood of publication). Third, the sample size is large enough to yield relatively accurate effect size estimate, based on the width of a 95% CI around the observed effect size estimate. There is always uncertainty around the effect size estimate, and entering the upper and lower limit of the 95% CI around the effect size estimate might be informative about the consequences of the uncertainty in the effect size estimate for an a-priori power analysis.

Using an Estimate from a Theoretical Model

When your theoretical model is sufficiently specific such that you can build a computational model, and you have knowledge about key parameters in your model that are relevant for the data you plan to collect, it is possible to estimate an effect size based on the effect size estimate derived from a computational model. For example, if one had strong ideas about the weights for each feature stimuli share and differ on, it could be possible to compute predicted similarity judgments for pairs of stimuli based on Tversky's contrast model

Table 6
Overview of recommendations when justifying the use of an effect size estimate from a single study.

What to take into account	How to take it into account?
Is the study sufficiently similar?	Consider if there are differences between the studies in terms of the population, the design, the manipulations, the measures, or other factors that should lead one to expect a different effect size.
Is there a risk of bias?	Evaluate the possibility that if the effect size estimate had been smaller, you would not have used it (or it would not have been published). Examine the difference when entering the reported, and a bias corrected, effect size estimate in a power analysis.
How large is the uncertainty?	Studies with a small number of observations have large uncertainty. Consider the possibility of using a more conservative effect size estimate to reduce the possibility of an underpowered study for the true effect size (such as a safeguard power analysis).

(Tversky, 1977), and estimate the predicted effect size for differences between experimental conditions. Although computational models that make point predictions are relatively rare, whenever they are available, they provide a strong justification of the effect size a researcher expects.

Compute the Width of the Confidence Interval around the Effect Size

If a researcher can estimate the standard deviation of the observations that will be collected, it is possible to compute an a-priori estimate of the width of the 95% confidence interval around an effect size (Kelley, 2007). Confidence intervals represent a range around an estimate that is wide enough so that in the long run the true population parameter will fall inside the confidence intervals 100 - α percent of the time. In any single study the true population effect either falls in the confidence interval, or it doesn't, but in the long run one can *act* as if the confidence interval includes the true population effect size (while keeping the error rate in mind). Cumming (2013) calls the difference between the observed effect size and the upper 95% confidence interval (or the lower 95% confidence interval) the margin of error.

If we compute the 95% CI for an effect size of $d = 0$ based on the t statistic and sample size (Smithson, 2003), we see that with 15 observations in each condition of an independent t test the 95% CI ranges from $d = -0.72$ to $d = 0.72$[5]. The margin of error is half the width of the 95% CI, 0.72. A Bayesian estimator who uses an uninformative prior would compute a credible interval with the same (or a very similar) upper and lower bound (Albers et al., 2018; Kruschke, 2011), and might conclude that after collecting the data they would be left with a range of plausible values for the population effect that is too large to be informative. Regardless of the statistical philosophy you plan to rely on when analyzing the data, the evaluation of what we can conclude based on the width of our interval tells us that with 15 observation per group we will not learn a lot.

One useful way of interpreting the width of the confidence interval is based on the effects you would be able to reject if the true effect size is 0. In other words, if there is no effect, which effects would you have been able to reject given the collected data, and which effect sizes would not be rejected, if there was no effect? Effect sizes in the range of $d = 0.7$ are findings such as "People become aggressive when they are provoked", "People prefer their own group to other groups", and "Romantic partners resemble one another in physical attractiveness" (Richard, Bond, & Stokes-Zoota, 2003). The width of the confidence interval tells you that you can only reject the presence of effects that are so large, if they existed, you would probably already have noticed them. If it is true that most effects that you study are realistically much smaller than $d = 0.7$, there is a good possibility that we do not learn anything we didn't already know by performing a study with n = 15. Even without data, in most research lines we would not consider certain large effects plausible (although the effect sizes that are plausible differ between fields, as discussed below). On the other hand, in large samples where researchers can for example reject the presence of effects larger than $d = 0.2$, if the null hypothesis was true, this analysis of the width of the confidence interval would suggest that peers in many research lines would likely consider the study to be informative.

[5]Confidence intervals around effect sizes can be computed using the MOTE Shiny app: https://www.aggieerin.com/shiny-server/

We see that the margin of error is almost, but not exactly, the same as the minimal statistically detectable effect (d = 0.75). The small variation is due to the fact that the 95% confidence interval is calculated based on the t distribution. If the true effect size is not zero, the confidence interval is calculated based on the non-central t distribution, and the 95% CI is asymmetric. Figure 8 visualizes three t distributions, one symmetric at 0, and two asymmetric distributions with a noncentrality parameter (the normalized difference between the means) of 2 and 3. The asymmetry is most clearly visible in very small samples (the distributions in the plot have 5 degrees of freedom) but remains noticeable in larger samples when calculating confidence intervals and statistical power. For example, for a true effect size of d = 0.5 observed with 15 observations per group would yield d_s = 0.50, 95% CI [-0.23, 1.22]. If we compute the 95% CI around the critical effect size, we would get d_s = 0.75, 95% CI [0.00, 1.48]. We see the 95% CI ranges from exactly 1.48 to 0.00, in line with the relation between a confidence interval and a p value, where the 95% CI excludes zero if the test is statistically significant. As noted before, the different approaches recommended here to evaluate how informative a study is are often based on the same information and should be seen as different ways to approach the same question.

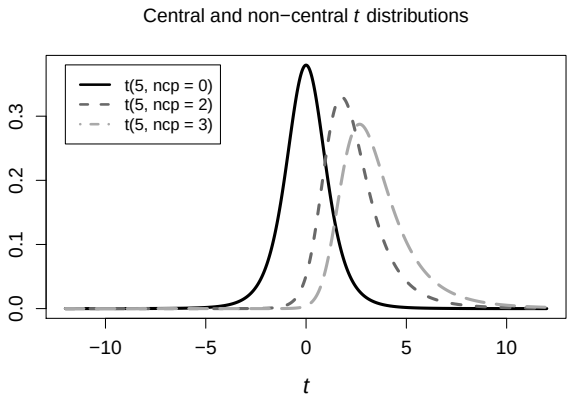

Figure 8. Central (black) and 2 non-central (darkgrey and lightgrey) t distributions.

Plot a Sensitivity Power Analysis

A sensitivity power analysis fixes the sample size, desired power, and alpha level, and answers the question which effect size a study could detect with a desired power. A sensitivity power analysis is therefore performed when the sample size is already known. Sometimes data has already been collected to answer a different research question, or the data is retrieved from an existing database, and you want to perform a sensitivity power analysis for a new statistical analysis. Other times, you might not have carefully considered the sample size when you initially collected the data, and want to reflect on the statistical power of the study for (ranges of) effect sizes of interest when analyzing the results. Finally, it is possible that the sample size will be collected in the future, but you know that due to resource constraints the maximum sample size you can collect is limited, and you want to reflect on whether the study has sufficient power for effects that you consider plausible and interesting (such as the smallest effect size of interest, or the effect size that is expected).

Assume a researcher plans to perform a study where 30 observations will be collected in total, 15 in each between participant condition. Figure 9 shows how to perform a sensitivity power analysis in G*Power for a study where we have decided to use an alpha level of 5%, and desire 90% power. The sensitivity power analysis reveals the designed study has 90% power to detect effects of at least d = 1.23. Perhaps a researcher believes that a desired power of 90% is quite high, and is of the opinion that it would still be interesting to perform a study if the statistical power was lower. It can then be useful to plot a sensitivity curve across a range of smaller effect sizes.

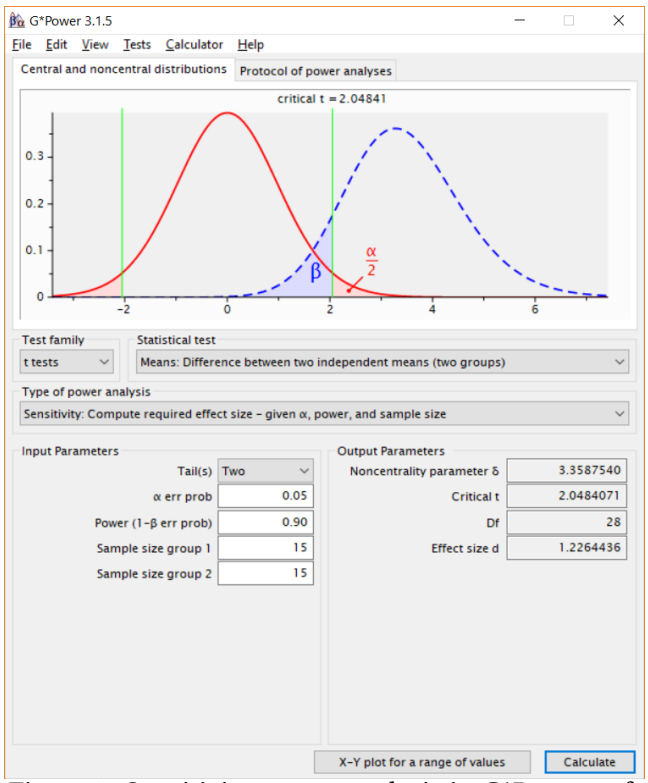

Figure 9. Sensitivity power analysis in G*Power software.

The two dimensions of interest in a sensitivity power

analysis are the effect sizes, and the power to observe a significant effect assuming a specific effect size. These two dimensions can be plotted against each other to create a sensitivity curve. For example, a sensitivity curve can be plotted in G*Power by clicking the 'X-Y plot for a range of values' button, as illustrated in Figure 10. Researchers can examine which power they would have for an a-priori plausible range of effect sizes, or they can examine which effect sizes would provide reasonable levels of power. In simulation-based approaches to power analysis, sensitivity curves can be created by performing the power analysis for a range of possible effect sizes. Even if 50% power is deemed acceptable (in which case deciding to act as if the null hypothesis is true after a non-significant result is a relatively noisy decision procedure), Figure 10 shows a study design where power is extremely low for a large range of effect sizes that are reasonable to expect in most fields. Thus, a sensitivity power analysis provides an additional approach to evaluate how informative the planned study is, and can inform researchers that a specific design is unlikely to yield a significant effect for a range of effects that one might realistically expect.

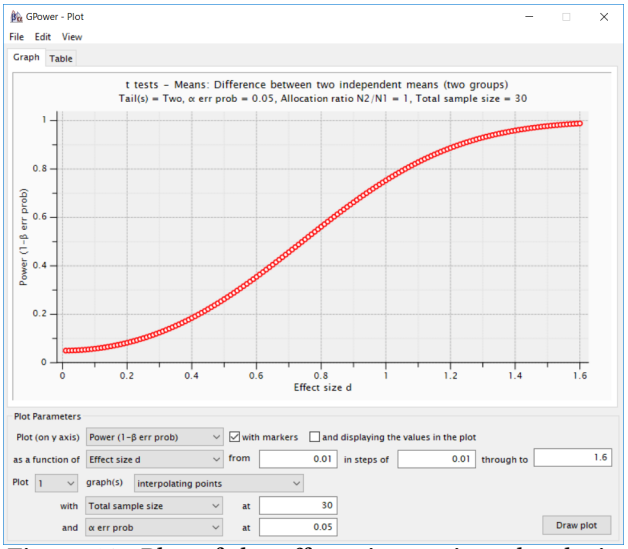

Figure 10. Plot of the effect size against the desired power when n = 15 per group and alpha = 0.05.

If the number of observations per group had been larger, the evaluation might have been more positive. We might not have had any specific effect size in mind, but if we had collected 150 observations per group, a sensitivity analysis could have shown that power was sufficient for a range of effects we believe is most interesting to examine, and we would still have approximately 50% power for quite small effects. For a sensitivity analysis to be meaningful, the sensitivity curve should be compared against a smallest effect size of interest, or a range of effect sizes that are expected. A sensitivity power analysis has no clear cut-offs to examine (Bacchetti, 2010). Instead, the idea is to make a holistic trade-off between different effect sizes one might observe or care about, and their associated statistical power.

The Distribution of Effect Sizes in a Research Area

In my personal experience the most commonly entered effect size estimate in an a-priori power analysis for an independent t test is Cohen's benchmark for a 'medium' effect size, because of what is known as the *default effect*. When you open G*Power, a 'medium' effect is the default option for an a-priori power analysis. Cohen's benchmarks for small, medium, and large effects should not be used in an a-priori power analysis (Cook et al., 2014; Correll, Mellinger, McClelland, & Judd, 2020), and Cohen regretted having proposed these benchmarks (Funder & Ozer, 2019). The large variety in research topics means that any 'default' or 'heuristic' that is used to compute statistical power is not just unlikely to correspond to your actual situation, but it is also likely to lead to a sample size that is more substantially misaligned with the question you are trying to answer with the collected data.

Some researchers have wondered what a better default would be, if researchers have no other basis to decide upon an effect size for an a-priori power analysis. Brysbaert (2019) recommends d = 0.4 as a default in psychology, which is the average observed in replication projects and several meta-analyses. It is impossible to know if this average effect size is realistic, but it is clear there is huge heterogeneity across fields and research questions. Any average effect size will often deviate substantially from the effect size that should be expected in a planned study. Some researchers have suggested to change Cohen's benchmarks based on the distribution of effect sizes in a specific field (Bosco, Aguinis, Singh, Field, & Pierce, 2015; Funder & Ozer, 2019; Hill, Bloom, Black, & Lipsey, 2008; Kraft, 2020; Lovakov & Agadullina, 2017). As always, when effect size estimates are based on the published literature, one needs to evaluate the possibility that the effect size estimates are inflated due to publication bias. Due to the large variation in effect sizes within a specific research area, there is little use in choosing a large, medium, or small effect size benchmark based on the empirical distribution of effect sizes in a field to perform a power analysis.

Having some knowledge about the distribution of effect sizes in the literature can be useful when interpreting the confidence interval around an effect size. If in a specific research area almost no effects are larger than the value you could reject in an equivalence test (e.g.,

if the observed effect size is 0, the design would only reject effects larger than for example $d = 0.7$), then it is a-priori unlikely that collecting the data would tell you something you didn't already know.

It is more difficult to defend the use of a specific effect size derived from an empirical distribution of effect sizes as a justification for the effect size used in an a-priori power analysis. One might argue that the use of an effect size benchmark based on the distribution of effects in the literature will outperform a wild guess, but this is not a strong enough argument to form the basis of a sample size justification. There is a point where researchers need to admit they are not ready to perform an a-priori power analysis due to a lack of clear expectations (Scheel, Tiokhin, Isager, & Lakens, 2020). Alternative sample size justifications, such as a justification of the sample size based on resource constraints, perhaps in combination with a sequential study design, might be more in line with the actual inferential goals of a study.

Additional Considerations When Designing an Informative Study

So far, the focus has been on justifying the sample size for quantitative studies. There are a number of related topics that can be useful to design an informative study. First, in addition to a-priori power analysis and sensitivity power analysis, it is important to discuss compromise power analysis (which is useful) and post-hoc power analysis (which is not useful). When sample sizes are justified based on an a-priori power analysis it can be very efficient to collect data in sequential designs where data collection is continued or terminated based on interim analyses of the data. Furthermore, it is worthwhile to consider ways to increase the power of a test without increasing the sample size. An additional point of attention is to have a good understanding of your dependent variable, especially it's standard deviation. Finally, sample size justification is just as important in qualitative studies, and although there has been much less work on sample size justification in this domain, some proposals exist that researchers can use to design an informative study. Each of these topics is discussed in turn.

Compromise Power Analysis

In a compromise power analysis the sample size and the effect are fixed, and the error rates of the test are calculated, based on a desired ratio between the Type I and Type II error rate. A compromise power analysis is useful both when a very large number of observations will be collected, as when only a small number of observations can be collected.

In the first situation a researcher might be fortunate enough to be able to collect so many observations that the statistical power for a test is very high for all effect sizes that are deemed interesting. For example, imagine a researcher has access to 2000 employees who are all required to answer questions during a yearly evaluation in a company where they are testing an intervention that should reduce subjectively reported stress levels. You are quite confident that an effect smaller than $d = 0.2$ is not large enough to be subjectively noticeable for individuals (Jaeschke, Singer, & Guyatt, 1989). With an alpha level of 0.05 the researcher would have a statistical power of 0.99, or a Type II error rate of 0.01. This means that for a smallest effect size of interest of $d = 0.2$ the researcher is 8.30 times more likely to make a Type I error than a Type II error.

Although the original idea of designing studies that control Type I and Type II error error rates was that researchers would need to justify their error rates (Neyman & Pearson, 1933), a common heuristic is to set the Type I error rate to 0.05 and the Type II error rate to 0.20, meaning that a Type I error is 4 times as unlikely as a Type II error. The default use of 80% power (or a 20% Type II or β error) is based on a personal preference of Cohen (1988), who writes:

> It is proposed here as a convention that, when the investigator has no other basis for setting the desired power value, the value .80 be used. This means that β is set at .20. This arbitrary but reasonable value is offered for several reasons (Cohen, 1965, pp. 98-99). The chief among them takes into consideration the implicit convention for α of .05. The β of .20 is chosen with the idea that the general relative seriousness of these two kinds of errors is of the order of .20/.05, i.e., that Type I errors are of the order of four times as serious as Type II errors. This .80 desired power convention is offered with the hope that it will be ignored whenever an investigator can find a basis in his substantive concerns in his specific research investigation to choose a value ad hoc.

We see that conventions are built on conventions: the norm to aim for 80% power is built on the norm to set the alpha level at 5%. What we should take away from Cohen is not that we should aim for 80% power, but that we should justify our error rates based on the relative seriousness of each error. This is where compromise power analysis comes in. If you share Cohen's belief

that a Type I error is 4 times as serious as a Type II error, and building on our earlier study on 2000 employees, it makes sense to adjust the Type I error rate when the Type II error rate is low for all effect sizes of interest (Cascio & Zedeck, 1983). Indeed, Erdfelder, Faul, and Buchner (1996) created the G*Power software in part to give researchers a tool to perform compromise power analysis.

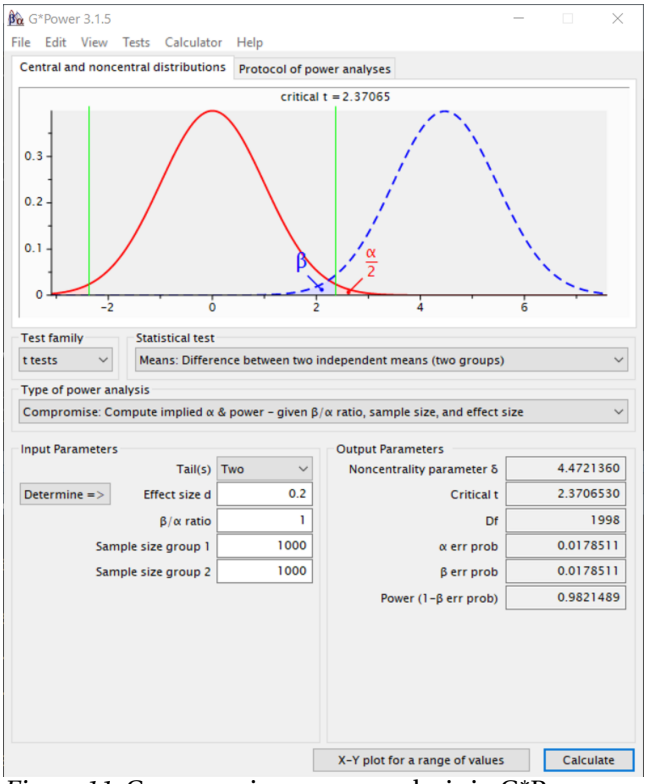

Figure 11. Compromise power analysis in G*Power.

Figure 11 illustrates how a compromise power analysis is performed in G*Power when a Type I error is deemed to be equally costly as a Type II error, which for for a study with 1000 observations per condition would lead to a Type I error and a Type II error of 0.0179. As Faul, Erdfelder, Lang, and Buchner (2007) write:

> Of course, compromise power analyses can easily result in unconventional significance levels greater than α = .05 (in the case of small samples or effect sizes) or less than α = .001 (in the case of large samples or effect sizes). However, we believe that the benefit of balanced Type I and Type II error risks often offsets the costs of violating significance level conventions.

This brings us to the second situation where a compromise power analysis can be useful, which is when we know the statistical power in our study is low. Although it is highly undesirable to make decisions when error rates are high, if one finds oneself in a situation where a decision must be made based on little information, Winer (1962) writes:

> The frequent use of the .05 and .01 levels of significance is a matter of convention having little scientific or logical basis. When the power of tests is likely to be low under these levels of significance, and when Type I and Type II errors are of approximately equal importance, the .30 and .20 levels of significance may be more appropriate than the .05 and .01 levels.

For example, if we plan to perform a two-sided t test, can feasibly collect at most 50 observations in each independent group, and expect a population effect size of 0.5, we would have 70% power if we set our alpha level to 0.05. We can choose to weigh both types of error equally, and set the alpha level to 0.149, to end up with a statistical power for an effect of $d = 0.5$ of 0.851 (given a 0.149 Type II error rate). The choice of α and β in a compromise power analysis can be extended to take prior probabilities of the null and alternative hypothesis into account (Miller & Ulrich, 2019; Murphy et al., 2014).

A compromise power analysis requires a researcher to specify the sample size. This sample size itself requires a justification, so a compromise power analysis will typically be performed together with a resource constraint justification for a sample size. It is especially important to perform a compromise power analysis if your resource constraint justification is strongly based on the need to make a decision, in which case a researcher should think carefully about the Type I and Type II error rates stakeholders are willing to accept. However, a compromise power analysis also makes sense if the sample size is very large, but a researcher did not have the freedom to set the sample size. This might happen if, for example, data collection is part of a larger international study and the sample size is based on other research questions. In designs where the Type II error rates is very small (and power is very high) some statisticians have also recommended to lower the alpha level to prevent Lindley's paradox, a situation where a significant effect ($p < \alpha$) is evidence for the null hypothesis (Good, 1992; Jeffreys, 1939). Lowering the alpha level as a function of the statistical power of the test can prevent this paradox, providing another argument for a compromise power analysis when sample sizes are large. Finally, a compromise power analysis needs a justification for the effect size, either based on a smallest effect

size of interest or an effect size that is expected. Table 7 lists three aspects that should be discussed alongside a reported compromise power analysis.

What to do if Your Editor Asks for Post-hoc Power?

Post-hoc (or observed) power is the statistical power of the test, assuming the effect size estimated from the data is the true effect size. Post-hoc power is therefore not performed before looking at the data based on effect sizes that are deemed interesting, as an a-priori power analysis is, and it is unlike a sensitivity power analysis where a range of effect sizes is evaluated. Because post-hoc power analysis is based on the observed effect size, it does not add any information beyond the reported p value, but it presents the same information in a different way. Despite this fact, editors and reviewers often ask authors to perform post-hoc power analysis to interpret non-significant results. This is not a sensible request, and whenever it is made, you should not comply with it. Instead, you should perform a sensitivity power analysis, and discuss the power for the smallest effect size of interest and a realistic range of expected effect sizes.

Post-hoc power is directly related to the p value of the statistical test (Hoenig & Heisey, 2001). For a z test where the p value is exactly 0.05, post-hoc power is always 50%. The reason for this relationship is that when a p value is observed that equals the alpha level of the test (e.g., 0.05), the observed z score of the test is exactly equal to the critical value of the test (e.g., $z = 1.96$ in a two-sided test with a 5% alpha level). Whenever the alternative hypothesis is centered on the critical value half the values we expect to observe if this alternative hypothesis is true fall below the critical value, and half fall above the critical value. Therefore, a test where we observed a p value identical to the alpha level will have exactly 50% power in a post-hoc power analysis, as the analysis assumes the observed effect size is true.

For other statistical tests, where the alternative distribution is not symmetric (such as for the t test, where the alternative hypothesis follows a non-central t distribution, see Figure 8), a $p = 0.05$ does not directly translate to an observed power of 50%, but by plotting post-hoc power against the observed p value we see that the two statistics are always directly related. As Figure 12 shows, if the p value is non-significant (i.e., larger than 0.05) the observed power will be less than approximately 50% in a t test. Lenth (2007) explains how observed power is also completely determined by the observed p value for F tests, although the statement that a non-significant p value implies a power less than 50% no longer holds.

When editors or reviewers ask researchers to report post-hoc power analyses they would like to be able to distinguish between true negatives (concluding there is no effect, when there is no effect) and false negatives (a Type II error, concluding there is no effect, when there actually is an effect). Since reporting post-hoc power is just a different way of reporting the p value, reporting the post-hoc power will not provide an answer to the question editors are asking (Hoenig & Heisey, 2001; Lenth, 2007; Schulz & Grimes, 2005; Yuan & Maxwell, 2005). To be able to draw conclusions about the absence of a meaningful effect, one should perform an equivalence test, and design a study with high power to reject the smallest effect size of interest (Lakens et al., 2018). Alternatively, if no smallest effect size of interest was specified when designing the study, researchers can report a sensitivity power analysis.

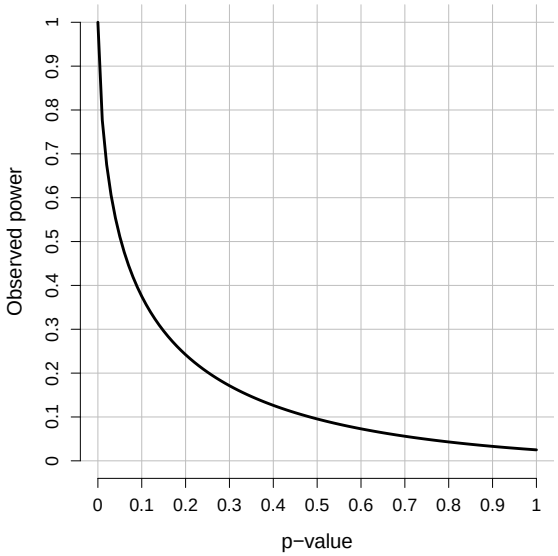

Figure 12. Relationship between p values and power for an independent t test with $\alpha = 0.05$ and n = 10.

Sequential Analyses

Whenever the sample size is justified based on an a-priori power analysis it can be very efficient to collect data in a sequential design. Sequential designs control error rates across multiple looks at the data (e.g., after 50, 100, and 150 observations have been collected) and can reduce the average expected sample size that is collected compared to a fixed design where data is only analyzed after the maximum sample size is collected (Proschan, Lan, & Wittes, 2006; Wassmer & Brannath, 2016). Sequential designs have a long history (Dodge & Romig, 1929), and exist in many variations,

Table 7
Overview of recommendations when justifying error rates based on a compromise power analysis.

What to take into account	How to take it into account?
What is the justification for the sample size?	Specify why a specific sample size is collected (e.g., based on resource constraints or other factors that determined the sample size).
What is the justification for the effect size?	Is the effect size based on a smallest effect size of interest or an expected effect size?
What is the desired ratio of Type I vs Type II error rates?	Weigh the relative costs of a Type I and a Type II error by carefully evaluating the consequences of each type of error.

such as the Sequential Probability Ratio Test (Wald, 1945), combining independent statistical tests (Westberg, 1985), group sequential designs (Jennison & Turnbull, 2000), sequential Bayes factors (Schönbrodt, Wagenmakers, Zehetleitner, & Perugini, 2017), and safe testing (Grünwald, de Heide, & Koolen, 2019). Of these approaches, the Sequential Probability Ratio Test is most efficient if data can be analyzed after every observation (Schnuerch & Erdfelder, 2020). Group sequential designs, where data is collected in batches, provide more flexibility in data collection, error control, and corrections for effect size estimates (Wassmer & Brannath, 2016). Safe tests provide optimal flexibility if there are dependencies between observations (ter Schure & Grünwald, 2019).

Sequential designs are especially useful when there is considerable uncertainty about the effect size, or when it is plausible that the true effect size is larger than the smallest effect size of interest the study is designed to detect (Lakens, 2014). In such situations data collection has the possibility to terminate early if the effect size is larger than the smallest effect size of interest, but data collection can continue to the maximum sample size if needed. Sequential designs can prevent waste when testing hypotheses, both by stopping early when the null hypothesis can be rejected, as by stopping early if the presence of a smallest effect size of interest can be rejected (i.e., stopping for futility). Group sequential designs are currently the most widely used approach to sequential analyses, and can be planned and analyzed using rpact (Wassmer & Pahlke, 2019) or gsDesign (Anderson, 2014).[6]

Increasing Power Without Increasing the Sample Size

The most straightforward approach to increase the informational value of studies is to increase the sample size. Because resources are often limited, it is also worthwhile to explore different approaches to increasing the power of a test without increasing the sample size. The first option is to use directional tests where relevant. Researchers often make directional predictions, such as 'we predict X is larger than Y'. The statistical test that logically follows from this prediction is a directional (or one-sided) t test. A directional test moves the Type I error rate to one side of the tail of the distribution, which lowers the critical value, and therefore requires less observations to achieve the same statistical power.

Although there is some discussion about when directional tests are appropriate, they are perfectly defensible from a Neyman-Pearson perspective on hypothesis testing (Cho & Abe, 2013), which makes a (preregistered) directional test a straightforward approach to both increase the power of a test, as the riskiness of the prediction. However, there might be situations where you do not want to ask a directional question. Sometimes, especially in research with applied consequences, it might be important to examine if a null effect can be rejected, even if the effect is in the opposite direction as predicted. For example, when you are evaluating a recently introduced educational intervention, and you predict the intervention will increase the performance of students, you might want to explore the possibility that students perform worse, to be able to recommend abandoning the new intervention. In such cases it is also possible to distribute the error rate in a 'lop-sided' manner, for example assigning a stricter error rate to effects in the negative than in the positive direction (Rice & Gaines, 1994).

Another approach to increase the power without increasing the sample size, is to increase the alpha level of the test, as explained in the section on compromise power analysis. Obviously, this comes at an increased probability of making a Type I error. The risk of making either type of error should be carefully weighed, which typically requires taking into account the prior

[6]Shiny apps are available for both rpact: https://rpact.shinyapps.io/public/ and gsDesign: https://gsdesign.shinyapps.io/prod/

probability that the null-hypothesis is true (Cascio & Zedeck, 1983; Miller & Ulrich, 2019; Mudge, Baker, Edge, & Houlahan, 2012; Murphy et al., 2014). If you *have* to make a decision, or want to make a claim, and the data you can feasibly collect is limited, increasing the alpha level is justified, either based on a compromise power analysis, or based on a cost-benefit analysis (Baguley, 2004; Field, Tyre, Jonzén, Rhodes, & Possingham, 2004).

Another widely recommended approach to increase the power of a study is use a within participant design where possible. In almost all cases where a researcher is interested in detecting a difference between groups, a within participant design will require collecting less participants than a between participant design. The reason for the decrease in the sample size is explained by the equation below from Maxwell, Delaney, and Kelley (2017). The number of participants needed in a two group within-participants design (NW) relative to the number of participants needed in a two group between-participants design (NB), assuming normal distributions, is:

$$NW = \frac{NB(1-\rho)}{2}$$

The required number of participants is divided by two because in a within-participants design with two conditions every participant provides two data points. The extent to which this reduces the sample size compared to a between-participants design also depends on the correlation between the dependent variables (e.g., the correlation between the measure collected in a control task and an experimental task), as indicated by the (1-ρ) part of the equation. If the correlation is 0, a within-participants design simply needs half as many participants as a between participant design (e.g., 64 instead 128 participants). The higher the correlation, the larger the relative benefit of within-participants designs, and whenever the correlation is negative (up to -1) the relative benefit disappears. Especially when dependent variables in within-participants designs are positively correlated, within-participants designs will greatly increase the power you can achieve given the sample size you have available. Use within-participants designs when possible, but weigh the benefits of higher power against the downsides of order effects or carryover effects that might be problematic in a within-participants design (Maxwell et al., 2017).[7] For designs with multiple factors with multiple levels it can be difficult to specify the full correlation matrix that specifies the expected population correlation for each pair of measurements (Lakens & Caldwell, 2019). In these cases sequential analyses might provide a solution.

In general, the smaller the variation, the larger the standardized effect size (because we are dividing the raw effect by a smaller standard deviation) and thus the higher the power given the same number of observations. Some additional recommendations are provided in the literature (Allison, Allison, Faith, Paultre, & Pi-Sunyer, 1997; Bausell & Li, 2002; Hallahan & Rosenthal, 1996), such as:

1. Use better ways to screen participants for studies where participants need to be screened before participation.
2. Assign participants unequally to conditions (if data in the control condition is much cheaper to collect than data in the experimental condition, for example).
3. Use reliable measures that have low error variance (Williams, Zimmerman, & Zumbo, 1995).
4. Smart use of preregistered covariates (Meyvis & Van Osselaer, 2018).

It is important to consider if these ways to reduce the variation in the data do not come at too large a cost for external validity. For example, in an *intention-to-treat* analysis in randomized controlled trials participants who do not comply with the protocol are maintained in the analysis such that the effect size from the study accurately represents the effect of implementing the intervention in the population, and not the effect of the intervention only on those people who perfectly follow the protocol (Gupta, 2011). Similar trade-offs between reducing the variance and external validity exist in other research areas.

Know Your Measure

Although it is convenient to talk about standardized effect sizes, it is generally preferable if researchers can interpret effects in the raw (unstandardized) scores, and have knowledge about the standard deviation of their measures (Baguley, 2009; Lenth, 2001). To make it possible for a research community to have realistic expectations about the standard deviation of measures they collect, it is beneficial if researchers within a research area use the same validated measures. This provides a reliable knowledge base that makes it easier to plan for a desired accuracy, and to use a smallest effect size of interest on the unstandardized scale in an a-priori power analysis.

In addition to knowledge about the standard deviation it is important to have knowledge about the correla-

[7] You can compare within- and between-participants designs in this Shiny app: http://shiny.ieis.tue.nl/within_between.

tions between dependent variables (for example because Cohen's d_z for a dependent t test relies on the correlation between means). The more complex the model, the more aspects of the data-generating process need to be known to make predictions. For example, in hierarchical models researchers need knowledge about variance components to be able to perform a power analysis (DeBruine & Barr, 2019; Westfall et al., 2014). Finally, it is important to know the reliability of your measure (Parsons, Kruijt, & Fox, 2019), especially when relying on an effect size from a published study that used a measure with different reliability, or when the same measure is used in different populations, in which case it is possible that measurement reliability differs between populations. With the increasing availability of open data, it will hopefully become easier to estimate these parameters using data from earlier studies.

If we calculate a standard deviation from a sample, this value is an estimate of the true value in the population. In small samples, our estimate can be quite far off, while due to the law of large numbers, as our sample size increases, we will be measuring the standard deviation more accurately. Since the sample standard deviation is an estimate with uncertainty, we can calculate a confidence interval around the estimate (Smithson, 2003), and design pilot studies that will yield a sufficiently reliable estimate of the standard deviation. The confidence interval for the variance σ^2 is provided in the following formula, and the confidence for the standard deviation is the square root of these limits:

$$(N-1)s^2/\chi^2_{N-1:\alpha/2}, (N-1)s^2/\chi^2_{N-1:1-\alpha/2}$$

Whenever there is uncertainty about parameters, researchers can use sequential designs to perform an *internal pilot study* (Wittes & Brittain, 1990). The idea behind an internal pilot study is that researchers specify a tentative sample size for the study, perform an interim analysis, use the data from the internal pilot study to update parameters such as the variance of the measure, and finally update the final sample size that will be collected. As long as interim looks at the data are blinded (e.g., information about the conditions is not taken into account) the sample size can be adjusted based on an updated estimate of the variance without any practical consequences for the Type I error rate (Friede & Kieser, 2006; Proschan, 2005). Therefore, if researchers are interested in designing an informative study where the Type I and Type II error rates are controlled, but they lack information about the standard deviation, an internal pilot study might be an attractive approach to consider (Chang, 2016).

Sample Size Justification in Qualitative Research

A value of information perspective to sample size justification also applies to qualitative research. A sample size justification in qualitative research should be based on the consideration that the cost of collecting data from additional participants does not yield new information that is valuable enough given the inferential goals. One widely used application of this idea is known as *saturation* and is indicated by the observation that new data replicates earlier observations, without adding new information (Morse, 1995). For example, let's imagine we ask people why they have a pet. Interviews might reveal reasons that are grouped into categories, but after interviewing 20 people, no new categories emerge, at which point saturation has been reached. Alternative philosophies to qualitative research exist, and not all value planning for saturation. Regrettably, principled approaches to justify sample sizes have not been developed for these alternative philosophies (Marshall, Cardon, Poddar, & Fontenot, 2013).

When sampling, the goal is often not to pick a representative sample, but a sample that contains a sufficiently diverse number of subjects such that saturation is reached efficiently. Fugard and Potts (2015) show how to move towards a more informed justification for the sample size in qualitative research based on 1) the number of codes that exist in the population (e.g., the number of reasons people have pets), 2) the probability a code can be observed in a single information source (e.g., the probability that someone you interview will mention each possible reason for having a pet), and 3) the number of times you want to observe each code. They provide an R formula based on binomial probabilities to compute a required sample size to reach a desired probability of observing codes.

A more advanced approach is used in Rijnsoever (2017), which also explores the importance of different sampling strategies. In general, purposefully sampling information from sources you expect will yield novel information is much more efficient than random sampling, but this also requires a good overview of the expected codes, and the sub-populations in which each code can be observed. Sometimes, it is possible to identify information sources that, when interviewed, would at least yield one new code (e.g., based on informal communication before an interview). A good sample size justification in qualitative research is based on 1) an identification of the populations, including any subpopulations, 2) an estimate of the number of codes in the (sub-)population, 3) the probability a code is encountered in an information source, and 4) the sampling strategy that is used.

Discussion

Providing a coherent sample size justification is an essential step in designing an informative study. There are multiple approaches to justifying the sample size in a study, depending on the goal of the data collection, the resources that are available, and the statistical approach that is used to analyze the data. An overarching principle in all these approaches is that researchers consider the value of the information they collect in relation to their inferential goals.

The process of justifying a sample size when designing a study should sometimes lead to the conclusion that it is not worthwhile to collect the data, because the study does not have sufficient informational value to justify the costs. There will be cases where it is unlikely there will ever be enough data to perform a meta-analysis (for example because of a lack of general interest in the topic), the information will not be used to make a decision or claim, and the statistical tests do not allow you to test a hypothesis with reasonable error rates or to estimate an effect size with sufficient accuracy. If there is no good justification to collect the maximum number of observations that one can feasibly collect, performing the study anyway is a waste of time and/or money (Brown, 1983; Button et al., 2013; Halpern et al., 2002).

The awareness that sample sizes in past studies were often too small to meet any realistic inferential goals is growing among psychologists (Button et al., 2013; Fraley & Vazire, 2014; Lindsay, 2015; Sedlmeier & Gigerenzer, 1989). As an increasing number of journals start to require sample size justifications, some researchers will realize they need to collect larger samples than they were used to. This means researchers will need to request more money for participant payment in grant proposals, or that researchers will need to increasingly collaborate (Moshontz et al., 2018). If you believe your research question is important enough to be answered, but you are not able to answer the question with your current resources, one approach to consider is to organize a research collaboration with peers, and pursue an answer to this question collectively.

A sample size justification should not be seen as a hurdle that researchers need to pass before they can submit a grant, ethical review board proposal, or manuscript for publication. When a sample size is simply stated, instead of carefully justified, it can be difficult to evaluate whether the value of the information a researcher aims to collect outweighs the costs of data collection. Being able to report a solid sample size justification means a researchers knows what they want to learn from a study, and makes it possible to design a study that can provide an informative answer to a scientific question.

References

Aberson, C. L. (2019). *Applied Power Analysis for the Behavioral Sciences* (Second). New York: Routledge.

Albers, C. J., Kiers, H. A. L., & Ravenzwaaij, D. van. (2018). Credible Confidence: A Pragmatic View on the Frequentist vs Bayesian Debate. *Collabra: Psychology*, *4*(1), 31. https://doi.org/10.1525/collabra.149

Albers, C. J., & 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. https://doi.org/10.1016/j.jesp.2017.09.004

Allison, D. B., Allison, R. L., Faith, M. S., Paultre, F., & Pi-Sunyer, F. X. (1997). Power and money: Designing statistically powerful studies while minimizing financial costs. *Psychological Methods*, *2*(1), 20–33. https://doi.org/10.1037/1082-989X.2.1.20

Anderson, K. M. (2014). Group Sequential Design in R. In *Clinical Trial Biostatistics and Biopharmaceutical Applications* (pp. 179–209). New York: CRC Press.

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. https://doi.org/10.1177/0956797617723724

Bacchetti, P. (2010). Current sample size conventions: Flaws, harms, and alternatives. *BMC Medicine*, *8*(1), 17. https://doi.org/10.1186/1741-7015-8-17

Baguley, T. (2004). Understanding statistical power in the context of applied research. *Applied Ergonomics*, *35*(2), 73–80. https://doi.org/10.1016/j.apergo.2004.01.002

Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology*, *100*(3), 603–617. https://doi.org/10.1348/000712608X377117

Bausell, R. B., & Li, Y.-F. (2002). *Power analysis for experimental research: A practical guide for the biological, medical and social sciences*. Cambridge University Press.

Berkeley, G. (1735). *A defence of free-thinking in mathematics, in answer to a pamphlet of Philalethes Cantabrigiensis entitled Geometry No Friend to Infidelity. Also an appendix concerning mr. Walton's Vindication of the principles of fluxions against the objections contained in The analyst. By the author of The minute philosopher* (Vol. 3).

Bosco, F. A., Aguinis, H., Singh, K., Field, J. G., & Pierce, C. A. (2015). Correlational effect size benchmarks. *The Journal of Applied Psychology*, *100*(2), 431–449. https://doi.org/10.1037/a0038047

Brown, G. W. (1983). Errors, Types I and II. *American Journal of Diseases of Children*, *137*(6), 586–591. https://doi.org/10.1001/archpedi.1983.02140320062014

Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *Journal of Cognition*, *2*(1), 16. https://doi.org/10.5334/joc.72

Brysbaert, M., & Stevens, M. (2018). Power Analysis and Effect Size in Mixed Effects Models: A Tutorial. *Journal of Cognition*, *1*(1). https://doi.org/10.5334/joc.10

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for Bias in Psychology: A Comparison of Meta-Analytic Methods. *Advances in Methods and Practices in Psychological Science*, *2*(2), 115–144. https://doi.org/10.1177/2515245919847196

Cascio, W. F., & Zedeck, S. (1983). Open a New Window in Rational Research Planning: Adjust Alpha to Maximize Statistical Power. *Personnel Psychology*, *36*(3), 517–526. https://doi.org/10.1111/j.1744-6570.1983.tb02233.x

Chang, M. (2016). *Adaptive Design Theory and Implementation Using SAS and R* (2nd edition). Chapman and Hall/CRC.

Cho, H.-C., & Abe, S. (2013). Is two-tailed testing for directional research hypotheses tests legitimate? *Journal of Business Research*, *66*(9), 1261–1266. https://doi.org/10.1016/j.jbusres.2012.02.023

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

Cook, J., Hislop, J., Adewuyi, T., Harrild, K., Altman, D., Ramsay, C., … Vale, L. (2014). Assessing methods to specify the target difference for a randomised controlled trial: DELTA (Difference ELicitation in TriAls) review. *Health Technology Assessment*, *18*(28). https://doi.org/10.3310/hta18280

Copay, A. G., Subach, B. R., Glassman, S. D., Polly, D. W., & Schuler, T. C. (2007). Understanding the minimum clinically important difference: A review of concepts and methods. *The Spine Journal*, *7*(5), 541–546. https://doi.org/10.1016/j.spinee.2007.01.008

Correll, J., Mellinger, C., McClelland, G. H., & Judd, C. M. (2020). Avoid Cohen's "Small", "Medium", and "Large" for Power Analysis. *Trends in Cognitive Sciences*, *24*(3), 200–207. https://doi.org/10.1016/j.tics.2019.12.009

Cousineau, D., & Chiasson, F. (2019). *Superb: Computes standard error and confidence interval of means under various designs and sampling schemes* [Manual].

Cumming, G. (2013). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge.

Cumming, G. (2014). The New Statistics: Why and How. *Psychological Science*, *25*(1), 7–29. https://doi.org/10.1177/0956797613504966

Cumming, G., & Calin-Jageman, R. (2016). *Introduction to the New Statistics: Estimation, Open Science, and Beyond*. Routledge.

DeBruine, L. M., & Barr, D. J. (2019). *Understanding mixed effects models through data simulation* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/xp5cy

Detsky, A. S. (1990). Using cost-effectiveness analysis to improve the efficiency of allocating funds to clinical trials. *Statistics in Medicine*, *9*(1-2), 173–184. https://doi.org/10.1002/sim.4780090124

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00781

Dodge, H. F., & Romig, H. G. (1929). A Method of Sampling Inspection. *Bell System Technical Journal*, *8*(4), 613–631. https://doi.org/10.1002/j.1538-7305.1929.tb01240.x

Eckermann, S., Karnon, J., & Willan, A. R. (2010). The Value of Value of Information. *PharmacoEconomics*, *28*(9), 699–709. https://doi.org/10.2165/11537370-000000000-00000

Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, & Computers*, *28*(1), 1–11. https://doi.org/10.3758/BF03203630

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

Ferron, J., & Onghena, P. (1996). The Power of Randomization Tests for Single-Case Phase Designs. *The Journal of Experimental Education*, *64*(3), 231–239. https://doi.org/10.1080/00220973.1996.9943805

Field, S. A., Tyre, A. J., Jonzén, N., Rhodes, J. R., & Possingham, H. P. (2004). Minimizing the cost of environmental management decisions by optimizing statistical thresholds. *Ecology Letters*, *7*(8), 669–675. https://doi.org/10.1111/j.1461-0248.2004.00625.x

Fraley, R. C., & Vazire, S. (2014). The N-Pact Factor: Evaluating the Quality of Empirical Journals with Respect to Sample Size and Statistical Power. *PLOS ONE*, *9*(10), e109019. https://doi.org/10.1371/journal.pone.0109019

Friede, T., & Kieser, M. (2006). Sample size recalculation in internal pilot study designs: A review. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, *48*(4), 537–555. https://doi.org/10.1002/bimj.200510238

Fugard, A. J. B., & Potts, H. W. W. (2015). Supporting thinking on sample sizes for thematic analyses: A quantitative tool. *International Journal of Social Research Methodology*, *18*(6), 669–684. https://doi.org/10.1080/13645579.2015.1005453

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, *2*(2), 156–168. https://doi.org/10.1177/2515245919847202

Good, I. J. (1992). The Bayes/Non-Bayes Compromise: A Brief Review. *Journal of the American Statistical Association*, *87*(419), 597–606. https://doi.org/10.2307/2290192

Green, P., & MacLeod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493–498. https://doi.org/10.1111/2041-210X.12504

Green, S. B. (1991). How Many Subjects Does It Take To Do A Regression Analysis. *Multivariate Behavioral Research*, *26*(3), 499–510. https://doi.org/10.1207/s15327906mbr2603_7

Grünwald, P., de Heide, R., & Koolen, W. (2019). Safe Testing. *arXiv:1906.07801 [Cs, Math, Stat]*. Retrieved from http://arxiv.org/abs/1906.07801

Gupta, S. K. (2011). Intention-to-treat concept: A review. *Perspectives in Clinical Research*, *2*(3), 109–112. https://doi.org/10.4103/2229-3485.83221

Hallahan, M., & Rosenthal, R. (1996). Statistical power: Concepts, procedures, and applications. *Behaviour Research and Therapy*, *34*(5), 489–499. https://doi.org/10.1016/0005-7967(95)00082-8

Halpern, J., Brown Jr, B. W., & Hornberger, J. (2001). The sample size for a clinical trial: A Bayesian decision theoretic approach. *Statistics in Medicine*, *20*(6), 841–858. https://doi.org/10.1002/sim.703

Halpern, S. D., Karlawish, J. H., & Berlin, J. A. (2002). The continuing unethical conduct of underpowered clinical trials. *Jama*, *288*(3), 358–362. https://doi.org/doi:10.1001/jama.288.3.358

Hedges, L. V., & Pigott, T. D. (2001). The power of statistical tests in meta-analysis. *Psychological Methods*, *6*(3), 203–217. https://doi.org/10.1037/1082-989X.6.3.203

Hilgard, J. (2021). Maximal positive controls: A method for estimating the largest plausible effect size. *Journal of Experimental Social Psychology*, *93*. https://doi.org/10.1016/j.jesp.2020.104082

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical Benchmarks for Interpreting Effect Sizes in Research. *Child Development Perspectives*, *2*(3), 172–177. https://doi.org/10.1111/j.1750-8606.2008.00061.x

Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *The American Statistician*, *55*(1), 19–24. https://doi.org/10.1198/000313001300339897

Jaeschke, R., Singer, J., & Guyatt, G. H. (1989). Measurement of health status: Ascertaining the minimal clinically important difference. *Controlled Clinical Trials*, *10*(4), 407–415. https://doi.org/10.1016/0197-2456(89)90005-6

Jeffreys, H. (1939). *Theory of probability* (1st ed). Oxford [Oxfordshire]: New York: Oxford University Press.

Jennison, C., & Turnbull, B. W. (2000). *Group sequential methods with applications to clinical trials*. Boca Raton: Chapman & Hall/CRC.

Julious, S. A. (2004). Sample sizes for clinical trials with normal data. *Statistics in Medicine*, *23*(12), 1921–1986. https://doi.org/10.1002/sim.1783

Keefe, R. S. E., Kraemer, H. C., Epstein, R. S., Frank, E., Haynes, G., Laughren, T. P., … Leon, A. C. (2013). Defining a Clinically Meaningful Effect for the Design and Interpretation of Randomized Controlled Trials. *Innovations in Clinical Neuroscience*, *10*(5-6 Suppl A), 4S–19S.

Kelley, K. (2007). Confidence Intervals for Standardized Effect Sizes: Theory, Application, and Implementation. *Journal of Statistical Software*, *20*(8). https://doi.org/10.18637/JSS.V020.I08

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods*, *17*(2), 137–152. https://doi.org/10.1037/a0028086

Kelley, K., & Rausch, J. R. (2006). Sample size planning for the standardized mean difference: Accuracy in parameter estimation via narrow confidence intervals. *Psychological Methods*, *11*(4), 363–385. https://doi.org/10.1037

Kenny, D. A., & Judd, C. M. (2019). The unappreciated heterogeneity of effect sizes: Implications for power, precision, planning of research, and replication. *Psychological Methods*, *24*(5), 578–589. https://doi.org/10.1037/met0000209

King, M. T. (2011). A point of minimal important difference (MID): A critique of terminology and methods. *Expert Review of Pharmacoeconomics & Outcomes Research*, *11*(2), 171–184. https://doi.org/10.1586/erp.11.9

Kish, L. (1965). *Survey Sampling*. New York: Wiley.

Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, *49*(4), 241–253. https://doi.org/10.3102/0013189X20912798

Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*, *6*(3), 299–312.

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, *142*(2), 573–603. https://doi.org/10.1037/a0029146

Kruschke, J. K. (2018). Rejecting or Accepting Parameter Values in Bayesian Estimation. *Advances in Methods and Practices in Psychological Science*, *1*(2), 270–280. https://doi.org/10.1177/2515245918771304

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

Lakens, D. (2017). Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses. *Social Psychological and Personality Science*, *8*(4), 355–362. https://doi.org/10.1177/1948550617697177

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., … Zwaan, R. A. (2018). Justify your alpha. *Nature Human Behaviour*, *2*, 168–171. https://doi.org/10.1038/s41562-018-0311-x

Lakens, D., & Caldwell, A. R. (2019). *Simulation-Based Power-Analysis for Factorial ANOVA Designs*. PsyArXiv. https://doi.org/10.31234/osf.io/baxsf

Lakens, D., & Etz, A. J. (2017). Too True to be Bad: When Sets of Studies With Significant and Nonsignificant Findings Are Probably True. *Social Psychological and Personality Science*, *8*(8), 875–881. https://doi.org/10.1177/1948550617693058

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259–269. https://doi.org/10.1177/2515245918770963

Lenth, R. V. (2001). Some practical guidelines for effective sample size determination. *The American Statistician*, *55*(3), 187–193. https://doi.org/10.1198/000313001317098149

Lenth, R. V. (2007). Post hoc power: Tables and commentary. *Iowa City: Department of Statistics and Actuarial Science, University of Iowa*.

Leon, A. C., Davis, L. L., & Kraemer, H. C. (2011). The Role and Interpretation of Pilot Studies in Clinical Research. *Journal of Psychiatric Research*, *45*(5), 626–629. https://doi.org/10.1016/j.jpsychires.2010.10.008

Lindsay, D. S. (2015). Replication in Psychological Science. *Psychological Science*, *26*(12), 1827–1832. https://doi.org/10.1177/0956797615616374

Lovakov, A., & Agadullina, E. (2017). Empirically Derived Guidelines for Interpreting Effect Size in Social Psychology. *PsyArXiv*. https://doi.org/10.17605/OSF.IO/2EPC4

Marshall, B., Cardon, P., Poddar, A., & Fontenot, R. (2013). Does Sample Size Matter in Qualitative Research?: A Review of Qualitative Interviews in is Research. *Journal of Computer Information Systems*, *54*(1), 11–22. https://doi.org/10.1080/08874417.2013.11645667

Maxwell, S. E., Delaney, H. D., & Kelley, K. (2017). *Designing Experiments and Analyzing Data: A Model Comparison Perspective, Third Edition* (3 edition). New York, NY: Routledge.

Maxwell, S. E., & Kelley, K. (2011). Ethics and sample size planning. In *Handbook of ethics in quantitative methodology* (pp. 179–204). Routledge.

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*(1), 537–563. https://doi.org/10.1146/annurev.psych.59.103006.093735

McIntosh, R. D., & Rittmo, J. Ö. (2020). *Power calculations in single case neuropsychology*. https://doi.org/10.31234/osf.io/fxz49

Meyners, M. (2012). Equivalence tests A review. *Food Quality and Preference*, *26*(2), 231–245. https://doi.org/10.1016/j.foodqual.2012.05.003

Meyvis, T., & Van Osselaer, S. M. J. (2018). Increasing the Power of Your Study by Increasing the Effect Size. *Journal of Consumer Research*, *44*(5), 1157–1173. https://doi.org/10.1093/jcr/ucx110

Miller, J., & Ulrich, R. (2019). The quest for an optimal alpha. *PLOS ONE*, *14*(1), e0208631. https://doi.org/10.1371/journal.pone.0208631

Morey, R. D. (2020). *Power and precision* [Blog]. https://medium.com/@richarddmorey/power-and-precision-47f644ddea5e.

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *38*(11), 2074–2102. https://doi.org/10.1002/sim.8086

Morse, J. M. (1995). The Significance of Saturation. *Qualitative Health Research*, *5*(2), 147–149. https://doi.org/10.1177/104973239500500201

Moshontz, H., Campbell, L., Ebersole, C. R., IJzerman, H., Urry, H. L., Forscher, P. S., … Antfolk, J. (2018). The Psychological Science Accelerator: Advancing psychology through a distributed collaborative network. *Advances in Methods and Practices in Psychological Science*, *1*(4), 501–515. https://doi.org/10.1177/2515245918797607

Mudge, J. F., Baker, L. F., Edge, C. B., & Houlahan, J. E. (2012). Setting an Optimal α That Minimizes Errors in Null Hypothesis Significance Tests. *PLOS ONE*, *7*(2), e32734. https://doi.org/10.1371/journal.pone.0032734

Murphy, K. R., Myors, B., & Wolach, A. H. (2014). *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests* (Fourth edition). New York: Routledge, Taylor & Francis Group.

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, *231*(694-706), 289–337. https://doi.org/10.1098/rsta.1933.0009

Olsson-Collentine, A., Wicherts, J. M., & van Assen, M. A. L. M. (2020). Heterogeneity in direct replications in psychology and its association with effect size. *Psychological Bulletin*, *146*(10), 922–940. https://doi.org/10.1037/bul0000294

Parker, R. A., & Berman, N. G. (2003). Sample Size. *The American Statistician*, *57*(3), 166–170. https://doi.org/10.1198/0003130031919

Parsons, S., Kruijt, A.-W., & Fox, E. (2019). Psychological Science Needs a Standard Practice of Reporting the Reliability of Cognitive-Behavioral Measurements. *Advances in Methods and Practices in Psychological Science*, *2*(4), 378–395. https://doi.org/10.1177/2515245919879695

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

Perugini, M., Gallucci, M., & Costantini, G. (2018). A Practical Primer To Power Analysis for Simple Experimental Designs. *International Review of Social Psychology*, *31*(1), 20. https://doi.org/10.5334/irsp.181

Phillips, B. M., Hunt, J. W., Anderson, B. S., Puckett, H. M., Fairey, R., Wilson, C. J., & Tjeerdema, R. (2001). Statistical significance of sediment toxicity test results: Threshold values derived by the detectable significance approach. *Environmental Toxicology and Chemistry*, *20*(2), 371–373. https://doi.org/10.1002/etc.5620200218

Proschan, M. A. (2005). Two-Stage Sample Size Re-Estimation Based on a Nuisance Parameter: A Review. *Journal of Biopharmaceutical Statistics*, *15*(4), 559–574. https://doi.org/10.1081/BIP-200062852

Proschan, M. A., Lan, K. K. G., & Wittes, J. T. (2006). *Statistical monitoring of clinical trials: A unified approach*. New York, NY: Springer.

Rice, W. R., & Gaines, S. D. (1994). 'Heads I win, tails you lose': Testing directional alternative hypotheses in ecological and evolutionary research. *Trends in Ecology & Evolution*, *9*(6), 235–237. https://doi.org/10.1016/0169-5347(94)90258-5

Richard, F. D., Bond, C. F., & Stokes-Zoota, J. J. (2003). One Hundred Years of Social Psychology Quantitatively Described. *Review of General Psychology*, *7*(4), 331–363. https://doi.org/10.1037/1089-2680.7.4.331

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

Rijnsoever, F. J. van. (2017). (I Can't Get No) Saturation: A simulation and guidelines for sample sizes in qualitative research. *PLOS ONE*, *12*(7), e0181689. https://doi.org/10.1371/journal.pone.0181689

Rogers, J. L., Howard, K. I., & Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin*, *113*(3), 553–565. https://doi.org/http://dx.doi.org/10.1037/0033-2909.113.3.553

Scheel, A. M., Tiokhin, L., Isager, P. M., & Lakens, D. (2020). Why Hypothesis Testers Should Spend Less Time Testing Hypotheses. *Perspectives on Psychological Science*, 1745691620966795. https://doi.org/10.1177/1745691620966795

Schnuerch, M., & Erdfelder, E. (2020). Controlling decision errors with minimal costs: The sequential probability ratio t test. *Psychological Methods*, *25*(2), 206–226. https://doi.org/10.1037/met0000234

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. 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. https://doi.org/10.1016/j.jrp.2013.05.009

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods*, *22*(2), 322–339. https://doi.org/10.1037/MET0000061

Schulz, K. F., & Grimes, D. A. (2005). Sample size calculations in randomised trials: Mandatory and mystical. *The Lancet*, *365*(9467), 1348–1353. https://doi.org/10.1016/S0140-6736(05)61034-3

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, *105*(2), 309–316. https://doi.org/10.1037/0033-2909.105.2.309

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2013). *Life after P-Hacking*. New Orleans, LA.

Simonsohn, U. (2015). Small Telescopes Detectability and the Evaluation of Replication Results. *Psychological Science*, *26*(5), 559–569. https://doi.org/10.1177/0956797614567341

Smithson, M. (2003). *Confidence intervals*. Thousand Oaks, Calif: Sage Publications.

Spiegelhalter, D. (2019). *The Art of Statistics: How to Learn from Data* (Illustrated edition). New York: Basic Books.

Taylor, D. J., & Muller, K. E. (1996). Bias in linear model power and sample size calculation due to estimating noncentrality. *Communications in Statistics-Theory and Methods*, *25*(7), 1595–1610. https://doi.org/10.1080/03610929608831787

Teare, M. D., Dimairo, M., Shephard, N., Hayman, A., Whitehead, A., & Walters, S. J. (2014). Sample size requirements to estimate key design parameters from external pilot randomised controlled trials: A simulation study. *Trials*, *15*(1), 264. https://doi.org/10.1186/1745-6215-15-264

ter Schure, J., & Grünwald, P. D. (2019). Accumulation Bias in Meta-Analysis: The Need to Consider Time in Error Control. *arXiv:1905.13494 [Math, Stat]*. Retrieved from http://arxiv.org/abs/1905.13494

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327–352. https://doi.org/10.1037/0033-295X.84.4.327

Valentine, J. C., Pigott, T. D., & Rothstein, H. R. (2010). How Many Studies Do You Need?: A Primer on Statistical Power for Meta-Analysis. *Journal of Educational and Behavioral Statistics*, *35*(2), 215–247. https://doi.org/10.3102/1076998609346961

Viechtbauer, W., Smits, L., Kotz, D., Budé, L., Spigt, M., Serroyen, J., & Crutzen, R. (2015). A simple formula for the calculation of sample size in pilot studies. *Journal of Clinical Epidemiology*, *68*(11), 1375–1379. https://doi.org/10.1016/j.jclinepi.2015.04.014

Wald, A. (1945). Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, *16*(2), 117–186. https://doi.org/https://www.jstor.org/stable/2240273

Wassmer, G., & Brannath, W. (2016). *Group Sequential and Confirmatory Adaptive Designs in Clinical Trials*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-32562-0

Wassmer, G., & Pahlke, F. (2019). *Rpact: Confirmatory adaptive clinical trial design and analysis*.

Westberg, M. (1985). Combining Independent Statistical Tests. *Journal of the Royal Statistical Society. Series D (the Statistician)*, *34*(3), 287–296. https://doi.org/10.2307/2987655

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General*, *143*(5), 2020–2045. https://doi.org/10.1037/xge0000014

Williams, R. H., Zimmerman, D. W., & Zumbo, B. D. (1995). Impact of Measurement Error on Statistical Power: Review of an Old Paradox. *The Journal of Experimental Education*, *63*(4), 363–370. https://doi.org/10.1080/00220973.1995.9943470

Wilson, E. C. F. (2015). A Practical Guide to Value of Information Analysis. *PharmacoEconomics*, *33*(2), 105–121. https://doi.org/10.1007/s40273-014-0219-x

Wilson VanVoorhis, C. R., & Morgan, B. L. (2007). Understanding power and rules of thumb for determining sample sizes. *Tutorials in Quantitative Methods for Psychology*, *3*(2), 43–50. https://doi.org/10.20982/tqmp.03.2.p043

Winer, B. J. (1962). *Statistical principles in experimental design*. New York : McGraw-Hill.

Wittes, J., & Brittain, E. (1990). The role of internal pilot studies in increasing the efficiency of clinical trials. *Statistics in Medicine*, *9*(1-2), 65–72. https://doi.org/10.1002/sim.4780090113

Yuan, K.-H., & Maxwell, S. (2005). On the Post Hoc Power in Testing Mean Differences. *Journal of Educational and Behavioral Statistics*, *30*(2), 141–167. https://doi.org/10.3102/10769986030002141