Meta-Psychology, 2022, vol 6, MP.2021.3078
https://doi.org/10.15626/MP.2021.3078
Article type: Tutorial

Open data: Not Applicable
Open materials: Not Applicable
Open and reproducible analysis: Not Applicable
Open reviews and editorial process: Yes
Preregistration: No

Edited by: Erin M Buchanan
Reviewed by: Nataly Beribisky, Daniel Baker
Analysis reproduced by: Not Applicable
All supplementary files can be accessed at OSF:
https://doi.org/10.17605/OSF.IO/J5V26

Power to the People: A Beginner's Tutorial to Power Analysis using jamovi

James E Bartlett
School of Psychology and Neuroscience, University of Glasgow, UK

Sarah J Charles
Department of Psychology, Institute of Psychiatry, Psychology & Neuroscience, King's College London, UK

Abstract

Authors have highlighted for decades that sample size justification through power analysis is the exception rather than the rule. Even when authors do report a power analysis, there is often no justification for the smallest effect size of interest, or they do not provide enough information for the analysis to be reproducible. We argue one potential reason for these omissions is the lack of a truly accessible introduction to the key concepts and decisions behind power analysis. In this tutorial targeted at complete beginners, we demonstrate *a priori* and sensitivity power analysis using jamovi for two independent samples and two dependent samples. Respectively, these power analyses allow you to ask the questions: "How many participants do I need to detect a given effect size?", and "What effect sizes can I detect with a given sample size?". We emphasise how power analysis is most effective as a reflective process during the planning phase of research to balance your inferential goals with your resources. By the end of the tutorial, you will be able to understand the fundamental concepts behind power analysis and extend them to more advanced statistical models.

Keywords: Power analysis, effect size, tutorial, a priori, sensitivity, jamovi

Introduction

"If he is a typical psychological researcher, not only has he exerted no prior control over his risk of committing a type II error, but he will have no idea what the magnitude of this risk is." (Cohen, 1965, pg. 96)[1]

For decades researchers have highlighted that empirical research has chronically low statistical power (Button et al., 2013; Cohen, 1962; Sedlmeier & Gigerenzer 1989). This means that the study did not include enough participants to reliably detect a realistic effect size (see Table 1 for a definition of key terms). One method to avoid low statistical power is to calculate how many participants you need for a given effect size in a process called "power analysis". Power analysis is not the only way to justify your sample size (see Lakens, 2022), but despite increased attention to statistical power, it is still rare to find articles that justified their sample size through power analysis (Chen & Liu, 2019; Guo et al., 2014; Larson & Carbine, 2017). Even for those that do report a power analysis, there are often other problems such as poor justification for the effect size, misunderstanding statistical power, or not making the power analysis reproducible (Bakker et al., 2020; Beribisky et al., 2019; Collins & Watt, 2021). Therefore, we present a beginner's tutorial which outlines the

[1] We are aware of the problem with using gendered language like in the original quote. Despite this issue, we think the quote still resonates.

key decisions behind power analysis and walk through how it applies to *t*-tests for two independent samples and two dependent samples. We expect no background knowledge as we will explain the key concepts and how to interact with the software we use.

Before beginning the tutorial, it is important to explain why we need power analysis. There is a negative relationship between the sample size of a study and the effect size the study can reliably detect. Holding everything else constant, a larger study can detect smaller effect sizes, and conversely, a smaller study can only detect larger effect sizes. A study can be described as underpowered if the effect size you are trying to detect is smaller than the effect size your study has the ability to detect.

If we published or shared the results of all the studies we ever conducted, underpowered research would be less of a problem. We would just see more statistically non-significant findings. However, since there is publication bias that favours significant findings (Dwan et al., 2008; Franco et al., 2014), underpowered studies warp whole fields of research (Button et al., 2013). Imagine five research groups were interested in the same topic and performed a similar study using 50 participants each. The results from the first four groups were not statistically significant, but the fifth group by chance observed a larger effect size which was statistically significant. We know that non-significant findings are less likely to be published (Franco et al., 2014), so only the fifth group published their findings which happened to observe a larger statistically significant effect.

Now imagine you wanted to build on this research and to inform your study, you review the literature. All you find is the fifth study reporting a larger statistically significant effect, and you use that effect size to inform your study, meaning you recruit a smaller sample than if you expected a smaller effect size. If studies systematically use small sample sizes, only larger more unrealistic effect sizes are published, and smaller more realistic effect sizes are hidden. Moreover, researchers tend to have poor intuitions about statistical power, where they underestimate what sample size they need for a given effect size (Bakker et al., 2016). In combination, this means researchers tend to power their studies for unrealistically large effect sizes and think they need a sample size which would be too small to detect more realistic effect sizes (Etz & Vandekerckhove, 2016). In short, systematically underpowered research is a problem as it warps researchers' understanding of both what constitutes a realistic effect size and what an appropriate sample size is.

A power analysis tutorial article is nothing new. There are comprehensive guides to power analysis (e.g., Brysbaert, 2019; Perugini et al., 2018), but from our perspective, previous tutorials move too quickly from beginner to advanced concepts. In research methods curricula, educators only briefly cover power analysis across introductory and advanced courses (Sestir et al., 2021; TARG Meta-Research Group, 2020). In their assessment of researcher's understanding of power analysis, Collins and Watt (2021) advise that clearer educational materials should be available. In response, our approach is presenting a beginner's tutorial that can support both students and established researchers who are unfamiliar with power analysis.

We have split our tutorial into three parts starting with a recap of the statistical concepts underlying power analysis. There are common misconceptions around useful types of power analysis (Beribisky et al., 2019), so it is important to outline what we are trying to achieve. In part two, we outline the decisions you must make when performing a power analysis, like choosing your alpha, beta, and smallest effect size of interest. We then present a walk-through in part three on performing *a priori* and sensitivity power analyses for two independent samples and two dependent samples. Absolute beginners unfamiliar with power analysis should start with part one, while readers with a general understanding of power analysis can start with part two. We conclude with recommendations for future reading that outlines power analysis for more advanced statistical tests.

Part One: The Statistics Behind Power Analysis

Type I and Type II Errors

The dominant theory of statistics in psychology is known as "frequentist" or "classical" statistics. Power analysis is used within this framework where probability is assigned to "long-run" frequencies of observations (many things happening over time). In contrast, Bayesian statistics uses another theory of probability that can be applied to individual events through combining prior belief with a likelihood function[2]. In this article, we are only covering the frequentist approach where the "long-run" probability is the basis of where you get *p*-values from.

Researchers often misinterpret the information provided by *p*-values (Goodman, 2008). In our following explanations, we focus on the Neyman-Pearson approach (Lakens, 2021), where the aim of the frequentist branch of statistics is to help you make decisions and limit the number of errors you will make in the long-run (Neyman, 1977). The formal definition of a *p*-value

[2]See Kruschke and Liddell, (2018) for how power analysis applies to Bayesian statistics.

by Cohen (1994) is the probability of observing a result *at least as extreme* as the one observed, assuming the null hypothesis (there is no effect) is true. This means a small *p*-value (closer to 0) indicates the results are unlikely if the null hypothesis is true, while a large *p*-value (closer to 1) indicates the results are more likely if the null is true.

The probabilities do not relate to individual studies but tell you the probability attached to the procedure if you repeated it many times. So, a *p*-value of .05 means that, if you were to keep taking lots of samples from the population, and the null hypothesis was true, the chance of finding a result at least as extreme as the one we have seen is 5%, or 1 in 20. So, this can be phrased as "if we conducted an infinite number of studies with the same design as this, 5% of all of the results would be at least this extreme". The reason to use long-run probability, is that any single measurement comes with it the possibility of some kind of 'error', or unaccounted for variance that are not assumed in our hypotheses. For example, the researcher's reaction times, the temperature of the room, the participant's level of sleep, the brightness of the screens on equipment being used, etc. could all cause slight changes to the accuracy of the measures we make. In the long-run, these are likely averaged out.

We used a *p*-value of .05 as an example because an alpha value of .05 tends to be used as the cut-off point in psychology to conclude "we are happy to say that this result is unlikely/surprising enough to make a note of". Alpha (sometimes written as "α") is the probability of concluding there is an effect when there is not one, known as a type I error (said, type one error) or false positive. This is normally set at .05 (5%) and it is the threshold we look at for a significant effect. Setting alpha to .05 means we are willing to make a type I error 5% of the time in the long-run. In the Neyman-Pearson approach, we create cutoffs to help us make decisions (Lakens, 2021). We want to know if we can reject the null hypothesis and conclude we have observed some kind of effect. By setting alpha, we are saying the *p*-value for this study must be smaller than alpha to reject the null hypothesis. If our *p*-value is larger than alpha, we cannot reject the null hypothesis. This is where the term "statistical significance" comes from. As a scientific community, we have come to the group conclusion that this cut-off point is enough to say "the null hypothesis may not be true" and we understand that in the long-run, we would be willing to make a mistake 5% of the time if the null was really true.

It is important to understand that the cut-off of 5% appears immutable now for disciplines like psychology that routinely use 5% for alpha, but it was never meant as a fixed standard of evidence. Fisher - one of the pioneers of hypothesis testing - commented that he accepted 5% as a low standard of evidence across repeated findings (Goodman, 2008). Fisher (1926) emphasised that individual researchers should consider which alpha is appropriate for the standard of evidence in their study, but this nuance has been lost over time. For example, Bakker et al. (2020) reported that for studies that specifically mention alpha, 91% of power analyses use 5%. This shows how, in psychology, alpha is synonymous with 5% and it is rare for researchers to use a different alpha value.

The opposite problem is where we say there is not an effect when there actually is one. This is known as a type II error (said, type two error) or a false negative. In the Neyman-Pearson approach, this is the second element of using hypothesis testing to help us make decisions. In addition to alpha limiting how many type I errors (false positives) we are willing to make, we set beta to limit how many type II errors (false negatives) we are willing to make. Beta (sometimes written as "β") is the probability of concluding there *is not an effect when there really is one*. This is normally set at .20 (20%), which means we are willing to make a type II error 20% of the time in the long-run. By setting these two values, we are stating rules to help us make decisions and trying to limit how often we will be wrong in the long-run. We will consider how you can approach these decisions in part two.

One- and two-tailed tests

In significance testing, we describe the null hypothesis as a probability distribution centred on zero. We can reject the null hypothesis if our observed result is greater than a critical value determined by our alpha value. The area after the critical value creates a rejection region in the outer tails of the distribution. If the observed result is in this rejection region, we conclude the data would be unlikely assuming the null hypothesis is true, and reject the null hypothesis.

There are two ways of stating the rejection region. These are based on the type of alternative hypothesis we are interested in. There are two types of alternative hypotheses: (1) non-directional, and (2) directional. A non-directional hypothesis is simply a statement that there will be any effect, irrespective of the direction of the effect, e.g., 'Group A is different from Group B'. In contrast to this, the assumed null-hypothesis is 'Group A is *not* different from Group B'. Group A could be smaller than Group B, or Group A could be bigger than Group B. In both situations, the null hypothesis could be rejected. A directional hypothesis, on the other hand, is a statement that there will be a *specific* effect, e.g., 'Group A is

bigger than Group B'. Now, the assumed null hypothesis is 'Group A is not bigger than Group B'. In this instance even if we find evidence that Group B is bigger than Group A, the null hypothesis could not be rejected.

This is where the number of tails in a test comes in. In a two-tailed test (also known as a non-directional test), when alpha is set to 5%, there are two separate 2.5% areas to create a rejection region in both the positive and negative tails. Together, the two tails create a total area of 5%. To be statistically significant, the observed result can be in either the positive or negative rejection regions. Group A could be higher than group B, or group B could be higher than group A, you are just interested in a difference in any direction.

In a one-tailed test (also known as a directional test), there is just one larger area totalling 5% to create a rejection region in either the positive or negative tail (depending on the direction you are interested in). To be statistically significant, the critical value is slightly smaller, but the result must be in the direction you predicted. This means you would only accept a result of 'group A is bigger than group B' as significant. You could still find that group B is bigger than group A, but no matter how big the difference is, you cannot reject the null hypothesis as it is contrary to your directional prediction.

Statistical Power

Statistical power is defined as the probability of correctly deciding to reject the null hypothesis when the null hypothesis is not true. In plain English: the likelihood of successfully detecting an effect that is actually there (see Baguley, 2009 for other lay definitions). When we have sufficient statistical power, we are making the study sensitive enough to avoid making too many type II errors. Statistical power is related to beta where it is 1-β and typically expressed as a percentage. If we use a beta value of .20, that means we are aiming to have statistical power of 80% (1 - .20 = .80 = 80%).

Effect Size

For statistical power, we spoke about "detecting an effect that is actually there". The final piece of the power analysis puzzle is the smallest effect size of interest. An effect size can be defined as a number that expresses the magnitude of a phenomenon relevant to your research question (Kelley & Preacher, 2012). Depending on your research question, this includes the difference between groups or the association between variables. For example, you could study the relationship between how much alcohol you drink and reaction time. We could say "alcohol has the effect of slowing down reaction time". However, there is something missing from that statement. *How much* does alcohol slow down reaction time? Is one drop of alcohol enough or do you need to consume a full keg of beer before your reaction time decreases by just 1 millisecond? The smallest effect size of interest outlines what effect size you would consider practically meaningful for your research question.

Effect sizes can be expressed in two ways: as an unstandardised effect, or as a standardised effect. An *unstandardised* effect size is expressed in the original units of measurement. For example, if you complete a Stroop task, you measure response times to congruent and incongruent colour words in milliseconds. The mean difference in response time to congruent and incongruent conditions is an unstandardised effect size and will remain consistent across studies. This means you could say one study reporting a mean difference of 106ms had a larger effect than a study reporting a mean difference of 79ms.

Unstandardised effect sizes are easy to compare if the measurement units are consistent, but in psychology we do not always have easily comparable units. Many subdisciplines use Likert scales to measure an effect of interest. For example, in mental health research, one might be interested in how much anxiety someone experiences each week (participants are often given options such as "not at all", "a little", "neither a little nor a lot", "a lot", and "all the time"). These responses are not in easily interpretable measurements but, as scientists, we would still like to provide a numerical value to explain what an effect means[3].

This is where *standardised* effect sizes are useful as they allow you to compare effects across contexts, studies, or slightly different measures. For example, if study one used a five-point scale to measure anxiety but study two used a seven-point scale, a difference of two points on each scale has a different interpretation. A standardised effect size allows you to convert these differences into common measures, making it easier to compare results across studies using different units of measurement. There are many types of standardised effect sizes, such as Cohen's d or η^2 (said, eta squared), which we use in different contexts (see Lakens, 2013 for an overview). In this tutorial, we mainly focus on Cohen's d as the standardised mean difference as it is the effect size used in jamovi, the software we use in part three below. Although there are different formulas, Cohen's d is normally the mean difference divided by the pooled standard deviation. This means it represents the difference between groups or conditions, expressed as standard deviations instead of the original units of measure-

[3]Note, we use this as an example of measurements with different scales, but it is normally better to analyse ordinal data with ordinal models (see Bürkner & Vuorre, 2019).

ment.

Standardised and unstandardised effect sizes each have their strengths and weaknesses (Baguley, 2009). Unstandardised effect sizes are easier to interpret, particularly for lay readers who would find it easier to understand a difference of 150ms instead of 0.75 standard deviations. However, it can be harder to compare unstandardised effect sizes across studies when there are different measurement scales. Standardised effect sizes help with this as they convert measures to standardised units, making it easier to compare effect sizes across studies. However, the standardisation process can cause problems, as effect sizes can change depending on whether the design was within- or between-subjects, if the measures are unreliable, and if sampling affects the variance of the measures through restricting the values to a smaller range of the scale (Baguley, 2009). Similarly, the frame of reference is important when interpreting standardised effect sizes. When classifying the magnitude of standardised effects, Cohen (1988, pg. 25) specifically says “the terms "small," "medium," and "large" are relative, not only to each other, but to the area of behavioural science or even more particularly to the specific content and research method being employed in any given investigation”. Cohen emphasised that the qualitative labels (“small”, “medium,” and “large”) are arbitrarily applied to specific values, and should be applied differently to different fields. This means that interpretation should not simply follow rules of thumb that were established outside of the research field of interest. Although the software we introduce in part three relies on standardised effect sizes, Baguley (2009) emphasises it is better to focus on interpreting unstandardised effect sizes wherever possible.

To bring this back to statistical power (successfully detecting a true effect), the bigger an effect is, the easier it is to detect. In the anxiety example, we could compare the effects of two types of therapy. If the difference between therapy A and therapy B was, on average, 3 points on an anxiety scale, it would be easier to detect than if the average difference between therapy A and therapy B was 1 point. The smaller decrease of 1 point would be harder to detect than the larger decrease of 3 points. You would need to test more people in each therapy group to successfully detect this weaker effect because of the greater level of overlap between the two sets of therapy outcomes. It is this principle that allows us to say that the bigger the effect size, the easier it is to detect. In other words, if you have the same number of participants, statistical power increases as the effect size increases.

We have now covered the five main concepts underlying power analysis: alpha, beta, sample size, effect size, and one- or two-tailed tests. For ease, we have provided a summary of these concepts, their meaning, and how we often use them in Table 1. It takes time to appreciate the interrelationship between these concepts, so we recommend using the interactive visualisation by Magnusson (https://rpsychologist.com/d3/nhst/).

Types of power analysis

As the four main concepts behind power analysis are related, we can calculate one as the outcome if we state the other three. The most common types of power analysis relating to these outcomes are (1) *a priori*, (2) sensitivity, and (3) post-hoc. If we want sample size as the outcome, we use *a priori* power analysis to determine how many participants we need to reliably detect a given smallest effect size of interest, alpha, and power. Alternatively, if we want the effect size as the outcome, we can use sensitivity power analysis to determine what effect size we can reliably detect given a fixed sample size, alpha, and power.

There is also post-hoc power analysis if we want statistical power as the outcome given an observed effect size, sample size, and alpha. Post-hoc power analysis is an attractive idea, but it should not be reported as it essentially expresses the *p*-value in a different way. There is a direct relationship between observed power and the *p*-value of your statistical test, where a *p*-value of .05 means your observed power is 50% (Lakens, 2022). Remember, probability in frequentist statistics does not apply to individual events, so using the observed effect size in a single study ignores the role of the smallest effect size of interest in the long-run. As post-hoc power is uninformative, we only focus on *a priori* and sensitivity power analysis in this tutorial.

Part Two: Decision Making in Power Analysis

Now that you are familiar with the concepts, we turn our focus to decision making in power analysis. In part one, we defined the main inputs used in power analysis, but now you must decide on a value for each one. Setting your inputs is the most difficult part of power analysis as you must understand your area of research and be able to justify your choices (Lakens, 2022). Power analysis is a reflective process that is most effective during the planning stage of research, meaning that you must balance your inferential goals (what you want to find out) with the resources you have available (time, money, equipment, etc.). In this part, we will outline different strategies for choosing a value for alpha, beta/power, one- or two-sided tests, your smallest effect size of interest, and your sample size.

Table 1
Table showing the basic concepts underlying power analysis, what they mean, and how they are often used.

Concept	What it is	How it is often used
Alpha (α)	Cut-off value for how frequently we are willing to accept a false-positive.	This is traditionally set to .05 (5% of the time), but it is often set to lower thresholds in disciplines like physics. The lower alpha is, the fewer false-positives there will be in the long-run.
Beta (β)	Cut-off value for how frequently we are willing to accept a false-negative.	In psychology, this is usually set to .20 (20% of the time), implicitly suggesting false negatives are less of a concern than false positives. The lower beta is, the fewer false-negatives there will be in the long-run.
Power (1-β)	The chances of detecting an effect that exists.	The opposite of beta, power is how likely you are to detect a given effect size. This is usually set to .80 (80% of the time). The higher power is, the more likely you are to successfully detect a true effect if it is there.
Effect size	A number that expresses the magnitude of a phenomenon relevant to your research question.	Unstandardised effect sizes express the difference or relationship in the original units of measurement, such as milliseconds. Standardised effect-sizes express the difference or relationship in standardised units, such Cohen's *d*. Higher absolute effect sizes mean a larger difference or stronger relationship.
One-tailed test	When the rejection region in null hypothesis significance testing is limited to one tail in a positive or negative direction.	If you have a (ideally preregistered) clear directional prediction, one-tailed tests mean you would only reject the null hypothesis if the result was in the direction you predicted. The observed result may be in the extreme of the opposite tail, but you would still fail to reject the null hypothesis.
Two-tailed test	When the rejection region in null hypothesis significance testing is present in the extremes of both the positive and negative tail area.	If you would accept a result in any direction, you can use a two-tailed test to reject the null hypothesis if the observed result is in the extremes of either the positive or negative tail.
a priori power analysis	How many participants do we need to reliably detect a given smallest effect size of interest, alpha, and power?	We tend to use the term '*a priori*' in front of a power analysis that is conducted before data is collected. This is because we are deducing the number of participants from information we already have.
Sensitivity power analysis	What effect size could we detect with our fixed sample size, alpha, and desired power?	We use a sensitivity power analysis when we already know how many participants we have (e.g., using secondary data, or access to a rare population). We use this type of analysis to evaluate what effect sizes we can reliably detect.

Alpha

The first input to set is your alpha value. Traditionally, we use .05 to say we are willing to accept making a type I error up to 5% of the time. There is nothing special about using an alpha of .05, it was only a brief suggestion by Fisher (1926) for what felt right, but he emphasised you should justify your alpha for each experiment. Decades of tradition mean the default alpha is set to .05, but there are different approaches you can take to argue for a different value.

You could start with the traditional alpha of .05 but adjust it for multiple comparisons. For example, if you were planning on performing four related tests and wanted to correct for multiple comparisons, you could use this corrected alpha value in your power analysis. If you used the Bonferroni-Holm method (Cramer et al., 2016), the most stringent alpha value would be set as .0125 instead of .05. Using this lower alpha value would require more participants to achieve the same power, but you would ensure your more stringent test had your desired level of statistical power.

Alternatively, you could argue your study requires deviating from the traditional .05 alpha value. One approach is to switch between a .05 alpha for suggestive findings and a .005 alpha for confirmatory findings (Benjamin et al., 2018). This means if you have a strong prediction, or one that has serious theoretical implications, you could argue your study requires a more stringent alpha value. Theoretical physicists take this approach of using a more stringent value even further, and use an alpha value of .0000003 (known as 'five sigma'). The reason for having such a stringent alpha level is that to make changes to our understanding of physics would have knock-on effects to all other sciences, so avoiding false positives is of the utmost importance. Another approach is to justify your bespoke alpha for each study (Lakens et al., 2018). The argument here is you should perform a cost-benefit analysis to determine your alpha based on how difficult it is to recruit your sample. See Maier and Lakens (2021) for a primer on justifying your alpha.

Beta

Beta also has a traditional value: most studies aim for a beta of .20, meaning they want 80% power. Cohen (1965) suggested the use of 80% as he felt that type II errors are relatively less serious than type I errors. At the time of writing, 80% power would lead to roughly double the sample sizes than the studies he critiqued in his review (Cohen, 1962). Aiming for 80% power has proved influential as Bakker et al. (2020) found it was the most common value researchers reported in their power analysis.

Aiming for 80% power was largely a pragmatic approach, so you may argue it is not high enough. Cohen (1965) explicitly stated that you should ignore his suggestion of 80% if you have justification for another value. Setting beta to .20 means you are willing to accept a type II error 20% of the time. This implicitly means type II errors are four times less serious than type I errors when alpha is set to .05. To match the error rates, Bakker et al. (2020) found the next most common value was 95% power (beta = .05), but it only represented 19% of power analyses in their sample.

Earlier, we mentioned working with rare populations. Many such populations (such as those with rare genetic conditions) may receive specialist care or support. If one were to assess the effectiveness of this specialist care/support, then not finding an effect that does exist (a type II error) might lead to this support being taken away. In such circumstances, you could argue that type II errors are just as important to avoid, if not more important, than type I errors. As such, in these circumstances, you might want to increase your power (have a lower beta value), to avoid undue harm to a vulnerable population.

Deciding on the beta value for your own study will involve a similar process to justifying your alpha. You must decide what your inferential goals are and whether 80% power is enough, knowing you could miss out on a real effect 20% of the time. However, if you increase power to 90% or 95%, it will require more participants, so you must perform a cost-benefit analysis based on how easy it will be to recruit your sample (Lakens, 2022).

One- and two-tailed tests

In tests comparing two values, such as the difference between two groups or the relationship between two variables, you can choose a one- or two-tailed test. Lakens (2016a) argued one-tailed tests are underused and offer a more efficient procedure. As the rejection region is one 5% area (instead of two 2.5% areas), the critical value is smaller, so holding everything else constant, you need fewer participants for a statistically significant result. One-tailed tests also offer a more severe test of a hypothesis since the observed result must reach the rejection region in the hypothesised direction. The *p*-value in your test may be smaller than alpha, but if the result is in the opposite direction to what you predicted, you still cannot reject the null hypothesis. This means one-tailed tests can be an effective option when you have a strong directional prediction.

One-tailed tests are not always appropriate though, so it is important you provide clear justification for why

you are more interested in an effect in one direction and not the other (Ruxton & Neuhäuser, 2010). If you would be interested in an effect in either a positive or negative direction, then a two-tailed test would be better suited. One-tailed tests have also been a source of suspicion since they effectively halve the p-value. For example, Wagenmakers et al. (2011) highlighted how some studies took advantage of an opportunistic use of one-tailed tests for their results to be statistically significant. This means one-tailed tests are most convincing when combined with preregistration (see Kathawalla et al. (2021) if preregistration is a procedure you are unfamiliar with) as you can demonstrate that you had a clear directional hypothesis and planned to test that hypothesis with a one-tailed test.

Effect size

In contrast to alpha and beta, there is not one traditional value for your choice of effect size. Many studies approach power analysis with a single effect size and value for power in mind, but as we will demonstrate in part three, power exists along a curve. In most cases, you do not know what the exact effect size is, or you would not need to study it. The effect size you use is essentially the inflection point for which effect sizes you want sufficient power to detect. If you want 80% power for an effect of Cohen's $d = 0.40$, you will be able to detect effects of 0.40 with 80% power. You will have increasingly *higher* levels of power for effects larger than 0.40, but increasingly *lower* levels of power for effects smaller than 0.40. This means it is important to think of your *smallest effect size of interest*, as you are implicitly saying you do not care about detecting effects smaller than this value.

The most difficult part of an *a priori* power analysis is justifying your smallest effect size of interest. Choosing an effect size to use in power analysis and interpreting effect sizes in your study requires subject matter expertise (Panzarella et al., 2021). You must decide what effect sizes you consider important or meaningful based on your understanding of the measures and designs in your area. For example, is there a relevant theory that outlines expected effects; what have studies testing similar hypotheses found; what are the practical implications of the results? Some of these decisions are difficult to make and all the strategies are not always available, but there are different sources of information you can consult for choosing and justifying your smallest effect size of interest (Lakens, 2022).

First, you could identify a meta-analysis relevant to your area of research which summarises the average effect across several studies. If you want to use the estimate to inform your power analysis, you must think about whether the results in the meta-analysis are similar to your planned study. Sometimes, meta-analyses will report a broad meta-analysis amalgamating all the results, then report moderator analyses for the type of studies they include, so you could check whether there is an estimate restricted to methods similar to your study. We also know meta-analyses can report inflated effect sizes due to publication bias (Lakens, 2022), meaning you can look for a more conservative estimate such as the lower bound of the confidence interval around the average effect or if the authors report a bias-corrected average effect.

Second, there may be one key study you are modelling your project on. You could use their effect size to inform your power analysis, but as Panzarella et al. (2021) warn, effect sizes are best interpreted in context, so question how similar your planned methods are. As a result from a single study, there will be uncertainty around their estimate, so think about the width of the confidence interval around their effect size and use a conservative estimate.

Finally, you can consult effect size distributions. The most popular guidelines are from Cohen (1988) who argued you should use $d = 0.2$ for new areas of research as the measures are likely to be imprecise, 0.5 for phenomena observable to the naked eye, and 0.8 for differences you hardly need statistics to detect. Cohen (1988) explicitly warned these guidelines were for new areas of research, when there was nothing else to go on. But, like many heuristics, the original suggestions have lost their nuance and are now taken as a ubiquitous 'rule of thumb'. It is important to consider what effect sizes mean for your subject area (Baguley, 2009), but researchers seldom critically choose an effect size. An analysis of studies that did justify their effect size found that the majority of studies simply cited Cohen's suggested values (Bakker et al., 2020). Relying on these rules of thumb can lead to strange interpretations, such as paradoxes where even incredibly small effect sizes (using Cohen's rule of thumb) can be meaningful. Abelson (1985) found that an R^2 of .003 was the effect size of the most significant characteristic predicting baseball success (batting average). In context, then, an R^2 of .003 is clearly meaningful, so it is important to interpret effect sizes in context rather than apply broad generalisations. If you must rely on effect size distributions, there are articles which are sub-field specific. For example, Gignac and Szodorai (2016) collated effects in individual differences research and Szucs and Ioannidis (2021) outlined effects in cognitive neuroscience research.

Effect size distributions can be useful to calibrate your understanding of effect sizes in different areas but

they are not without fault. Panzarella et al. (2021) demonstrated that in the studies that cited effect size distributions, most used them to directly interpret the effect sizes they observed in their study (e.g., "in this study we found a 'large' effect, which means..."). However, as seen in Abelson's paradox, small effects in one context can be meaningful in another context. Effect size distributions can help to understand the magnitude of effect sizes within and across subject areas, but comparing your observed effect size to an amalgamation of effects across all of psychology leads to a loss in nuance. If you have no other information, effect size distributions can help to inform your smallest effect size of interest, but when it comes to interpretation it is important to put your effect size in context compared to studies investigating a similar research question.

With these strategies in mind, it is important to consider what represents the smallest effect size of interest for your specific study. It is the justification that is important as there is no single right or wrong answer. Power analysis is always a compromise between designing an informative study and designing a feasible study for the resources at your disposal. You could always set your effect size to $d = 0.05$, but the sample size required would often be unachievable and effects this small may not be practically meaningful. Therefore, you must explain and justify what represents the smallest effect size of interest for your area of research.

Sample size

The final input you can justify is your sample size. You may be constrained by resources or the population you study, meaning you know the sample size and want to know what effect sizes you could detect in a sensitivity power analysis.

Lakens (2022) outlined that two strategies for sample size justification include measuring an entire population and resource constraints. If you study a specific population, such as participants with a rare genetic condition, you might know there are only 30 participants in your country which you regularly study, placing a limit on the sample size you can recruit. Alternatively, in many student projects, the time or money available to conduct research is limited, so the sample size may be influenced by resource constraints. You might have £500 for recruitment and if you pay them £10 for an hour of their time, you only have enough money for 50 participants.

In both scenarios, you start off knowing what your sample size will be. This does not mean you can ignore statistical power, but it changes from calculating the necessary sample size to detect a given effect size, to what effect size you can detect with a given sample size. This allows you to decide whether the study you plan on conducting is informative, or if it would be uninformative, you would have the opportunity to change the design or use more precise measures to produce larger effect sizes.

Part Three: Power Analysis using jamovi

For this tutorial, we will be using the open source software jamovi (2021). Although it currently offers a limited selection for power analysis, it is perfect for an introduction for three reasons. First, it is free and accessible on a wide range of devices. Historically, G*Power (Faul et al., 2009) was a popular choice, but it is no longer under active development which presents accessibility issues. Second, the output in jamovi contains written guidance on how to interpret the results and emphasises underrepresented concepts like power existing along a curve. Finally, Bakker et al. (2020) observed authors often fail to provide enough information to reproduce their power analysis. In jamovi, you save your output and options in one file, meaning you can share this file to be fully reproducible. Combined, these features make jamovi the perfect software to use in this tutorial.

To download jamovi to your computer, navigate to the download page (https://www.jamovi.org/download.html) and install the solid version to be the most stable. Once you open jamovi, click Modules (in the top right, Figure 1a in red), click jamovi library (Figure 1a in blue), and scroll down in the Available tab until you see jpower and click INSTALL (Figure 1b in green). This is an additional module written by Morey and Selker which appears in your jamovi toolbar.

In the following sections, imagine we are designing a study to build on Irving et al. (2022) who tested an intervention to correct statistical misinformation. Participants read an article about a new fictional study where one passage falsely concludes watching TV causes cognitive decline. In the correction group, participants receive an extra passage where the fictional researcher explains they only reported a correlation, not a causal relationship. In the no-correction group, the extra passage just explains the fictional researcher was not available to comment. Irving et al. then tested participants' comprehension of the story and coded their answers for mistaken causal inferences. They expected participants in the correction group to make fewer causal inferences than those in the no-correction group, and found evidence supporting this prediction with an effect size equivalent to Cohen's $d = 0.64$, 95% CI = [0.28, 0.99]. Inspired by their study, we want to design an experiment to correct another type of misinformation in articles.

Irving et al. (2022) themselves provide an excellent

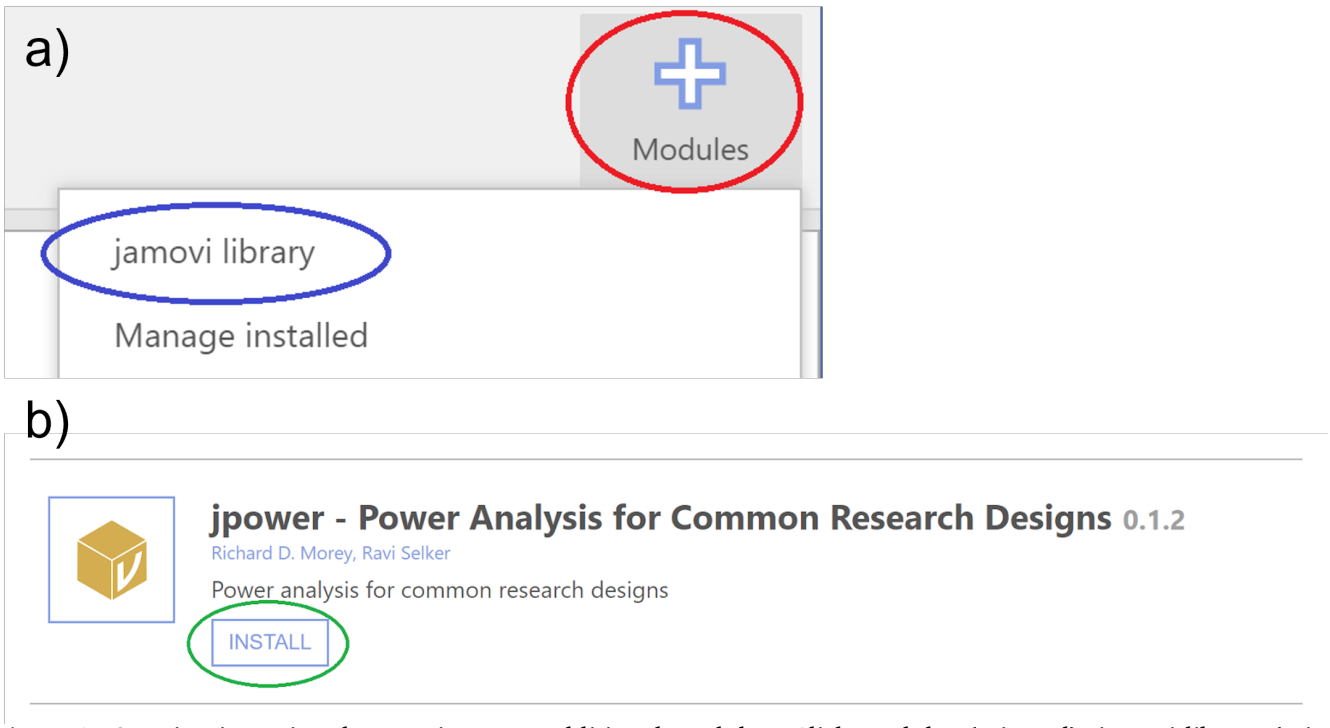


Figure 1. Opening jamovi and managing your additional modules. Click modules (a in red), jamovi library (a in blue) to manage your modules, and scroll down to install jpower (b in green) to be able to follow along to the tutorial.

example of explaining and justifying the rationale behind their power analysis, so we will walk through the decision making process and how it changes the outputs. For our smallest effect size of interest, our starting point is the estimate of $d = 0.64$. However, it is worth consulting other sources to calibrate our understanding of effects in the area, such as Irving et al. citing a meta-analysis by Chan et al. (2017). For debunking, the average effect across 30 studies was $d = 1.14$, 95% CI = [0.68, 1.61], so we could use the lower bound of the confidence interval, but this may still represent an overestimate. Irving et al. used the smallest effect ($d = 0.54$) from the studies most similar to their design which was included in the meta-analysis. As a value slightly smaller than the other estimates, we will also use this as the smallest effect of interest for our study.

Now we have settled on our smallest effect size of interest, we will use $d = 0.54$ in the following demonstrations. We start with *a priori* and sensitivity power analysis for two independent samples, exploring how the outputs change as we alter inputs like alpha, power, and the number of tails in the test. For each demonstration, we explain how you can transparently report the power analysis to your reader. We then repeat the demonstrations for two dependent samples to show how you require fewer participants when the same participants complete multiple conditions instead of being allocated to separate groups.

Two Independent samples

A priori power analysis in independent samples. If you open jamovi, you should have a new window with no data or output. For an independent samples *t*-test, make sure you are on the analyses tab, click on jpower, and select Independent Samples T-Test. This will open the window shown in Figure 2.

We will start by calculating power *a priori* for an independent samples *t*-test. On the left side, you have your inputs and on the right side, you have the output from your choices. Depending on the type of analysis you select under Calculate, one of the inputs will be blanked out in grey. This means it is the parameter you want as the output on the right side and you will not be able to edit it. To break down the main menu options in this window:

- Calculate: Your choice of calculating one of (a) the minimum number of participants (N per group) needed for a given effect size, (b) what your power is given a specific effect size and sample size, or (c) the smallest effect size that you could reliably detect given a fixed sample size.

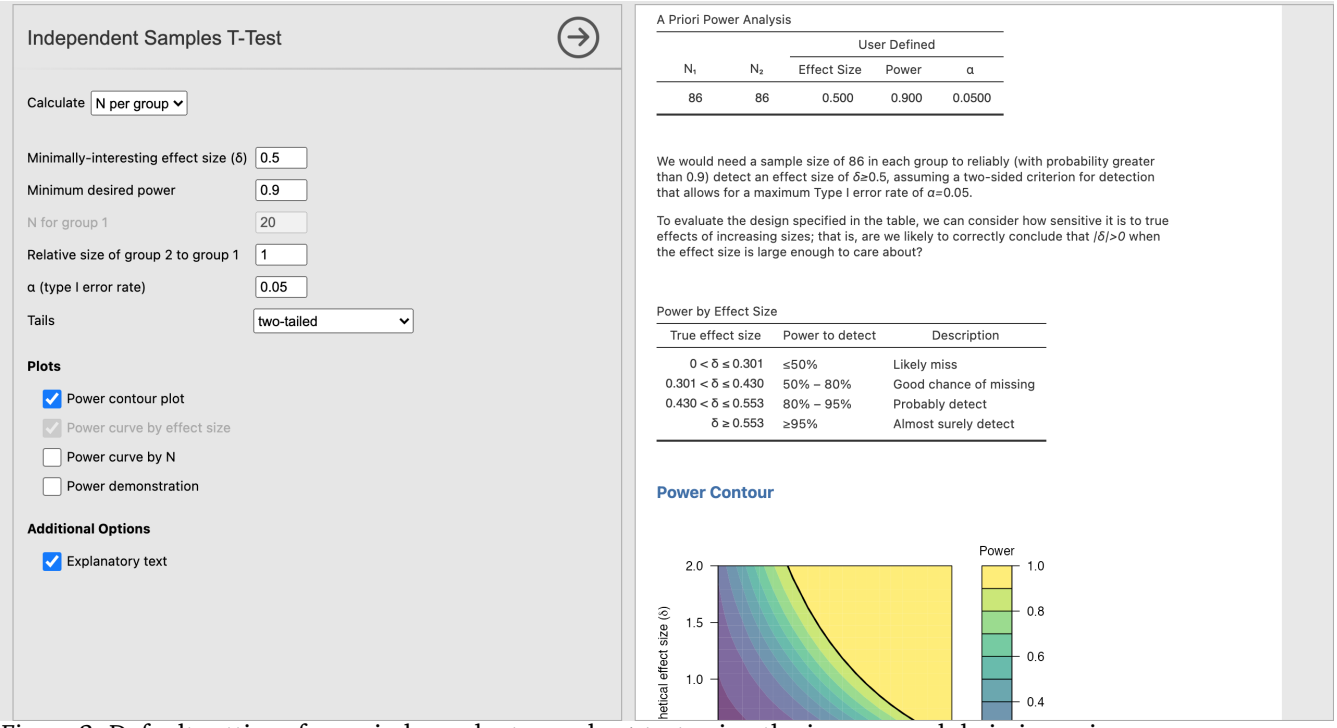


Figure 2. Default settings for an independent samples *t*-test using the jpower module in jamovi.

- Minimally-interesting effect size: This is the standardised effect size known as Cohen's *d*. Here we can specify our smallest effect size of interest.
- Minimum desired power: This is our long-run power. Power is traditionally set at .80 (80%) but some researchers argue that this should be higher at .90 (90%) or .95 (95%). The default setting in jpower is .90 (90%) but see part two for justifying this value.
- N for group 1: This input is currently blanked out as in this example we are calculating the minimum sample size, but here you would define how many participants are in the first group.
- Relative size of group 2 to group 1: If this is set to 1, the sample size is calculated by specifying equal group sizes. You can specify unequal group sizes by changing this input. For example, 1.5 would mean group 2 is 1.5 times larger than group 1, whereas 0.5 would mean group 2 is half the size of group 1.
- α (type I error rate): This is your long-run type one error rate which is conventionally set at .05. See part two for strategies on justifying a different value.
- Tails: Is the test one- or two-tailed? You can specify whether you are looking for an effect in just one direction or you would be interested in any significant result.

For this example, our smallest effect size of interest will be $d = 0.54$ following our discussion of building on Irving et al. (2022). We can enter the following inputs: effect size $d = 0.54$, alpha = .05, power = .90, relative size of group 2 to group 1 = 1, and two-tailed. You should get the output in Figure 3. The first table "A Priori Power Analysis" tells us that to detect our smallest effect size of interest, we would need two groups of 74 participants (N = 148) to achieve 90% power in a two-tailed test.

In jamovi, it is clear how statistical power exists along a curve. The second table in Figure 3 "Power by Effect Size" shows us what range of effect sizes we would likely detect with 74 participants per group. We would have 80-95% power to detect effect sizes between $d = 0.46$-0.60. However, we would only have 50-80% power to detect effects between $d = 0.32$-0.46. This shows our smallest effect size of interest could be detected with 90% power, but smaller effects have lower power and larger effects would have higher power.

This is also reflected in the power contour plot, which is reported by default (Figure 4). If you cannot see the plot, make sure the "Power contour plot" option is ticked under Plots and scroll down in the Results window as it is included at the bottom. The level of power you choose is the black line that curves from the top

Independent Samples T-Test

The purpose of a *power analysis* is to evaluate the sensitivity of a design and test. You have chosen to calculate the minimum sample size needed to have an experiment sensitive enough to consistently detect the specified hypothetical effect size.

A Priori Power Analysis

		User Defined		
N_1	N_2	Effect Size	Power	α
74	74	0.540	0.900	0.0500

We would need a sample size of 74 in each group to reliably (with probability greater than 0.9) detect an effect size of $\delta \geq 0.54$, assuming a two-sided criterion for detection that allows for a maximum Type I error rate of $\alpha = 0.05$.

To evaluate the design specified in the table, we can consider how sensitive it is to true effects of increasing sizes; that is, are we likely to correctly conclude that $|\delta| > 0$ when the effect size is large enough to care about?

Power by Effect Size

True effect size	Power to detect	Description
$0 < \delta \leq 0.324$	≤50%	Likely miss
$0.324 < \delta \leq 0.464$	50% – 80%	Good chance of missing
$0.464 < \delta \leq 0.597$	80% – 95%	Probably detect
$\delta \geq 0.597$	≥95%	Almost surely detect

Figure 3. *priori* power analysis results for a two-tailed independent samples *t*-test using $d = 0.54$ as the smallest effect size of interest. We would need 74 participants per group (N = 148) for 90% power.

Power Contour

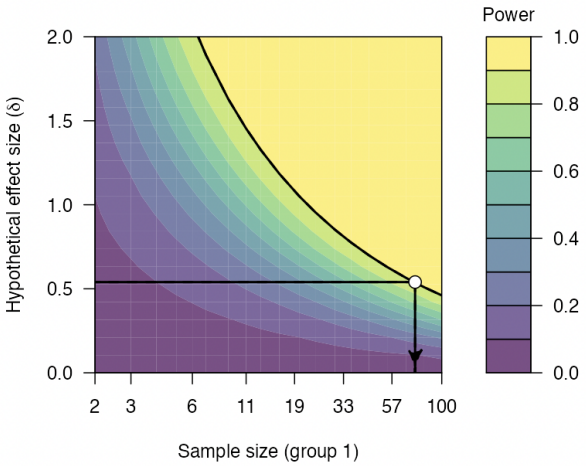


Figure 4. A power contour to show how as the effect size decreases (smaller values on the y-axis), the number of participants required to detect the effect increases (higher values on the x-axis). Our desired level of 90% power is indicated by the black curved line.

left to the bottom right. For our effect size, we travel along the horizontal black line until we reach the curve, and the down arrow tells us we need 74 participants per group. For larger effects as you travel up the curve, we would need fewer participants and for smaller effects down the curve we would need more participants.

Now that we have explored how many participants we would need to detect our smallest effect size of interest, we can alter the inputs to see how the number of participants changes. Wherever possible, it is important to perform a power analysis before you start collecting data, as you can explore how changing the inputs impacts your sample size.

- Tail(s): If you change the number of tails to one, this decreases the number of participants in each group from 74 to 60. This saves a total of 28 participants (14 in each group). If your experiment takes 30 minutes per participant, that is saving you 14 hours' worth of work or cost while still providing your experiment with sufficient power.
- α: If you change α to .01, we would need 104 participants in each group (for a two-tailed test), 60 more participants than our first estimate and 30 more hours of data collection.
- Minimum desired power: If we decreased power to the traditional 80%, we would need 55 participants per group (for a two-tailed test; alpha = .05). This would be 38 fewer participants than our first estimate, saving 19 hours of data collection.

It is important to balance creating an informative experiment with the resources available. Therefore, it is crucial that, where possible, you perform a power analysis in the planning phase of a study as you can make these kinds of decisions before you recruit any participants. You can make fewer type one (decreasing alpha) or type two (increasing power) errors, but you must recruit more participants.

In the original power analysis by Irving et al. (2022), they used inputs of $d = 0.54$, alpha = .05, power = .95, and one-tailed for a directional prediction, and they aimed for two groups of 75 (N = 150) participants. In these demonstrations, we are walking through changing the inputs to see how it affects the output, but you can look at their article for a good example of justifying and reporting a power analysis.

How can this be reported? Bakker et al. (2020) warned that only 20% of power analyses contained enough information to be fully reproducible. To report your power analysis, the reader needs the following four key pieces of information:

- The type of test being conducted,
- The software used to calculate power,

- The inputs that you used, and
- Why you chose those inputs.

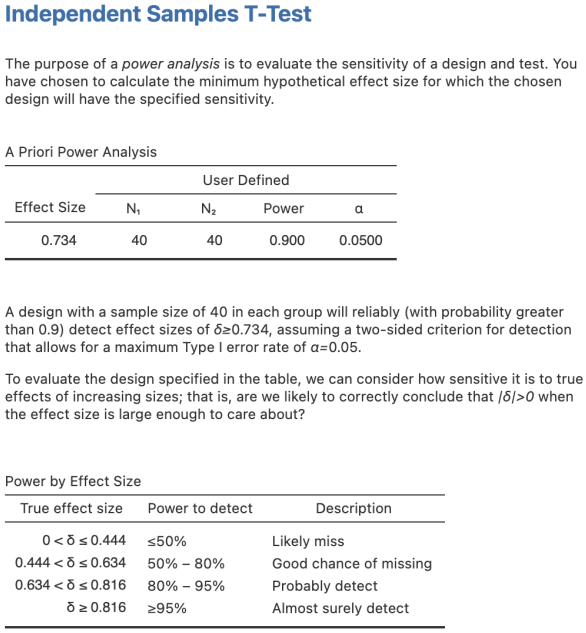
For the original example in Figure 3, we could report it like this: *"To detect an effect size of Cohen's d = 0.54 with 90% power (alpha = .05, two-tailed), the jpower module in jamovi suggests we would need 74 participants per group (N = 148) for an independent samples t-test. Similar to Irving et al. (2022), the smallest effect size of interest was set to d = 0.54, but we used a two-tailed test as we were less certain about the direction of the effect."*

This provides the reader with all the information they would need in order to reproduce the power analysis and ensure you have calculated it accurately. The statement also includes your justification for the smallest effect size of interest. Please note there is no single 'correct' way to report a power analysis. Just be sure that you have the four key pieces of information.

Sensitivity power analysis in independent samples. Selecting the smallest effect size of interest for an *a priori* power analysis would be an effective strategy if you wanted to calculate how many participants you need when designing your study. Now imagine you already knew the sample size or had access to a population of a known size. In this scenario, you would conduct a sensitivity power analysis. This would tell you what effect sizes your study would be powered to detect for a given alpha, power, and sample size. This is helpful for interpreting your results as you can outline what effect sizes your study was sensitive to and which effects would be too small for you to reliably detect. If you change the "Calculate" input to Effect size, "Minimally-interesting effect size" will now be greyed out.

Imagine we had finished collecting data and we knew we had 40 participants in each group but did not conduct a power analysis when designing the study. If we enter 40 for N for group 1, 1 for relative size of group 2 to group 1, alpha = .05, power = .90, and two-tailed, we get the output in Figure 5.

The first table in Figure 5 "A Priori Power Analysis" tells us that the study is sensitive to detect effect sizes of $d = 0.73$ with 90% power (note, the table is still referred to as *a priori*, despite it being a sensitivity power analysis. This is a quirk of the software. Do not worry, it is running a sensitivity analysis). This helps us to interpret the results if we did not plan with power in mind or we had a rare sample. The second table in Figure 5 "Power by Effect Size" shows we would have 80-95% power to detect effect sizes between $d = 0.63$-0.82, but 50-80% power to detect effect sizes between $d = 0.44$-0.63. As the effect size gets smaller, there is less chance of detecting it with 40 participants per group,



Independent Samples T-Test

The purpose of a *power analysis* is to evaluate the sensitivity of a design and test. You have chosen to calculate the minimum hypothetical effect size for which the chosen design will have the specified sensitivity.

A Priori Power Analysis

	User Defined			
Effect Size	N_1	N_2	Power	α
0.734	40	40	0.900	0.0500

A design with a sample size of 40 in each group will reliably (with probability greater than 0.9) detect effect sizes of $\delta \geq 0.734$, assuming a two-sided criterion for detection that allows for a maximum Type I error rate of $\alpha = 0.05$.

To evaluate the design specified in the table, we can consider how sensitive it is to true effects of increasing sizes; that is, are we likely to correctly conclude that $|\delta| > 0$ when the effect size is large enough to care about?

Power by Effect Size

True effect size	Power to detect	Description
$0 < \delta \leq 0.444$	≤50%	Likely miss
$0.444 < \delta \leq 0.634$	50% – 80%	Good chance of missing
$0.634 < \delta \leq 0.816$	80% – 95%	Probably detect
$\delta \geq 0.816$	≥95%	Almost surely detect

Figure 5. Sensitivity power analysis results for an independent samples *t*-test when there is a fixed sample size of 40 per group (N = 80). We would be able to detect an effect size of $d = 0.73$ with 90% power.

but we would have greater than 90% power to detect effect sizes larger than $d = 0.73$.

To acknowledge how power exists along a curve, we also get a second type of graph. We now have a power curve (Figure 6) with the x-axis showing the potential effect size and the y-axis showing what the power would be for that potential effect size. If this plot is not visible in the output, make sure you have "Power curve by effect size" ticked in the Plots options. This tells us how power changes as the effect size increases or decreases, with our other inputs held constant.

At 90% power, we can detect effect sizes of $d = 0.73$ or larger. If we follow the black curve towards the bottom left, power decreases for smaller effect sizes. This shows that once we have a fixed sample size, power exists along a curve for different effect sizes. When interpreting your results, it is important you have sufficient statistical power to detect the effects you do not want to miss out on. If the sensitivity power analysis suggests you would miss effects you would consider meaningful, you would need to calibrate your expectations of how informative your study is.

How can this be reported? We can also state the results of a sensitivity power analysis in a report. If you did not perform an *a priori* power analysis, you could report this in the method to comment on your final sample size. If you are focusing on interpreting how informa-

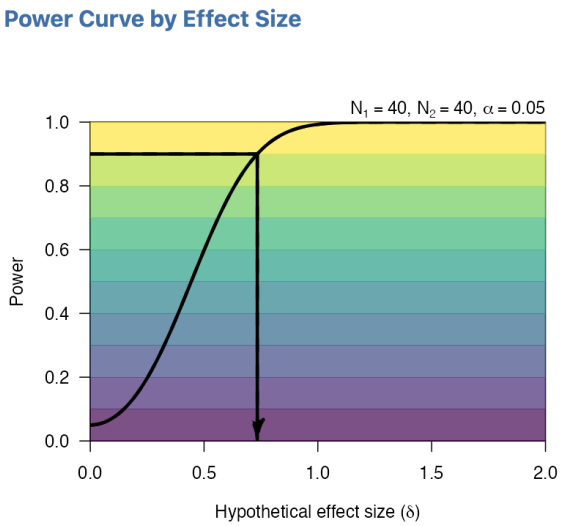


Figure 6. A power curve to show how as the effect size decreases (smaller values on the x-axis), we would have less statistical power (lower values on the y-axis) for our fixed sample size. Our desired level of 90% power is indicated by the intersection of the black horizontal line and the black curved line.

tive your results are, you could explore it in the discussion. Much like an *a priori* power analysis, there are key details that must be included to ensure it is reproducible and informative. For the example in Figure 5, you could report:

"The jpower module in jamovi suggests an independent samples t-test with 40 participants per group (N = 80) would be sensitive to effects of Cohen's d = 0.73 with 90% power (alpha = .05, two-tailed). This means the study would not be able to reliably detect effects smaller than Cohen's d = 0.73".

As with an *a priori* power analysis, there are multiple ways you can describe the sensitivity power analysis with the example above demonstrating one way of doing so. The main goal is to communicate the four key pieces of information to ensure your reader could reproduce the sensitivity power analysis and confirm you calculated it accurately.

Two dependent samples

A priori power analysis in dependent samples. Now we will demonstrate how you can conduct a power analysis for a within-subjects design. This time, you need to select Paired Samples T-Test from the jpower menu to get a window like Figure 7.

The inputs are almost identical to what we used for the independent samples *t*-test, but this time we only have four inputs as we do not need to worry about the ratio of group 2 to group 1. In a paired samples *t*-test, every participant must contribute a value for each condition. If we repeat the inputs from our independent samples *t*-test *a priori* power analysis (d = 0.54, alpha = .05, power = .90, two-tailed), your window should look like Figure 8.

The table "A Priori Power Analysis" suggests we would need 39 participants to achieve 90% power to detect our smallest effect size of interest (d = 0.54) inspired by Irving et al. (2022). We would need 109 fewer participants than our first estimate, saving 54.5 hours of data collection assuming your experiment takes 30 minutes. We also have the second table "Power by Effect Size" to show how power changes for different effect size ranges.

Before we move on to how to report the power analysis, we will make a note of the important lesson that using a within-subjects design will always save you participants. The reason for this is that instead of every participant contributing just one value (which may have measurement error because of extraneous variables), they are contributing two values (one to each condition). The error caused by many of the extraneous variables (such as their age, eye sight, strength, or any other participant-specific variable that might cause error) are the same for both conditions for the same person. The less error in our measurements there is, the more sure we can be that the results we see are due to our manipulation. As within-participants designs lower the amount of error compared to between-participants design, they need fewer participants to achieve the same amount of power. The amount of error accounted for in a within-participants design means you need approximately half the number of participants you need to detect the same effect size in a between-subjects design (Lakens, 2016b). When you are designing a study, think about whether you could convert the design to within-subjects to make it more efficient.

While it helps save on participants, it is not always possible, or practical, to use a within-subjects design. For example, in the experiment we are designing here, participants are shown two versions of a news story with a subtle manipulation. A between-subjects design might be a better choice as participants are randomised into one of two groups and they do not see the alternative manipulation. This means participants would find it more difficult to work out the aims of the study and change their behaviour. In a within-subjects design you would need at least two versions of the news story to create one 'correction' condition and one 'no correction' condition. This means participants would experience both conditions and they could work out the aims of

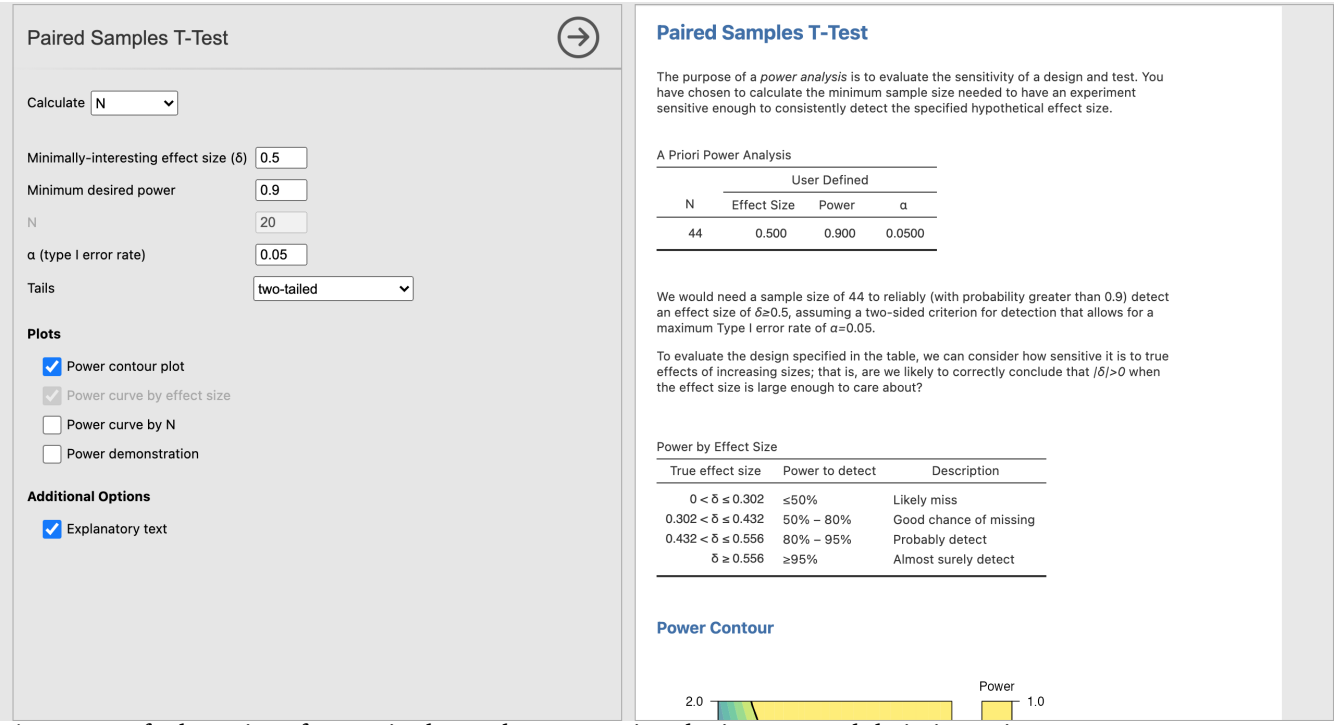


Figure 7. Default settings for a paired samples *t*-test using the jpower module in jamovi.

the study and potentially change their behaviour. In addition, you would need to ensure the two versions of the news story were different enough that participants did not simply provide the same answer, but comparable enough to ensure you are not introducing a confound. This is another example of where thinking of statistical power in the design stage of research is most useful. You can decide whether a within- or between-subjects design is best suited to your procedure.

How can this be reported? For the example in Figure 8, you could report: *"To detect an effect size of Cohen's d = 0.54 with 90% power (alpha = .05, two-tailed), the jpower module in jamovi suggests we would need 39 participants for a paired samples t-test. Similar to Irving et al. (2022), the smallest effect size of interest was set to d = 0.54, but we used a two-tailed test as we were less certain about the direction of the effect."*.

Sensitivity power analysis in dependent samples. If you change "Calculate" to Effect size, we can see what effect sizes a within-subjects design is sensitive enough to detect. Imagine we sampled from 30 participants without performing an *a priori* power analysis. Setting the inputs to power = .90, N = 30, alpha = .05, and two-tailed; you should get the output in Figure 9.

The "A Priori Power Analysis" table shows us that the design would be sensitive to detect an effect size of d = 0.61 with 90% power using 30 participants. This helps us to interpret the results if we did not plan with power in mind or had a limited sample. The second table in Figure 9 "Power by Effect Size" shows we would have 80-95% power to detect effect sizes between d = 0.53-0.68, but 50-80% power to detect effect sizes between d = 0.37-0.53. As the effect size gets smaller, there is less chance of detecting it with 30 participants, but we would have greater than 90% power to detect effect sizes larger than d = 0.61.

How can this be reported? For the example in Figure 9, you could report: *"The jpower module in jamovi suggests a paired samples t-test with 30 participants would be sensitive to effects of Cohen's d = 0.61 with 90% power (alpha = .05, two-tailed). This means the study would not be able to reliably detect effects smaller than Cohen's d = 0.61"*.

Conclusion

In this tutorial, we demonstrated how to perform a power analysis for both independent and paired samples *t*-tests using the jpower module in jamovi. We outlined two of the most useful types of power analysis: (1) *a priori*, for when you want to know how many participants you need to detect a given effect size, and (2) sensitivity, for when you want to know what effect sizes you can detect with a given sample size. We also emphasised the key information you must report to ensure power analyses are reproducible. Our aim was to provide a beginner's tutorial to learn the fundamental

Paired Samples T-Test

The purpose of a *power analysis* is to evaluate the sensitivity of a design and test. You have chosen to calculate the minimum sample size needed to have an experiment sensitive enough to consistently detect the specified hypothetical effect size.

A Priori Power Analysis

	User Defined		
N	Effect Size	Power	α
39	0.540	0.900	0.0500

We would need a sample size of 39 to reliably (with probability greater than 0.9) detect an effect size of $\delta \geq 0.54$, assuming a two-sided criterion for detection that allows for a maximum Type I error rate of $\alpha = 0.05$.

To evaluate the design specified in the table, we can consider how sensitive it is to true effects of increasing sizes; that is, are we likely to correctly conclude that $|\delta| > 0$ when the effect size is large enough to care about?

Power by Effect Size

True effect size	Power to detect	Description
$0 < \delta \leq 0.322$	≤50%	Likely miss
$0.322 < \delta \leq 0.460$	50% – 80%	Good chance of missing
$0.460 < \delta \leq 0.592$	80% – 95%	Probably detect
$\delta \geq 0.592$	≥95%	Almost surely detect

Figure 8. A priori power analysis results for a paired samples *t*-test using $d = 0.54$ as the smallest effect size of interest. We would need 39 participants for 90% power.

Paired Samples T-Test

The purpose of a *power analysis* is to evaluate the sensitivity of a design and test. You have chosen to calculate the minimum hypothetical effect size for which the chosen design will have the specified sensitivity.

A Priori Power Analysis

	User Defined		
Effect Size	N	Power	α
0.612	30	0.900	0.0500

A design with a sample size of 30will reliably (with probability greater than 0.9) detect effect sizes of $\delta \geq 0.612$, assuming a two-sided criterion for detection that allows for a maximum Type I error rate of $\alpha = 0.05$.

To evaluate the design specified in the table, we can consider how sensitive it is to true effects of increasing sizes; that is, are we likely to correctly conclude that $|\delta| > 0$ when the effect size is large enough to care about?

Power by Effect Size

True effect size	Power to detect	Description
$0 < \delta \leq 0.370$	≤50%	Likely miss
$0.370 < \delta \leq 0.529$	50% – 80%	Good chance of missing
$0.529 < \delta \leq 0.681$	80% – 95%	Probably detect
$\delta \geq 0.681$	≥95%	Almost surely detect

Figure 9. Sensitivity power analysis results for a paired samples *t*-test when there is a fixed sample size of 30 participants. We would be able to detect an effect size of $d = 0.61$ with 90% power.

concepts of power analysis, so you can build on these lessons and apply them to more complicated designs.

There are three key lessons to take away from this tutorial. First, you can plan to make fewer type one (decreasing alpha) or type two (increasing power) errors, but it will cost more participants assuming you want to detect the same effect size. Second, using a one-tailed test offers a more severe test of a hypothesis and requires fewer participants to achieve the same level of power. Finally, using a within-subjects design requires fewer participants than a between-subjects design.

Power analysis is a reflective process, and it is important to keep these three lessons in mind when designing your study. Designing an informative study is a balance between your inferential goals and the resources available to you (Lakens, 2022). That is why we framed changes in the inputs around how many hours of data collection your study would take assuming it lasted 30 minutes. You will rarely have unlimited resources as a researcher, either from the funding body supporting your research, or from the number of participants in your population of interest. Planning your study with statistical power in mind provides you with the most flexibility as you can make decisions, like considering a one-tailed test or using a within-subjects design, before you can preregister and conduct your study.

The number of participants required for a sufficiently powered experiment might have surprised you. Depending on the inputs and design, we needed between 39 and 208 participants to detect the same smallest effect size of interest ($d = 0.54$) to build on Irving et al. (2022). For resource-limited studies like student dissertations or participant-limited studies on rare populations, getting so many participants may be unachievable. As a result, changing to a within-participants design, or changing the other inputs might be needed, where possible. In circumstances where within-participants designs are not possible, and changing inputs (e.g., alpha) does not work, you can still conduct the study, providing you adjust your expectations and make the results available. If you do make your results available, while your sample size may be too small in isolation to detect your smallest effect size of interest, your results can then be collated into meta-analyses providing they are available to other researchers.

Alternative solutions include studying larger effect sizes, and/or focusing on 'team science'. Cohen (1973) quipped that instead of chasing smaller effects, psychology should emulate older sciences by creating larger effects through stronger manipulations or using more precise measurements. Alternatively, if you cannot conduct an informative study individually, you could pool resources and engage in team science. For example,

student dissertations can benefit from projects where multiple students work together, with each student contributing one component and collecting data for a larger network/project (Creaven et al., 2021; Wagge et al., 2019), or labs across the world pool their resources together (Moshontz et al., 2018). In the past, students have been encouraged to work on a project by themselves to gain experience conducting a science experiment by themselves. However, encouraging students to work in groups may now be just as useful, as such group work reflects the paradigm shift towards 'team science' seen in the wider research community in recent years (Wuchty et al., 2007).

To conclude our tutorial, we present a list of resources you can refer to for additional applications of power analysis. We limited our tutorial to two independent samples and two paired samples for maximum accessibility, so it is important to outline resources for additional designs.

G*Power

Although G*Power (Faul et al., 2009) is no longer in active development, it supports power analysis for a range of statistical tests, such as correlation, nonparametric tests, and ANOVA. There is a longer companion guide to this manuscript that walks through power analysis for correlation and ANOVA (Bartlett, 2021).

Superpower

G*Power can calculate power for ANOVA models, but it does not accurately scale to factorial ANOVA and pairwise comparisons. To target these limitations, Lakens and Caldwell (2021) developed an R package and Shiny app called Superpower. Superpower provides more flexibility and scales up to factorial ANOVA as you enter means and standard deviations per cell for your smallest effect sizes of interest. For guidance, see our companion guide for the Shiny app (Bartlett, 2021) and the authors' ebook for the R package (Caldwell et al., 2021).

pwr R package

If you use R, the pwr package (Champely et al., 2020) supports many of the same tests as G*Power such as *t*-tests, correlation, and regression. The arguments are also similar to G*power's inputs, such as omitting one of numerator and denominator degrees of freedom, effect size as f^2, alpha, or power for your desired output.

Simulation

Packages like pwr are user-friendly as they only require you to define inputs to calculate power analytically, but one of the benefits of a programme like R is the flexibility to simulate your own bespoke power analysis. The starting point is simulating a dataset with known attributes - like the mean and standard deviation of each group or correlation between variables - and applying your statistical test. You then repeat this simulation process many times and store the *p*-values from each iteration. As probability in frequentist statistics relates to long-run frequencies, you calculate what percentage of those *p*-values were lower than your alpha, providing your statistical power. See Quandt (2020) and Sleegers (2021) for demonstrations of simulation applied to power analysis in R and the summer school workshop series organised by PsyPAG (https://simsummerschool.github.io/).

Simulation approaches also scale to more advanced techniques such as accounting for the number of trials in a task instead of solely the number of participants (Baker et al., 2021) or mixed-effects models, which are growing in popularity in psychology. Power analysis procedures for mixed effects models rely on simulation, so see Brysbaert and Stevens (2018), DeBruine and Barr (2021), and Kumle et al. (2021) for guidance.

Author Contact

ORCiD JEB: 0000-0002-4191-5245; Email JEB: james.bartlett@glasgow.ac.uk; ORCiD SJC: 0000-0002-3559-1141; Email SJC: sarah.charles@kcl.ac.uk.

Conflict of Interest and Funding

We have no conflicts of interest to disclose. Writing this article was not supported by any funding sources.

Author Contributions

Conceptualization (JEB, SJC); Writing - Original Draft (JEB, SJC); Writing - Review & Editing (JEB; SJC); Visualization (JEB; SJC). JEB placed first due to original conceptualization, but fully shared authorship between JEB and SJC.

Open Science Practices

This article is theoretical and as such provides no new data or materials, and was not pre-registered. The entire editorial process, including the open reviews, is published in the online supplement.

References

Abelson, R. P. (1985). A variance explanation paradox: When a little is a lot. *Psychological Bulletin*, *97*(1), 129–133. https://doi.org/10.1037/0033-2909.97.1.129

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

Baker, D. H., Vilidaite, G., Lygo, F. A., Smith, A. K., Flack, T. R., Gouws, A. D., & Andrews, T. J. (2021). Power contours: Optimising sample size and precision in experimental psychology and human neuroscience. *Psychological Methods*, *26*(3), 295–314. https://doi.org/http://dx.doi.org/10.1037/met0000337

Bakker, M., Hartgerink, C. H. J., Wicherts, J. M., & van der Maas, H. L. J. (2016). Researchers' Intuitions About Power in Psychological Research. *Psychological Science*, *27*(8), 1069–1077. https://doi.org/10.1177/0956797616647519

Bakker, M., Veldkamp, C. L. S., Akker, O. R. v. d., Assen, M. A. L. M. v., Crompvoets, E., Ong, H. H., & Wicherts, J. M. (2020). Recommendations in pre-registrations and internal review board proposals promote formal power analyses but do not increase sample size. *PLoS ONE*, *15*(7), e0236079. https://doi.org/10.1371/journal.pone.0236079

Bartlett, J. E. (2021). *Introduction to Power Analysis: A Guide to G*Power, jamovi, and Superpower*. https://osf.io/zqphw/

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., Cesarini, D., Chambers, C. D., Clyde, M., Cook, T. D., De Boeck, P., Dienes, Z., Dreber, A., Easwaran, K., Efferson, C., … Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*(1), 6–10. https://doi.org/10.1038/s41562-017-0189-z

Beribisky, N., Davidson, H., & Cribbie, R. A. (2019). Exploring perceptions of meaningfulness in visual representations of bivariate relationships. *PeerJ*, *7*, e6853. https://doi.org/10.7717/peerj.6853

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), 9. https://doi.org/10.5334/joc.10

Bürkner, P.-C., & Vuorre, M. (2019). Ordinal Regression Models in Psychology: A Tutorial. *Advances in Methods and Practices in Psychological Science*, *2*(1), 77–101. https://doi.org/10.1177/2515245918823199

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

Caldwell, A. R., Lakens, D., & Parlett-Pelleriti, C. M. (2021). *Power Analysis with Superpower*. Retrieved November 23, 2021, from https://aaroncaldwell.us/SuperpowerBook/

Champely, S., Ekstrom, C., Dalgaard, P., Gill, J., Weibelzahl, S., Anandkumar, A., Ford, C., Volcic, R., & Rosario, H. D. (2020). Pwr: Basic Functions for Power Analysis. Retrieved November 23, 2021, from https://CRAN.R-project.org/package=pwr

Chan, M.-P. S., Jones, C. R., Hall Jamieson, K., & Albarracín, D. (2017). Debunking: A Meta-Analysis of the Psychological Efficacy of Messages Countering Misinformation. *Psychological Science*, *28*(11), 1531–1546. https://doi.org/10.1177/0956797617714579

Chen, L.-T., & Liu, L. (2019). Content Analysis of Statistical Power in Educational Technology Research: Sample Size Matters. *International Journal of Technology in Teaching and Learning*, *15*(1), 49–75. Retrieved July 8, 2021, from https://eric.ed.gov/?id=EJ1276088

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

Cohen, J. (1965). Some Statistical Issues in Psychological Research. In W. Benjamin B (Ed.), *Handbook of clinical psychology*. McGraw-Hill.

Cohen, J. (1973). Statistical Power Analysis and Research Results. *American Educational Research Journal*, *10*(3), 225–229. https://doi.org/10.2307/1161884

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

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

Collins, E., & Watt, R. (2021). Using and Understanding Power in Psychological Research: A Survey Study. *Collabra: Psychology*, *7*(1), 28250. https://doi.org/10.1525/collabra.28250

Cramer, A. O. J., van Ravenzwaaij, D., Matzke, D., Steingroever, H., Wetzels, R., Grasman, R. P. P. P., Waldorp, L. J., & Wagenmakers, E.-J. (2016). Hidden multiplicity in exploratory multiway ANOVA: Prevalence and remedies. *Psychonomic Bulletin & Review*, *23*(2), 640–647. https://doi.org/10.3758/s13423-015-0913-5

Creaven, A.-M., Button, K., Woods, H., & Nordmann, E. (2021). Maximising the educational and research value of the undergraduate dissertation in psychology. Retrieved April 4, 2022, from https://psyarxiv.com/deh93/

DeBruine, L. M., & Barr, D. J. (2021). Understanding Mixed-Effects Models Through Data Simulation. *Advances in Methods and Practices in Psychological Science*, *4*(1), 2515245920965119. https://doi.org/10.1177/2515245920965119

Dwan, K., Altman, D. G., Arnaiz, J. A., Bloom, J., Chan, A.-W., Cronin, E., Decullier, E., Easterbrook, P. J., Elm, E. V., Gamble, C., Ghersi, D., Ioannidis, J. P. A., Simes, J., & Williamson, P. R. (2008). Systematic Review of the Empirical Evidence of Study Publication Bias and Outcome Reporting Bias. *PLoS ONE*, *3*(8), e3081. https://doi.org/10.1371/journal.pone.0003081

Etz, A., & Vandekerckhove, J. (2016). A Bayesian Perspective on the Reproducibility Project: Psychology (D. Marinazzo, Ed.). *PLoS ONE*, *11*(2), 1–12. https://doi.org/10.1371/journal.pone.0149794

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

Fisher, R. A. (1926). The Arrangement of Field Experiments. *Journal of the Ministry of Agriculture*, *33*, 503–515.

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*. https://doi.org/10.1126/science.1255484

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences*, *102*, 74–78. https://doi.org/10.1016/j.paid.2016.06.069

Goodman, S. (2008). A dirty dozen: Twelve p-value misconceptions. *Seminars in Hematology*, *45*(3), 135–140. https://doi.org/10.1053/j.seminhematol.2008.04.003

Guo, Q., Thabane, L., Hall, G., McKinnon, M., Goeree, R., & Pullenayegum, E. (2014). A systematic review of the reporting of sample size calculations and corresponding data components in observational functional magnetic resonance imaging studies. *NeuroImage*, *86*, 172–181. https://doi.org/10.1016/j.neuroimage.2013.08.012

Irving, D., Clark, R. W. A., Lewandowsky, S., & Allen, P. J. (2022). Correcting statistical misinformation about scientific findings in the media: Causation versus correlation. *Journal of Experimental Psychology. Applied*. https://doi.org/10.1037/xap0000408

Kathawalla, U.-K., Silverstein, P., & Syed, M. (2021). Easing Into Open Science: A Guide for Graduate Students and Their Advisors. *Collabra: Psychology*, *7*(18684). https://doi.org/10.1525/collabra.18684

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

Kruschke, J. K., & Liddell, T. M. (2018). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*, *25*(1), 178–206. https://doi.org/10.3758/s13423-016-1221-4

Kumle, L., Võ, M. L.-H., & Draschkow, D. (2021). Estimating power in (generalized) linear mixed models: An open introduction and tutorial in R. *Behavior Research Methods*, *53*(6), 2528–2543. https://doi.org/10.3758/s13428-021-01546-0

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

Lakens, D. (2021). The Practical Alternative to the p Value Is the Correctly Used p Value. *Perspectives on Psychological Science*, *16*(3), 639–648. https://doi.org/10.1177/1745691620958012

Lakens, D. (2022). Sample Size Justification. *Collabra: Psychology*, *8*(1), 33267. https://doi.org/10.1525/collabra.33267

Lakens, D. (2016a). One-sided tests: Efficient and Underused. Retrieved March 29, 2022, from http://daniellakens.blogspot.com/2016/03/one-sided-tests-efficient-and-underused.html

Lakens, D. (2016b). Why Within-Subject Designs Require Fewer Participants than Between-Subject Designs. Retrieved November 21, 2021, from http://daniellakens.blogspot.com/2016/11/why-within-subject-designs-require-less.html

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., Baguley, T., Becker,

R. B., Benning, S. D., Bradford, D. E., Buchanan, E. M., Caldwell, A. R., Van Calster, B., Carlsson, R., Chen, S.-C., Chung, B., Colling, L. J., Collins, G. S., Crook, Z., ... Zwaan, R. A. (2018). Justify your alpha. *Nature Human Behaviour*, *2*(3), 168–171. https://doi.org/10.1038/s41562-018-0311-x

Lakens, D., & Caldwell, A. R. (2021). Simulation-Based Power Analysis for Factorial Analysis of Variance Designs. *Advances in Methods and Practices in Psychological Science*, *4*(1), 2515245920951503. https://doi.org/10.1177/2515245920951503

Larson, M. J., & Carbine, K. A. (2017). Sample size calculations in human electrophysiology (EEG and ERP) studies: A systematic review and recommendations for increased rigor. *International Journal of Psychophysiology*, *111*, 33–41. https://doi.org/10.1016/j.ijpsycho.2016.06.015

Maier, M., & Lakens, D. (2021). Justify Your Alpha: A Primer on Two Practical Approaches. Retrieved June 24, 2021, from https://psyarxiv.com/ts4r6/

Moshontz, H., Campbell, L., Ebersole, C. R., IJzerman, H., Urry, H. L., Forscher, P. S., Grahe, J. E., McCarthy, R. J., Musser, E. D., Antfolk, J., Castille, C. M., Evans, T. R., Fiedler, S., Flake, J. K., Forero, D. A., Janssen, S. M. J., Keene, J. R., Protzko, J., Aczel, B., ... Chartier, C. R. (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

Neyman, J. (1977). Frequentist Probability and Frequentist Statistics. *Synthese*, *36*(1), 97–131. http://www.jstor.org/stable/20115217

Panzarella, E., Beribisky, N., & Cribbie, R. A. (2021). Denouncing the use of field-specific effect size distributions to inform magnitude. *PeerJ*, *9*, e11383. https://doi.org/10.7717/peerj.11383

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

Quandt, J. (2020). Power Analysis by Data Simulation in R - Part I. Retrieved April 1, 2022, from https://julianquandt.com/post/power-analysis-by-data-simulation-in-r-part-i/

Ruxton, G. D., & Neuhäuser, M. (2010). When should we use one-tailed hypothesis testing? *Methods in Ecology and Evolution*, *1*(2), 114–117. https://doi.org/10.1111/j.2041-210X.2010.00014.x

Sedlmeier, P., & Gigerenzer, G. (1989). Do Studies of Statistical Power Have an Effect on the Power of Studies? *Psychological Bulletin*, *105*(2), 309–316.

Sestir, M. A., Kennedy, L. A., Peszka, J. J., & Bartley, J. G. (2021). New Statistics, Old Schools: An Overview of Current Introductory Undergraduate and Graduate Statistics Pedagogy Practices. *Teaching of Psychology*, 00986283211030616. https://doi.org/10.1177/00986283211030616

Sleegers, W. (2021). Simulation-based power analyses. https://willemsleegers.com/content/posts/9-simulation-based-power-analyses/simulation-based-power-analyses.html

Szucs, D., & Ioannidis, J. P. A. (2021). Correction: Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLOS Biology*, *19*(3), e3001151. https://doi.org/10.1371/journal.pbio.3001151

TARG Meta-Research Group. (2020). Statistics education in undergraduate psychology: A survey of UK course content. https://doi.org/10.31234/osf.io/jv8x3

The jamovi Project. (2021). Jamovi. https://www.jamovi.org

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology*, *100*(3), 426–432. https://doi.org/10.1037/a0022790

Wagge, J. R., Brandt, M. J., Lazarevic, L. B., Legate, N., Christopherson, C., Wiggins, B., & Grahe, J. E. (2019). Publishing Research With Undergraduate Students via Replication Work: The Collaborative Replications and Education Project. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.00247

Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The Increasing Dominance of Teams in Production of Knowledge. *Science*, *316*(5827), 1036–1039. https://doi.org/10.1126/science.1136099