Power to Detect What? Considerations for Planning and Evaluating Sample Size

August 6, 2019

Roger Giner-Sorolla
School of Psychology, University of Kent

Christopher L Aberson
Department of Psychology, Humboldt State University

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

Tom Carpenter
Department of Psychology, Seattle Pacific University

Beverly G. Conrique
Department of Psychology, University of Pittsburgh

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

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

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

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

Alexander M. Schoemann
Department of Psychology, East Carolina University

Courtney Soderberg
Center for Open Science

Abstract

In the wake of the replication crisis, psychologists have begun devoting increased attention to the statistical power of psychological studies, and to other means of determining and evaluating their sample size. Unfortunately, there seems to be some misunderstanding about (a) what statistical power is, (b) how it should be evaluated, and (c) how researchers should think about sample size when designing their own studies. In this article, we ask readers to reflect on a central question - what exactly is it that a study is supposed to be "powered" to detect? With that question in mind, we discuss common misconceptions about statistical power, as well as alternatives based on inferential outcomes other than statistical significance. We also make concrete recommendations that researchers, reviewers, and journal editors can incorporate into their respective practices to move the field toward more sound statistical inferences.

Power to Detect What? Considerations for Planning and Evaluating Sample Size

The recent movement toward reform in psychological research has renewed interest in how studies' sample size is determined. For a number of reasons, small sample sizes jeopardize statistical conclusions, and higher sample size predicts whether findings replicate (Open Science Collaboration, 2015). In this article we discuss various methods of sample size determination, both when planning and when evaluating research. Because most published research in psychology uses null hypothesis testing, we focus mainly on power analysis. We offer recommendations for its use, and address four common misunderstandings in the course of our discussion. One major theme is that analyses are never high or low powered by themselves, but always critically depend on design and expected effect size. In offering guidelines to evaluate power for prospective and completed research, we emphasize the importance of ***what*** the analysis is powered to detect, and how to determine this in a principled and well-informed way.

Fundamental Concepts in Power Analysis

Power derives from the Neyman-Pearson approach to statistical testing (Pearson, 1933). High power reduces the "false negative" error rate (β), which represents the chance to fail to detect an effect as significant, conditional on the null hypothesis being false (and the alternative hypothesis being true). Conversely, statistical power is defined as (1 – β). (Cohen, 1988, 1992). In Cohen's writings, and in most current psychology research, an ideal level of power is conventionally recommended to be 80%, yielding a false negative error rate (β) of only 20%. This admittedly arbitrary convention stems from Neyman and Pearson's (1933) claim that false positives were four times worse for science than false negatives; hence, "false positives" (α) should be kept at 5% whereas "false negatives" (β) can approach 20%.

Beyond basics, we feel it is important to begin by addressing two prevailing misunderstandings about power analysis.

Misunderstanding #1: "Power is just a matter of numbers of participants." Power is actually also tied to design. A small-*N* study with many within-subjects data points can have far greater power than a large-*N* between-subjects study. Given this, sample size evaluation should consider the number of *data points*, not just subjects. Further, power functions are nonlinear (Cohen, 1988) and analysis-specific. Therefore, intuitions and heuristics based on the number of participants overall, or per design cell (e.g. van Voorhis & Morgan, 2007), will not be as accurate as actual power analysis.

Power also cannot be evaluated separately from effect size, which determines *what* a study has power to detect. All studies have excellent power to detect *some* (large) effect, and low power to detect *some* (small) effect. The same $N = 100$ test has 99% power to detect a correlation effect size of $\rho = .50$ yet only 52% power to detect $\rho = .10$. Later in the section, "Determining Target Effect Size," we will explain how to make principled decisions about expected effect size.

Misunderstanding #2: "Power is only important for controlling false negative rates." It is widely understood that running adequately-powered studies controls the risk of false negative conclusions. However, power also improves the replicability of *positive* results (Szucs & Ioannidis, 2017). If the truth about a null hypothesis is unknown, then given that "false positives" naturally occur, it is critical that "true positives" also be detected, ideally in far greater quantities. When power is low, any significant results obtained are more likely to be false positives, owing to the low probability of "true" positives. We explain these concepts later in more detail, in the section "Using Power in Evaluating Existing Research."

Types of Power Analyses

Four parameters define power analysis: α *level*, *population effect size*, *sample size*, and *power*. When three are known, the fourth can be determined.

Because α is usually fixed by convention, common types of power analyses specify two inputs and an output. The examples below all use α = .05, two-tailed, a common criterion in psychology. Researchers can improve power by committing to a one-tailed test, although this requires they restrict all inference to effects in the predicted direction. In the case of pre-registered confirmatory analyses, however, one-tailed testing may be useful (Nosek, Ebersole, DeHaven, & Mellor, 2018).

- An *a priori* power analysis (Cohen, 1988) inputs the desired power and the effect size for which power is desired. It returns a target sample size, ideal for determining sample size ahead of time. For example, to detect an effect size $\rho = .40$ with 80% power, *a priori* power analysis returns $N = 44$ observations.
- An *effect-size sensitivity* analysis inputs the desired power and likely (or achieved) usable sample size. It returns the minimum effect size detectable at this power. For example, with 100 observations and 80% desired power, effect-size sensitivity analysis returns that this power will be achieved for correlations $\rho \geq .27$. This is useful after data are collected, to assess the adequacy of the achieved sample.
- A *power-determination* analysis (sometimes referred to as "post-hoc power" in tools such as G*Power) inputs *n*, effect size, and α, and outputs power. Assume $n = 100$ can be collected; what is the power to detect a correlation of $\rho = .10$? $\rho = .30$? A power-

determination analysis reveals that power is terrible (16%) in the first case but excellent (87%) in the latter.

Misunderstanding #3: "The observed effect size from the current sample can be a basis for power analysis." Power-determination analysis is *not* correctly used with the present sample's observed effect sizes (Cohen, 1988; Gelman, 2019). It needs to assess the study against a previously-known or hypothetical effect size. Unfortunately, some statistical packages wrongly calculate power determination analyses using *sample-observed* effect size (sometimes referred to as *observed power* or *post-hoc* power). This kind of power estimate is uninformative, because it is a monotonic function of the *p*-value (see Goodman & Berlin, 1994).

In practice, all three kinds of power analysis may be run, even before a study begins. A researcher might begin *a priori,* determining sample size necessary to detect the effect size of interest (e.g., $N = 788$ for 80% power to detect $d = 0.20$ in an independent samples *t*-test). Using this example, they might then realize that $N = 788$ is unrealistic, because 500 is the maximum sample size achievable with current funding. Power-determination analysis then reveals that only 61% power is achieved for $N = 500$ and $d = 0.20$. Begrudgingly, the researcher gives up on detecting such small effects, but what effect sizes *can* be detected with $n = 500$? An effect-size sensitivity analysis reveals that $d = 0.25$ can be detected with 80% power. The researcher might decide this is close enough to the intended effect size, and proceed to gather 500 participants. In summary, the different kinds of power analysis can reinforce each other when planning research.

Alternatives to Power Analysis in Sample Size Determination

Precision Analysis

Sometimes researchers will want to do more than reject the null hypothesis via power analysis. For example, they may be confident that an effect is not zero and, instead, focus on estimating its size. For situations like these, sample size planning should be based on precision rather than power.

Precision in data means that the confidence interval (CI) for the effect is narrow. The confidence interval gives a range of effect size values around a mean, based on standard error, that is likely at a certain level (1-α; most commonly, 95%) to contain the result of a re-sampling of the study, assuming its observed parameters are true in the population. It becomes narrower as the sample size increases, but wider if the desired confidence level increases.

For some tests such as correlations, specific guidelines for precision are available. Schönbrodt and Perugini (2013) found that when $N > 250$, the width of CIs for correlations stabilized and increasing sample size did not appreciably decrease the width of CIs. The Accuracy in Parameter Estimation (AIPE) approach, however, is an approach to precision that can be used with many different statistical tests (Maxwell, Kelley & Rausch, 2008), alone or in conjunction with power analysis.

Sample size planning with AIPE aims to reach a pre-specified width of the confidence interval around a parameter. Unlike power, this width can vary separately from the size of the effect. Maxwell et al. (2008) provide an example of a study comparing two means with $d = 0.50$; a sample size of 128 (64 per group) provides 80% power, but that sample size results in a predicted 95% confidence interval ranging from .15 to .85. Neither will a study with narrow confidence intervals necessarily have high power. For example, when comparing two means with

d = .05, a sample size of 342 results in a predicted 95% confidence interval of -.10 to .20 but only 9.5% power.

AIPE requires deciding when a confidence interval is sufficiently narrow to be desirable, analogous to selecting an effect size in power analysis. A researcher should consider factors such as the maturity of the research area, and the need for a practically useful range, to select a confidence interval. Note, however, that the number of cases required for AIPE is often much greater than in *a priori* power analysis.

Optional Stopping (Sequential Design)

Traditionally, a researcher specifies one sample size *a priori*. But uncertainty about the population effect size could lead to such a study being underpowered and missing effects in the population, or being overpowered and needlessly exhausting resources. To balance power and feasibility concerns, a number of ***optional stopping*** techniques let researchers make data-dependent changes to their sample size while correcting for an increased false positive rate. Specifically, participants are collected in "waves." Between waves, an interim decision is made: whether to continue collecting data or to stop, based on the significance test and/or the achieved *N*. This method, importantly, is not the same as *undisclosed* optional stopping without controlling for false positive inflation, which has been criticized as a practice leading to low replicability in psychology (Simmons, Nelson & Simonsohn, 2011).

Optional stopping techniques do have drawbacks. Sample sizes from studies stopped early will be smaller, and so effect-size estimates will be less precise. Also, studies that stop early will have some degree of effect-size inflation, because only larger effect sizes will pass the lower significance bounds with the smaller samples of early interim analyses. There are methods to correct for inflation (see Lakens, 2016 for calculations), and we suggest that researchers report

the corrected effect size when using these designs. Another potential downside of some kinds of sequential analyses is that, if their maximum *N* is reached, they are somewhat less powerful than a traditional design, because their significance criterion is more stringent.

Planning Future Research

Whether precision or error control is the goal, analyses can play an important part in planning a study. More than sample size, a priori analyses can also test the relative power of different designs, such as within- versus between- subjects, or the number of levels in a proposed manipulation. Systematically planning these aspects of research lowers the risk of failure, whether defined in terms of precision (coming to a conclusion far from the truth) or power ("missing" an effect that exists in the population). To control these risks, *a priori* power analysis has become required in recent decades by many funders, and by ethical bodies charged with determining whether research is worthwhile (Vollmer & Howard, 2010). However (as savvy applicants know) such analyses can deliver seemingly high-powered results, if a suitably optimistic effect size is input (Maxwell & Kelley, 2011). Without taking a principled approach to effect sizes and other decisions, power analysis' usefulness will be lost.

Determining the Power Criterion

We suggest that in psychology, the 80% power suggested by Cohen (e.g., 1988), and generally used, is a bare minimum, implying that a false positive is four times more important to avoid (5% risk given H0) than a false negative (20% given H1). However, 90% represents a more rigorous standard, and 95% -- exactly balancing false-negatives with false-positives -- a strong ideal. At 99% power or above, there may be a case to adopt a precision approach, because rejecting the null is almost a foregone conclusion. Just as it is important to justify the trade-offs

involved in adopting any particular value of alpha (Lakens, Adolfi, et al., 2018), so with any particular power criterion. Perhaps the best practice is to report results for multiple reasonable power levels (e.g., 80, 90, and 95%).

Determining the Target Effect Size

Effect size precedent. When using previously known effect sizes as a guide, one might first look for the most specific precedent available for the planned analyses. For completely novel research, one might assume that the effect size will be similar to what studies in the relevant sub-discipline of psychology typically find. For example, Richard, Bond, and Stokes-Zoota (2003) conducted a meta-analysis of meta-analyses to estimate the average effect size in social psychology, drawing on over 25,000 studies. The average effect size across all of those studies was $r = .21$ with a standard deviation of .15. So in the absence of any other information, social psychologists designing new studies could assume that the effect size will be about $r = .21$. Similarly, personality psychologists could assume that their effect falls close to $r = .19$, as the appropriate field-wide median estimate suggests (Gignac & Szodorai, 2016).

Greater focus can help researchers choose more accurately. Within Richard et al.'s (2003) $r = .21$ average, there was substantial variability across topics (and even within topics). For example, studies in group processes tended to produce larger ($r = .32$) effects than those in social influence ($r = .13$). Therefore, knowing the research's general topic area might improve the estimate. Research topics can be defined even more precisely, such as correspondence bias or cognitive therapy effectiveness, and the appropriate meta-analytic estimate used. In the most focused case, researchers performing a direct replication of a given study often use its reported effect size as a ready target (Brandt et al., 2014).

It is underappreciated that a study's methodological paradigm, not just topic, can influence effect sizes. Meta-analyses of paradigms cutting across multiple substantive topics, unfortunately, are few. Still, consider studies of interracial threat. At the subtle end of the spectrum, one can manipulate minimal features of vignettes, such as when the United States will become a "majority-minority" nation (Craig & Richeson, 2014). That subtle manipulation should produce a smaller effect size, all else equal, than a more vivid experiment in which, for example, White male undergraduates and a Black male are paired to chat about racial profiling (Goff, Steele, & Davies, 2008). Thinking about the tasks and measures used in previous studies, then, can also assist informed choices about effect size, although for now more in an impressionistic than a precise way.

Publication bias importantly limits all literature-based approaches. Most meta-analytic estimates come from published studies, plus some unpublished data collected *ad hoc*. Even single studies for replication have usually been published in an environment where non-significant results do not see the light of day. We never know how many other, less impressive study results live in a file drawer, biasing our estimates of the true effect size. Because significant, hence larger, results are more likely to be published, literature-based estimates are often too large (Dickersin, 1990). However, how biased a particular estimate is, and thus how much to adjust it, can be difficult to infer, though some heuristics and methods have been proposed (see Lewis & Michalak, 2019; Simonsohn, 2015).

Smallest Effect Size of Interest. If reluctant to rely on the literature, one could anchor power analysis instead on the smallest effect size of interest (SESOI) for the research team (Lakens, Scheel, & Isager, 2018). The challenge lies in actually determining the SESOI. Research teams that study more tangible outcome measures may have it easier. For instance, an

education researcher might design a study to detect whether an intervention changes grade point averages (GPA) by at least 0.25 units on a 4.0 grading scale because, in their view, a lesser effect would not be worth investing large amounts of resources disseminating and scaling the intervention. For cheaper interventions, the SESOI might be lower.

Without clear benchmarks, it can be more difficult for research teams to determine their SESOI. Are interventions that produce $d = .1$ changes on 7-point Likert scales important to study, or does d need to be at least .2? Scholars have suggested various criteria for how appreciable an effect size is. These include: societal importance (e.g., lives saved; Rosenthal, 1990); perceived difficulty in affecting an outcome variable (Prentice & Miller, 1992); whether multiple small effects might build on each other to have large effects on an outcome (Abelson, 1985); and how the presence of multiple predictors can limit mathematically the potency (i.e., effect size) of any one predictor (Ahadi & Diener, 1989; Strube, 1991). Though they are rarely exact, these criteria may help researchers who are planning new studies decide what range of effect sizes they would consider worthwhile.

For basic research topics that are primarily meant to inspire further research studies, it may also be reasonable to assume a maximum resource level, and use sensitivity analysis to consider what effect sizes are clearly outside the grasp of a researcher using that paradigm. For example, in a study using a half-hour lab session, one might judge it hard to justify running any more than 100 individuals. In that case, effects not detectable at our minimal power criterion of .80 (here, $r < .24$) would need special justification to serve as a SESOI. In areas of research where the meaning of effect sizes is not well understood, a pragmatically-rooted approach would start from method, leaving it to the reader to interpret the sensitivity of that method.

Power Analysis Techniques

As the importance of power analysis has grown, tools for conducting it have flourished. Where algorithmic (that is, analytic) approaches would be too difficult to compute, Monte Carlo power determination has gained popularity. In Monte Carlo techniques, many random simulations are run to assess the probability of observing significant outcomes given an underlying effect of a certain size. These techniques may look difficult because they require the input of many parameters, such as means, standard deviations, and/or correlations between variables. However, the first two parameters can be handled by assuming standardized data (standard deviation = 1) and expressing mean differences in terms of these standard units (that is, to represent a medium effect size d = 0.5, or half a standard deviation, you might input one mean as -0.25 and the other as 0. 25). Assumed correlations among variables can be handled by looking at what is typical in similar research, or by putting in a variety of plausible correlations and seeing how they affect the result.

Although a full review of power techniques is too detailed for the present treatment, we have made available a critical review of tools for commonly used analyses in psychology as an online preprint (Aberson, Bostyn … & Soderberg, 2019). We take the freely available and highly cited software, G*Power (Faul, Erdfelder, Lang, & Buchner, 2007) as a reference point. Table 1 summarizes our review, outlining where G*Power gives a good answer, where it needs special considerations as of this writing, and where other resources need to be consulted. Citations for R packages and online resources are listed in the Appendix.

To briefly explain some special considerations when using G*Power for regression and ANOVA (see preprint for details):

- Regression: When planning to interpret more than one regression coefficient in the same analysis, G*Power's estimates do not consider correlations among independent variables. It is recommended instead to use the R package, pwr2ppl.
- ANOVA: In all ANOVA applications, G*Power uses the biased effect size estimate of partial eta-squared (η^2), but the unbiased estimates ω^2 or ε^2 are preferable (Lakens, 2013, 2015). MOTE and ANOVApower R packages allow calculation from unbiased effect sizes.
- Repeated measures ANOVA: This procedure is currently not documented in G*Power. It is important for users to do two non-obvious things: select the option "effect sizes as in SPSS," and enter "number of measures" in factorial repeated designs as the numerator degrees of freedom of the ANOVA plus one (not the total number of measures in the design), or power will be greatly over-estimated.
- Factorial ANOVA: G*Power may give startlingly low sample size recommendations for factorial designs, which are technically correct, but with caveats. First, if part of your argument rests on simple effects or multiple comparisons, you should base your power on cell size for those tests, not just the overall ANOVA (Giner-Sorolla, 2018). Second, effect sizes for two-way and higher interactions are usually smaller than the effect size of the simple effects they are based on, by a factor of 0.5 or less (Simonsohn, 2014; Westfall, 2015a, 2015b). Only when there is "cross-over", such that (in a 2x2 design) one simple effect is the reverse of the other, do interaction effect sizes approach simple effect sizes.

Table 1. Summary of Specific Power Analysis Methods Recommendations from Supplementary Article

Technique	G*Power OK as is?	G*Power considerations	Other resources
Correlation, chi square, t-tests	Yes		R package: pwr
Multiple regression: model and change tests	Yes		R package: pwr2ppl
Multiple regression: multiple single coefficients	No	Need to know correlations among IV	R package: pwr2ppl
ANOVA: general	No	Use unbiased effect sizes, ω^2 or ε^2	MOTE, ANOVApower R packages; MOTE app
ANOVA: Repeated measures & mixed	No	1. Multicollinearity is double-counted; use effect size "as in SPSS" 2. "Number of measurements" input unclear in factorial RM, use num. df +1	1. GLIMMPSE, online 2. PANGEA app, online 3. R package: ANOVApower
ANOVA: Factorial	Yes, but	1. Interactions often involve lower effect sizes than main / simple effects 2. Power also needs determining for comparisons and simple effects	R package: ANOVApower
Mediation	No	Not available	Various Monte Carlo, see Appendix
SEM	No	Not available	Various; see Appendix
Multilevel	No	Not available	Various; see Appendix

Reporting Power Analyses

Current writing guides (e.g. JARS, Appelbaum, Cooper, Kline, Mayo-Wilson, Nezu, & Rao, (2018); APA Publication Manual, 2010) leave unclear how power considerations should be reported in manuscripts. We offer recommendations here.

If sample size was decided *a priori* via power analysis, make sure to report the software used, effect size (with units, e.g. d, f^2), rationale for effect size, target power (including, as we have suggested, values for 80%, 90% and 95% power), and any other parameters used in that analysis. Full reporting of parameters and decisions is also recommended for precision and sequential analysis choices. As noted already, "observed power" should not be reported, as it conflates the sample and population effect sizes.

However, *a priori* power is less appropriate to report if it was not used to determine sample size. Often, sample size is decided by resource availability, rules of thumb, or emulation of prior sample sizes. In such cases, an effect-size sensitivity analysis is the most useful and honest tool (Cohen, 1988). Even when sample size is planned, missing or incomplete responses may reduce the amount of *usable* data below original intent, reducing achieved power and again making effect-size sensitivity analysis necessary. For example, an author who planned for 300 participants and only could keep 243 might state that the final analysis had "80% power to detect effects as small as …".

Power within a single study may vary, if multiple statistical analyses use different tests and designs. To acknowledge this point, we suggest that power *analyses* (plural) be described in the Results section close to each type of analysis (following Sleegers, 2019). The *Participants*

section need only specify which of these analyses the overall sample size was based on – typically, the most demanding analysis in terms of numbers suggested.

Using Power in Evaluating Reported Research

Power is not just useful for research planning. Reviewers of manuscripts, and readers of published work, need to assess the value of research when it is disseminated. Power bears on this task, but many people do not have a clear idea about why and how.

Evaluating power accurately, first of all, requires full and transparent reporting of the results. A result from a study where many outcomes were analyzed, but only significant results reported, cannot be evaluated in the same way as the identical result from a single-analysis study. Multiple testing increases the likelihood of making a type I error, and selective reporting inflates the effect size estimate. The second example will lead to a more accurate representation of power and effect size. Given previously accepted practices of selective reporting, a statement that all measures, manipulations, and even relevant studies are disclosed, gives greater confidence (Simmons, Nelson, & Simonsohn, 2012). Other practices that inflate type I error, such as *undisclosed* optional stopping, reduce confidence in the study's accuracy. Conversely, other practices that ensure disclosure, such as preregistration or Registered Reports, increase confidence.

Misunderstanding #4: "Effects that are significant despite a low-powered design are clearly very large in the population." Although some may admire a "heroic" effect that survives every attempt of poor methodology to kill it, this evaluation is wrong (Loken & Gelman, 2017). A significant result only means that if the null hypothesis is true, the probability of the observed effect (or stronger) is low. But the survivor myth depends on wrongly concluding

the reverse: that if a significant effect is observed, the null hypothesis is unlikely (see Ioannidis, 2005). In fact, low power reduces the likelihood that an observed significant result reflects a true positive effect.

More concretely, if $\alpha = .05$, the fallacious assumption is that there is only a 5% risk of a false positive. Setting $\alpha = .05$ does indeed allow only 5% of null effects to appear significant. However, it does *not* restrict false positives overall to 5%. Researchers may wish to know what percent of all observed significant results are false positives, known as the False Discovery Rate (FDR; Ioannidis, 2005). An argument could then be made for attempting to restrict the FDR to 5% or some other low number. Critically, the FDR depends on both the frequency of false positives (set by alpha), true positives (determined by power), and the odds of an effect actually being true (prior odds). For example, in a study with a very low power of 10% and uninformed prior odds of 1:1, the FDR is 33%, whereas an identical study with power of 80% has FDR of 5.9%. High power to detect a given effect size means the FDR is closer to an acceptable number. Even a middling power of 40% leads to a FDR of 11.1%. Put another way, the α level needed in the 40% study to reach the same FDR as the 80% study would be .025, not .05. So, in low-powered studies, significant results are more likely to be false-positive errors, and *p*-values close to .05 are particularly untrustworthy, because they are unlikely to reach the alpha level required in an underpowered study to achieve an acceptable false discovery rate. (All calculations were facilitated by the online resources at Schönbrodt, 2019.)

The downsides of false positives are well known. Creating interventions and further studies based on false positives is a misuse of resources. Even if the underlying effect is true, underpowered studies, when subject to publication bias on the basis of significance, yield overestimates of effect sizes (e.g., Sterne, Gavaghan & Egger, 2000). Researchers might then

run studies that they think have adequate power, but are actually underpowered to detect the effect, leading to replication failure and a confused literature.

In evaluating the power of a published article, an effect-size sensitivity analysis gives the best information. If the authors do not provide it, it can be calculated using the information on *N* and design in the article, setting a desired power level. The question is how to evaluate the effect size output. As we have seen, criteria for typical and minimal effect sizes are not well established, and depend on the topic, method, and application of the research. As a tentative guide for research in psychology, we suggest that studies with less than 80% power to detect a conventionally medium-sized effect (i.e., a difference between groups $d = .50$ standard deviations, or $\rho = .30$) will require justification: by defending the assumption of a larger effect, or by invoking considerations of research difficulty (see next section). This target effect size may be modified if more is known about minimal or typical effect sizes relevant to the research.

Cautions and Solutions about Power for Difficult Research Cases

Although it may be justified to be cautious about the results of studies low in power, simply excluding such manuscripts from publication and other forms of dissemination can also limit a scientific field in unexpected ways. Because publication is a major metric of hiring, tenure, and promotion, these decisions will also impact scholars' judgments about whether to pursue a particular type of research in the first place.

A policy of rejecting low-powered research could discourage work on hard-to-reach and diverse populations. It would also perpetuate the long-standing file-drawer problem, an issue that becomes particularly pernicious for groups that are already underrepresented in the literature. This includes underserved groups, and ones that are simply more diverse or difficult to study

than typical samples from relatively affluent Western citizens (WEIRD populations; Henrich, Heine, & Norenzayan, 2010).

Conversely, standards requiring high power are most easily reached through samples such as undergraduate college students, who can be recruited relatively easily, quickly, and in large numbers. But these samples are simply not appropriate (or possible) to utilize for some research questions. Likewise, researchers have recently turned to crowdsourced participant pools online for data collection (e.g., Buhrmester, Kwang, & Gosling, 2011; Buhrmester, Talaifar, & Gosling, 2018; Paolacci & Chandler, 2014; Sassenberg & Ditrich, 2019). But online samples also are not very valid for research involving, for example, immersive face-to-face social environments or tangible behavioral outcomes (Anderson et al., 2019).

Consider the hypothetical example of a researcher interested in prejudice experiences of Asian Americans, who also wants to represent the diversity of backgrounds within this category (East Asian Americans versus Southeast Asian Americans, for example; Leong & Okazaki, 2009). Doing so could require recruiting enough participants to represent, say, five or six ethnic backgrounds, some of which might be relatively small in numbers or hard to reach.

Consider a second example of a researcher who studies population health disparities intersectionally. While existing literature has shown meaningful population health disparities between people of color and Whites in the United States, researchers are only just beginning to examine how the intersection of multiple identities (Crenshaw, 1989) may exacerbate existing health disparities (e.g., Lewis & Van Dyke, 2018). Perhaps this researcher is interested in group differences in depression, not just between Whites and people of color in the United States, but additionally how these ethnic disparities may be exacerbated in elderly populations.

In both cases, conducting research at high power to detect smallish effects would be very difficult. The investigators would need time and resources to ensure the validity of their materials; adequate participant-payment funds; and, most likely, longer-term partnerships with people in their communities to locate participants. They would be limited, critically, by the numbers of reachable participants fitting the target demographics. And eligible individuals may not want to participate, for a variety of reasons—time, general wariness, specific concerns about the research process.

Given the barriers facing such researchers, the detection of some effects is likely to be underpowered despite their most assiduous efforts. In this case, a rigid decision to reject the work based on power may do more harm than good. It would perpetuate the exclusion from research literature of hard-to-reach populations who are already severely under-represented. A file-drawer problem based on statistical power is still a file-drawer problem.

Rejection, then, should not be the only possible outcome for an underpowered but methodologically difficult study. Indeed, by definition every study has sufficient power to detect some effects, but lacks power to detect others. Editors and reviewers must consider the effects that a study is adequately powered to detect, weighing the clarity of the finding against the importance of doing research at all in the context. They might consciously adopt a different threshold for reporting research that uses difficult methods or studies difficult-to-reach populations (e.g., a different power criterion, or higher α), allowing published results to be more tentative than for questions that can be studied through large and multiple repeated studies. Of course, the authors should then be encouraged to express uncertainty in their writing, without having to oversell the findings to get published.

For less convenient methods or populations, researchers also need to plan around power issues. They may want to concentrate on larger, rather than smaller, effects (for example, in studies where a policy or health-related intervention may be involved). They can also choose methods for stronger effect size and hence power: for instance, using a more robust vs. subtle experimental manipulation, or adopting a within-subjects vs. between-subjects design. Those conducting similar work may benefit from collaborations pooling together resources and samples from many labs to maximize power (e.g., the Psychological Science Accelerator, Moshontz et al., 2018). Finally, researchers may also choose to share unpublished data through preprints, remedying distorted perceptions of effects and aiding future meta-analyses.

Conclusion

Within psychology, there has been increased recognition over the years of statistical power and related considerations (e.g., effect size, precision). Determining statistical power can be daunting, however, due to statistical and mathematical complexity and the multitude of different approaches, depending on one's research design and statistical test. To remedy this, we have provided background on statistical power and other approaches, tackled the main question of what effect sizes should be assumed ("power to detect what?"), and explained the role of power analysis in evaluating completed as well as prospective research, all with the aim to make power-related issues more clear to researchers. With our overview of specific techniques and software, we hope we have also further empowered researchers to conduct and evaluate research in line with optimal sample-size considerations.

Author contributions: The first author (RG-S) organized the Working Group and did planning, coordination, writing, and editing on the manuscript. After the first author, the remaining authors appear alphabetically, reflecting their roughly equal contributions to the writing of different sections of the paper and online supplement, as well as comments and revision.

Conflicts of Interest: The authors declare that there were no conflicts of interest with respect to the authorship or the publication of this article.

Acknowledgments: We would like to acknowledge the Society for Personality and Social Psychology's and in particular its Executive Director, Chad Rummel, who initiated a call for working groups at the 2019 meeting, approved our application, and facilitated our in-person meeting at the conference.

Supplemental Material: Posted on OSF: https://osf.io/9bt5s/, "Power Analysis Working Group supplement Aug 6 19".

Prior versions: The initially submitted version of this article has been posted as a preprint on OSF at https://osf.io/9bt5s/.

References

Abelson, R. P. (1985). A variance explanation paradox: When a little is a lot. *Psychological Bulletin, 97,* 129-133**.**

Aberson, C. L. (2019). *Applied power analysis for the behavioral sciences (2nd edition).* New York: Routledge.

Aberson, C. L., Bostyn, D. H., Carpenter, T. Conrique, B. G., Giner-Sorolla, R., Lewis, Jr., N. A., Montoya, A. K., Ng, B. W., Reifman, A., Schoemann, A. M., & Soderberg, C. (2019). Techniques and solutions for sample size determination in Psychology: A critical review. Preprint available online at https://osf.io/9bt5s/.

Ahadi, S., & Diener, E. (1989). Multiple determinants and effect size. *Journal of Personality and Social Psychology, 56*, 398-406.

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: American Psychological Association.

Anderson, C. A., Allen, J. J., Plante, C., Quigley-McBride, A., Lovett, A., & Rokkum, J. N. (2019). The MTurkification of social and personality psychology. *Personality and Social Psychology Bulletin*, *45*, 842-850.

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board Task Force report. *American Psychologist*, *73*, 3–25.

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods, 37*, 379-384.

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., ... & Van't Veer, A. (2014). The replication recipe: What makes for a convincing replication?. *Journal of Experimental Social Psychology*, *50*, 217-224.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data?. *Perspectives on Psychological Science*, *6*(1), 3-5.

Buhrmester, M. D., Talaifar, S., & Gosling, S. D. (2018). An evaluation of Amazon's Mechanical Turk, its rapid rise, and its effective use. *Perspectives on Psychological Science*, *13*(2), 149-154.

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

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

Craig, M. A., & Richeson, J. A. (2014). On the precipice of a "majority-minority" America: Perceived status threat from the racial demographic shift affects White Americans' political ideology. *Psychological Science*, *25*(6), 1189-1197.

Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: A Black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *University of Chicago Legal Forum*, *1989*(8), 139–167.=

Dickersin, K. (1990). The existence of publication bias and risk factors for its occurrence. *Jama*, *263*(10), 1385-1389.

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

Frick, R. W. (1998). A better stopping rule for conventional statistical tests. *Behavioral Research Methods, Instrucuments, & Computers, 30*, 690-697.

G*Power 3.1 Manual (March 1, 2017). Retrieved from http://www.gpower.hhu.de/fileadmin/redaktion/Fakultaeten/Mathematisch-Naturwissenschaftliche_Fakultaet/Psychologie/AAP/gpower/GPowerManual.pdf .

Gelman, A. (2019). Don't calculate post-hoc power using observed estimate of effect size. *Annals of Surgery, 269*, e9. https://doi.org/10.1097/SLA.0000000000002908

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences, 102*, 74-78.

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

Goff, P. A., Steele, C. M., & Davies, P. G. (2008). The space between us: Stereotype threat and distance in interracial contexts. *Journal of Personality and Social Psychology, 94*(1), 91.

Goodman, S. N., & Berlin, J. A. (1994). The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. *Annals of Internal Medicine*, *121*(3), 200-206.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world?. *Behavioral and Brain Sciences*, *33*(2-3), 61-83.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine, 2*(8), e124.

Konstantopoulos, S. (2010). Power analysis in two-level unbalanced designs. *The Journal of Experimental Education*, *78*(3), 291-317. Doi: 10.1080/00220970903292876

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology, 4,* 863.

Lakens, D. (2015, June 8). Why you should use omega-squared instead of eta-squared [Blog Post]. Retrieved from http://daniellakens.blogspot.com/2015/06/why-you-should-use-omega-squared.html

Lakens, D. (2016). Sequential analyses. Retrieved from https://osf.io/uygrs/.

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A., Argamon, S. E., ... & Buchanan, E. M. (2018). Justify your alpha. *Nature Human Behaviour*, *2*(3), 168.

Lakens, D. & Evers, E. R. (2014). Sailing from the seas of chaos into the corridor of stability: Practical recommendations to increase the information value of studies. *Perspectives on Psychological Science, 9*(3), 278-292.

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

Leong, F. T., & Okazaki, S. (2009). History of Asian American psychology. *Cultural Diversity and Ethnic Minority Psychology*, *15*(4), 352-362.

Lewis, N. A., Jr., & Michalak, N. M. (2019, April 8). Has stereotype threat dissipated over time? A cross-temporal meta-analysis. Preprint retrieved from https://doi.org/10.31234/osf.io/w4ta2.

Lewis, T. T., & Van Dyke, M. E. (2018). Discrimination and the health of African Americans: The potential importance of intersectionalities. *Current Directions in Psychological Science*, *27*(3), 176-182.

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science, 355*(6325), 584–585.

Maxwell, S. E., & Kelley, K. (2011). Ethics and sample size planning. *Handbook of Ethics in Quantitative Methodology*, 159-184.

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

Moshontz, H., Campbell, L., Ebersole, C. R., IJzerman, H., Urry, H. L., Forscher, P. S., ... & Castille, C. M. (2018). The Psychological Science Accelerator: Advancing psychology through a distributed collaborative network. *Advances in Methods and Practices in Psychological Science*, *1*(4), 501-515.

Neyman, J., & Pearson, E. S. (1933, October). The testing of statistical hypotheses in relation to probabilities a priori. In *Mathematical Proceedings of the Cambridge Philosophical Society* (Vol. 29, No. 4, pp. 492-510). Cambridge University Press.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences, 115*(11), 2600-2606.

Okada, K. (2013). Is omega squared less biased? A comparison of three major effect size indices in one-way ANOVA. *Behaviormetrika, 40*(2), 129-147.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science*, *23*(3), 184-188.

Prentice, D. A., & Miller, D. T. (1992). When small effects are impressive. *Psychological Bulletin, 112,* 160-164.

Richard, F. D., Bond Jr, C. F., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology*, *7*(4), 331-363.

Rosenthal, R. (1990). How are we doing in soft psychology? *American Psychologist, 45*(6), 775-777.

Sassenberg, K., & Ditrich, L. (2019). Research in social psychology changed between 2011 and 2016: Larger sample sizes, more self-report measures, and more online studies. *Advances in Methods and Practices in Psychological Science, 2*(2), 107–114.

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize?. *Journal of Research in Personality*, *47*(5), 609-612.

Schönbrodt, F. D. (2019). When does a significant p-value indicate a true effect? Understanding the Positive Predictive Value (PPV) of a p-value [Web Page]. Retrieved from http://alturl.com/k3do9

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22,* 1359–1366. https://doi.org/10.1177/0956797611417632

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2012, October 14). A 21 Word Solution. Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2160588

Simonsohn, U. (2014, March 12). No-way interaction [Blog post]. Retrieved from http://datacolada.org/17

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science, 26*(5), 559-569.

Sleegers, W. (February 25, 2019) [Twitter Post]. Retrieved from https://twitter.com/willemsleegers/status/1100087024785244161

SPSP Power Analysis Working Group (2019). Techniques and solutions for sample size determination in psychology: A critical review. Available online at **********.

Sterne, J. A., Gavaghan, D., & Egger, M. (2000). Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. *Journal of clinical epidemiology*, *53*(11), 1119-1129.

Strube, M. J. (1991). Multiple determinants and effect size: A more general method of discourse. *Journal of Personality and Social Psychology*, *61*, 1024-1027.

Szucs, D., & Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLOS Biology*, *15*, e2000797. https://doi.org/10.1371/journal.pbio.2000797

van Voorhis, C. W., & Morgan, B. L. (2007). Understanding power and rules of thumb for determining sample sizes. *Tutorials in Quantitative Methods for Psychology*, *3*(2), 43-50.

Vollmer, S. H., & Howard, G. (2010). Statistical power, the Belmont report, and the ethics of clinical trials. *Science and Engineering Ethics*, *16*(4), 675-691.

Westfall, J. (2015a, May 26). Think about total N, not n per cell [Blog post]. Retrieved from http://jakewestfall.org/blog/index.php/2015/05/26/think-about-total-n-not-n-per-cell/

Westfall, J. (2015b, May 27). Follow-up: What about Uri's 2n rule? [Blog post]. Retrieved from http://jakewestfall.org/blog/index.php/2015/05/27/follow-up-what-about-uris-2n-rule/

Appendix: Reference list of computational resources

Precision analysis

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

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

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

Sequential analysis

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

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

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

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

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

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

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

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

Basic analyses including correlation, T test, regression

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

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

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

ANOVA

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

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

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

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

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

Mediation analysis

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

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

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

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

Structural equation modeling

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

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

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

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

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

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

Multilevel / hierarchical / mixed model analysis

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

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

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

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

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