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

May 28, 2020

Roger Giner-Sorolla
School of Psychology, University of Kent

Tom Carpenter
Department of Psychology, Seattle Pacific University

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

Christopher L Aberson
Department of Psychology, Humboldt State University

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

Beverly G. Conrique
Department of Psychology, University of Pittsburgh

Brandon W. Ng
Department of Psychology, University of Richmond

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

Manuscript under review, please do not cite without permission of first/corresponding author:
rsg@kent.ac.uk

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

May 7, 2020

Abstract

In the wake of the replication crisis, social and personality psychologists have increased attention to the sample sizes and statistical power of their studies. Nonetheless, there remain misunderstandings about what statistical power is, how to evaluate it, and how researchers should think about sample size . Further, the realities of some research areas (e.g., limited resources) and goals (e.g., testing hypotheses vs. precisely estimating effects) limit the utility of generic recommendations dictating sample sizes. We address common misconceptions about power and sample-size adequacy, and highlight relevant statistical tools and approaches. We also provide concrete recommendations for improving the practices of researchers, reviewers, and journal editors in social andpersonality psychology. In doing so, we go beyond current practice byadvocating an *effect-size-sensitivity* approach to power analysis for most, but not all, situations, based on the indeterminacy of our current knowledge about effect sizes in most fields of research.

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

The recent movement toward reform in psychological research has renewed interest in studies' sample sizes and statistical power. Small sample sizes have been shown to jeopardize the accuracy of statistical conclusions, as well as the replicability of findings (Open Science Collaboration, 2015). Although researchers have various intuitions about sample sizes, the related concept of statistical power is still poorly understood. Currently, psychologists assess whether a study's sample size is adequate by relying on intuition, rules of thumb (e.g., 20 observations per cell; Simmons, Nelson, & Simonsohn, 2011), existing studies, various forms of power analyses (Cohen, 1988), and precision-based approaches (e.g., Rothman & Greenland, 2018). In this environment, researchers, reviewers, and editors may struggle to know how many observations (cases and measurements) are required to adequately answer a given research question.

Uncertainty about these topics in social and personality psychology led the authors to meet as a Power Analysis Working Group at the 2019 meeting of the Society for Personality and Social Psychology, in response to a call from Executive Director Chad Rummel. In this paper, we address common misconceptions about power and sample size adequacy, discuss different kinds of sample size determination methods, propose standards for reporting them, and summarize tools and approaches for power analysis. While not all approaches are suited for all research contexts, in many contexts calling for power analysis we advocate for an *effect-size sensitivity* approach. In this approach, researchers calculate the effect sizes for which their analyses are adequately powered, given a typical or feasible sample size, and a range of useful power levels. Because of the many difficulties in determining a likely effect size ahead of time, it

may make more sense to look at effect size as an *output* of power analysis, and let other factors weigh in when determining sample size.

Fundamental Concepts and Misunderstandings in Power Analysis

Power derives from the Neyman-Pearson approach to statistical hypothesis testing (Pearson, 1933). Statistical power is defined as 1 - β (Cohen, 1988, 1992), where β is the *false negative* error rate (the probability of failing to declare an effect as significant, if the alternative hypothesis is true). In other words, higher power means that true effects, if present, are detected more frequently. In Cohen's writings (e.g., Cohen, 1988), and in most current psychology research, the recommended level of power is conventionally 80%, yielding a false-negative error rate (β) of 20%. Based on Neyman and Pearson's (1933) observation that false positive rates are traded off against false negatives, the 80% power level, as Cohen (1988) explains, assumes that admitting false findings at the conventional 5% rate is four times worse than missing true findings. Consequently, if the rate of "false positives" assuming the null is kept at 5% (α, the criterion for statistical significance)[1], then "false negatives" assuming the alternative (β) can approach 20%.

Although many quantitative researchers are familiar with these fundamental concepts, a nuanced understanding of power is not always evident in the planning, discussion, and evaluation of quantitative research. In the sections that follow, we address several common misconceptions in power analysis and sample-size planning.

[1] Although the choice of α is increasingly seen as an analytic choice (Lakens et al., 2017) with an argument to be made for values below .05 (e.g., Benjamin et al., 2018; Nosek et al., 2018), we assume α = .05 in this paper as it remains the most commonly applied criterion.

Misunderstanding #1: "Power Analysis Can Only Determine Sample Size"

Power analysis is sometimes conflated with a specific kind of *a priori* power analysis in which sample size is determined, given a desired power level (and other features of the study). Because commonly used software (e.g., G*Power; Faul, Erdfelder, Lang, & Buchner, 2007) refers to this kind of analysis as "a priori," researchers may be tempted to presume this is the best kind of analysis that can be done before conducting a study. In fact, there are several kinds of power analysis, and all can be useful at all stages of the research process (planning, analysis, evaluation).

Four parameters define power analysis: α *level*, *population effect size, sample size,* and *power*. When three are known, the fourth can be determined, creating four distinct types of power analysis. Because α is usually fixed by convention, common types of power analyses specify two inputs and one output. The examples below all use α = .05, two-tailed[2]—a common criterion in psychology. It should be emphasized that in all three analysis types below, effect size refers to the (unknown) population effect size, not the observed effect size in the sample.

- 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 study methodology ahead of time. For example, to detect a correlation's effect size $\rho \geq .40$ with at least 80% power, *a priori* power analysis requires $N = 44$ observations.
- An *effect-size sensitivity* analysis inputs the desired power and likely (or achieved) usable sample size. It returns the minimum population effect size detectable at or above

[2] 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). Arguments also exist for adopting a more stringent criterion α (e.g., .005; Benjamin et al., 2018).

this power. For example, with 100 observations, an effect-size sensitivity analysis identifies that 80% or greater power will be achieved for correlations that have size $\rho \geq .27$ in the population.

- A *power-determination* analysis, sometimes referred to as "post-hoc power" in tools such as G*Power, inputs N and a *population* effect size, and returns power. For example, given that $N = 100$ participants have been recruited, a power-determination analysis reveals power is terrible (16%) to detect a population correlation of $\rho = .10$ and excellent (87%) to detect a slightly larger population correlation of $\rho = .30$. (This analysis type seems particularly prone to the misconception that the observed sample effect size can be useful for input; see Misunderstanding #5, below.)

Which of these analyses to use depends on the researcher's goals. If the only goal is to test an empirical hypothesis, then an *a priori* power analysis will return the minimum necessary sample size given a particular effect size. However, this approach may be inaccurate in many research contexts within social-personality psychology for which expected effect size is not easily determined. As noted in Misunderstanding #3, power is critically influenced by the (unknown) population effect size. Thus, any a priori power analysis requires that researchers input a smallest effect size they wish to detect, the value of which helps to determine sample size. A researcher who wishes to detect "small" ($d = .20$) effects must collect $N = 787$ observations, whereas a researcher who is happy to detect only slightly larger effects ($d = .35$) need only collect *one third* that sample size—$N = 258$.

We provide some guidelines for thinking about effect sizes below (Misunderstanding #6) but acknowledge that there is often not enough information available on this matter before running a study. Though many researchers may have intuitions about the kind of effect size that

is not worth bothering about (e.g., setting the lower bound at a conventionally "small" size corresponding to $d = 0.2$), theories in the field of social and personality psychology are not typically well-defined enough to precisely predict effect sizes in novel studies, or to state what a smallest effect size of interest should be (e.g., $d = 0.2$ or 0.35?). This is particularly important in sample size determination because the sample sizes required to detect $d = 0.2$ and $d = 0.35$ are very different from each other. And, as we discuss later, determining a realistic typical or minimal effect size from the literature is often uncertain.

In the absence of clear criteria for effect sizes, many researchers would be happy knowing if there existed *any* effect in a definite direction. If this is the case, then the first question to answer is "What is the smallest effect size that I can afford to detect?" while the second question follows,"...and is that effect size reasonable or important?" Researchers with limited time, money, and mental energy may be tempted to enter many different effect sizes into an a priori analysis by trial-and-error until an acceptable sample size if found.

However, this is simply an inefficient way of running an *effect-size sensitivity* analysis (Cohen, 1988). In this kind of analysis, the researcher enters a given sample size – for example, the largest sample they can afford to collect -- and the desired level of power (e.g., 80%). The analysis returns the smallest effect size that can be tested, given that sample size and design, at that level of power. Presuming that this effect size is a reasonable lower bound for the range of effects in a research area (see Misunderstanding #6), then the sample size is likely adequate. If this effect size is larger than effect sizes that would be reasonable or of interest, then the sample is not sufficiently powerful.

Using effect-size sensitivity has a number of advantages. First, it lets researchers explicitly consider sample-size criteria other than power (e.g., their resources available) when

planning research. These are often realities of research design, yet an *a priori* analysis does not allow for their consideration. Relatedly, to the degree that it is *already* the practice of researchers to find an effect size that has good power for a given feasible *N*, then reporting it as such is more transparent and accurate. Third, to focus on the largest possible sample size given one's resources may improve precision: Larger samples estimate effects more precisely and accurately (Maxwell, Kelley & Rausch, 2008), so researchers have more to say about an effect when they collect more than a minimal *N* to reject the null hypothesis (recall that rejecting the null typically means declaring an effect nonzero). Further, even if an *a priori* analysis is used, the final usable number of cases may differ from its recommendations. For example, researchers who collect some reaction-time tests routinely reject nontrivial proportions of their samples for fast, slow, or erroneous responding (see Greenwald et al., 2003). An effect-size sensitivity analysis may be needed to evaluate this revised sample size.

In practice, all three kinds of power analysis may be run in concert, even before a study begins. For example, a researcher might decide it best to be able to detect $d \geq .20$. An *a priori* analysis can be used to determine the necessary sample size (e.g., finding that $N = 787$ for 80% power to detect $d = 0.20$ with an independent-samples *t*-test). Using this example, they might then realize that $N = 787$ is unrealistic, because 500 is the maximum sample size achievable with current funding. Power-determination analysis could then be used to assess the adequacy of that sample size (e.g., revealing that only 61% power is achieved for $N = 500$ and $d = 0.20$). Begrudgingly, the researcher might give up on detecting such small effects and instead ask what effect sizes *can* be detected with $N = 500$? An effect-size sensitivity analysis could then reveal that $N = 500$ can detect $d = 0.25$ with 80% power. The researcher might decide this is close enough to the original intended effect size, and proceed to gather 500 participants. Alternatively,

if $d = 0.25$ seems inadequate, the researcher could seek additional resources in order to collect the initially recommended sample size for $d = .20$, thus avoiding wasting their existing resources seeking an effect which the study would be underpowered to detect.

Misunderstanding #2: "Power Can Be Intuited Based on the Number of Participants"

In practice, research psychologists have often used the sample size reported in a study as a proxy for power. However, sample size is not the whole story about power. Power is also tied to study design, analytic choices, and other features of the research. For example, a repeated-measures study with few participants, but many data points per participant, can have far greater power than a between-subjects study with many participants, but with few data points per participant. Given this, we caution against intuitions and heuristics based on the number of participants overall, or on *N*-per-design-cell guidelines (e.g. van Voorhis & Morgan, 2007), which are often calibrated for between-subjects designs. Instead, sample size evaluation based on power analysis will properly consider the number of *data points*, not just individuals, and will also properly consider the greater informational efficiency of repeated-measures analysis. For example, a within-subjects design comparing participants' evaluation of four emotion-eliciting scenarios will give the same number of data points (and power) as a between-subjects design that addresses the same question with four times as many participants, each evaluating one scenario. Furthermore, if responses to the scenarios are correlated within each individual, repeated-measures analysis will increase power even more by allowing individual-level "noise" to be removed, focusing on differences between scenarios. More points about within-participants analyses can be found in this article's Supplementary Materials.

After data are collected, a given research sample's power also cannot be properly evaluated based on a mere count of participants, as power functions are nonlinear and analysis-specific (Cohen, 1988). Therefore, any heuristic that deems a study inadequate from N alone has the potential to do serious disservice to designs that use participants efficiently. Even in between-subject designs, heuristics may be misleading because they rely on an assumption of effect size (for example, *N*-per-design-cell = 20 assumes an effect size $d = 0.91$ to achieve 80% power, and *N*-per-design-cell = 40 assumes $d = 0.64$). As we will see, it is unlikely that one effect size estimate can accurately characterize all relationships and manipulations across a single field of research, let alone many such fields in a discipline.

Misunderstanding #3: "There is a Single Power Level for a Study"

Researchers may be tempted, or requested, to report on "the" power level for a study. The most obvious problem with this nomenclature is that a single study may encompass different analyses and designs, which require different power analyses. But more meaningfully, because population effect sizes are unknown, researchers need to consider a range of possible effect sizes to accurately assess power for each study. Calling a study "high-powered" or "low-powered" is problematic because power is entirely dependent on the unknown population effect size for a given analysis. Although hidden, the population effect size is also crucial, because it determines *what* an analysis has power to detect. For example, a sample test with $N = 100$ has 99% power to detect a correlation effect size of $\rho = .50$ yet only 52% power to detect $\rho = .10$. Similar calculations can be made across a range of all possible effect sizes. For this reason, power is best thought of as a curve across this range rather than a single value. A study can only be reported as

having 80% power in the context of a given effect size and a given analysis (e.g., "the final sample had at least 80% power to detect correlations $\rho > .28$ at $\alpha = .05$").

One corollary of this property is that all analyses have 80% (or 90%, or 95%) power *for some effect size*. By this logic, all analyses have adequate power -- and inadequate power! That is, a given analysis has high power to detect *some* (large) effect, and low power to detect *some* (small) effect. The question then becomes whether the analysis has adequate power for effect sizes that are meaningfully likely to occur in the context of interest. We address the question of how to determine target effect size under Misunderstanding #6.

Misunderstanding #4: "Power is Only Important for Controlling False Negative Rates"

It is widely understood that running adequately powered studies controls the risk of falsely arriving at negative outcomes, which in many research traditions are seen as failures. However, this is only a partial understanding. Power also improves the replicability of *positive* results (Szucs & Ioannidis, 2017). In a world where power to detect true effects is low, any given positive result is less likely to be a true positive, and thus relatively more likely to be a false positive.

For example, assume that 100 research studies on varied topics examine 30 true effects and 70 null effects in the population. If the power across all studies to detect their effect's population size is low (e.g., 12%), then $30 \times 12\% = 3.6$ true positives are expected (Szucs & Ioannidis, 2017). However, $70 \times 5\% = 3.5$ false positives arealso expected, so that nearly half of all observed significant effects are not true in the population. But at power of 80%, 24 true positives are found versus 3.5 false positives, and we can be much more confident that any

observed significant effect is true. Thus, power is important in controlling the field-wide dissemination of false positive results, as well as controlling study-level false negative rates.

Misunderstanding #5: "Power Analysis Based on the Currently Obtained Effect Size is Meaningful for Evaluating the Current Study"

Researchers may wish to know whether a given analysis—theirs or someone else's—was adequately powered. In such cases, it is incorrect to use the sample's *observed* effect size to determine the power of that analysis (Cohen, 1988; Gelman, 2019). Unfortunately, some users of statistical packages wrongly interpret power analyses using the *sample-observed* effect size in exactly this way. For example, some procedures in the SPSS package allow for a calculation of the "observed power" on the basis of the sample effect size (IBM Corp., 2017). This kind of power estimate is uninformative, because it is a monotonic function of the *p*-value (see Goodman & Berlin, 1994). In particular, if $p > .05$, then observed power will always be less than 50%. Power based on observed values does not add new information, beyond echoing the already-known significance test. Instead, we recommend that researchers in this situation should, enter the given sample size and desired power into an *effect-size sensitivity* analysis to determine whether the analysis was powered to detect meaningful effect sizes.

Misunderstanding #6: "One Should Always Power for Effect Size X"

Because required *N* depends on effect size, all rules of thumb for sample size (e.g., *N* per cell; Simmons et al., 2011; van Voorhis & Morgan, 2007) also imply an effect size that researchers are, by default, powering for. As argued above, instead of relying on these heuristics for picking a sample size, researchers should conduct a power analysis that requires either (1)

inputting an effect size or (2) evaluating the effect size of the resulting analysis. However, this advice solves one problem and replaces it with another—what effect size should researchers use? We contend that it would be a mistake to replace one heuristic with another. A principled way of thinking about effect sizes is necessary. We review two options here: effect size precedent and smallest effect size of interest.

Effect size precedent. When using previously observed effect sizes as a guide, it seems reasonable to 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 sizes that 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 (this corresponds to a $d = 0.43$ and an $f = .215$[3]). 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 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. For example, studies in group processes tended to produce larger effects ($r = .32$) 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 moral licensing, and the appropriate meta-analytic estimate used. Finally, in the most focused case, researchers

[3] The *d* and *f*, like the *r*, are typical effect size statistics that power analysis software will request to make a sample size determination. They correspond to ANOVA results and, in the case of *d,* also to t-tests.

performing a direct replication of a given study often use its reported effect size as a ready target (Brandt et al., 2014).

But in making more precise calibrations, it is underappreciated that a study's methodological paradigm, not just topic, can influence effect sizes. Consider methods of manipulating interracial threat. At the subtle end of the spectrum, one can manipulate minimal features of vignettes, such as the year when the United States is expected to become a "majority-minority" nation (Craig & Richeson, 2014). That manipulation should produce a smaller effect size, all else equal, than a more vivid procedure in which, for example, White and Black men are paired to chat about racial profiling (Goff, Steele, & Davies, 2008).

In fact, we might question the wisdom of taking effect size estimates from a meta-analysis at all to guide our own efforts. A novel study by definition will not share the exact combination of independent variable, dependent variable, and setting in any published study. What's more, one paradigm with a tendency toward stronger or weaker effect sizes might dominate any given meta-analysis for arbitrary reasons, biasing its overall estimate. What might be useful in triangulating our expectations are meta-analyses of effects from paradigms that cut across multiple substantive topics, but these, unfortunately, are few (e.g., Forscher, Lai et al. 2019 on bias reduction techniques across attitude topics; Shaffer & Postlethwaite, 2012, on the predictive effects of contextual vs. non-contextual personality instruments).

Likewise, effect sizes might be increased by an efficient design or analysis that, for example, uses repeated measures to reduce theoretically uninteresting variance from between-participant differences, or uses covariates to dampen "noise" in the outcome variable. And finally, a manipulation that ropes together many different effective mechanisms may be conceptually imprecise, but it will probably give stronger effects than an intervention that uses

only one isolated mechanism. Thinking about tasks and measures used in previous studies, then, should inform choices about effect size, but for now this process happens more in an impressionistic than a precise way.

Another major limitation to the effect-size precedent approach comes from the distorting effects that publication bias and questionable research practices (QRPs) have on effect sizes in a literature. Most meta-analytic estimates come from published studies, plus 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 study results live in a file drawer, biasing our estimates of the true effect size. What's more, within a study, the main conclusion may be based on the most favorable of many possible analyses chosen *post hoc.* This QRP will tend to inflate effect sizes compared to analyses chosen *a priori*. Because significant, hence larger, results are more likely to be published, literature-based estimates are often too large (Dickersin, 1990).

Simply put, in a world where only large effects can be significant (i.e., low power), then all significant effects will be large. Consider a literature that tests a true effect whose size is d = 0.20, but with inadequate power (e.g., 16%). Over repeated testing, the average effect size over *all* tests will indeed be d = 0.20. However, most tests will not be significant; by definition, only 16% of tests will be significant, and they will be the tests with the highest sample estimates of effect size. If only significant findings are published under publication bias, these "surviving" sample effects will be higher (on average) than the population d = .20, biasing the literature toward larger effect sizes. As power increases, or publication bias is mitigated, then more sample estimates will be declared fit to publish and fewer 'weak' results will be filtered from the literature--making the literature more representative of the population. Thus, low power in

combination with publication bias may be a contributor to biased effect sizes in the literature. This situation can help explain, among other things, why studies with smaller sample sizes are less likely to show significant independent replication compared to those with higher sample sizes, even when their observed effect sizes look healthy (e.g., Open Science Collaboration, 2015).

The amount of publication bias that needs to be adjusted for can be difficult to infer, though some heuristics and methods have been proposed (see Lewis & Michalak, 2019; Simonsohn, 2015). As less-biased effect size estimates emerge from large-scale replication projects, and from articles in the Registered Report format that are approved for publication prior to knowing results, more literatures – especially in social psychology -- should begin to support accurate effect size determination by precedent (Klein et al., 2014; Scheel, Schijen, & Lakens, 2019).

Smallest Effect Size of Interest. If these arguments make researchers reluctant to rely on the literature, they could anchor power analysis instead on their idea of the smallest effect size of interest (SESOI) (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 developing and disseminating the intervention. For cheaper interventions, the SESOI might be lower.

But without clear benchmarks from applied outcomes, it can be difficult for teams conducting research on basic questions to determine their SESOI. Are interventions that produce

d = 0.10 changes on 7-point Likert scales important to study, or does d need to be at least 0.20? Anecdotally, many researchers take Cohen's "small" effect size guidelines such as $\rho = .1$ or $d = .2$ as a SESOI, but this choice seems to be based more on its availability as the smallest named value in that system, than on any detailed analysis. With more justification, if not necessarily clarity, 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).

The problem with these factors is that they are rarely possible to specify exactly or with consensus. More tangible criteria, perhaps, could be based on the typical methods and resources available to other basic-question researchers in a given field and research population. For example, a $\rho = .028$ effect needing 10,000 participants for 80% power may be unreachable outside of the largest data collection contexts, but 500 might reasonably be found to give the same power to detect $\rho = .124$. These latter kinds of consideration may help researchers who are planning new studies decide what range of effect sizes they would consider worthwhile. Spending wagonloads of resources on a huge sample, just to detect effect sizes that very few labs can practically replicate or build upon, might not be the best use of those resources in terms of the larger community of basic researchers.

Misunderstanding #7: **"Effects That are Significant Despite a Low-Powered Analysis 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. However, the survivor myth depends on wrongly concluding the reverse, following the logical fallacy, *affirming the consequent*: that if a significant effect is observed, the null hypothesis is unlikely (Ioannidis, 2005). In fact, low power reduces the likelihood that an observed significant result reflects a true positive effect, as discussed in Misunderstanding #4.

Misunderstanding #8: "Power is the Only Way to Plan Sample Size."

Researchers might assume that *a priori* power analysis is the only way to quantitatively determine a sample size ahead of time. Here, we present two alternatives that may be more attractive depending on a researcher's goals: precision analysis and sequential analysis.

Precision Analysis. Sometimes researchers will want to do more than reject the null hypothesis. 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 analysis means that the confidence interval (CI) for the effect is narrow. The CI gives a range of effect size values around the effect size estimate, based on the standard error. Assuming that the observed parameters are true in the population, a replication study drawn from the same underlying population should find parameter estimates that fall within the specified CI, $(1-\alpha)$% of the time (most commonly, 95%, in keeping with the conventional 5% α

level). The CI 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 et al., 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. Further, a study with good power will not necessarily have a narrow confidence interval. Maxwell et al. (2008) provide an example of a study comparing two means with $d = 0.50$. A sample size of $N = 128$ (64 per group) provides 80% power, but that sample size results in a predicted 95% confidence interval ranging widely from 0.15 to 0.85. Similarly, a study with narrow confidence intervals does not 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 -0.10 to 0.20 but only 9.5% power. Thus, an AIPE analysis can be useful for selecting the appropriate sample size for the desired level of precision, but not for determining 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 desired confidence interval.

Sequential Analysis / Optional Stopping. Traditionally, a researcher specifies one sample size *a priori*. However, uncertainty about the population effect size could lead to a study being underpowered and missing effects in the population, or being overpowered and needlessly exhausting resources. To balance power and feasibility concerns, optional stopping techniques let researchers make data-dependent changes to their sample size while correcting for an increased false positive rate. In these *sequential analysis* designs, 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 corrected for multiple testing, and/or the achieved *N*. This method, if done openly and correctly, is by no means the same as *undisclosed* optional stopping without correction, which has been rightly criticized as a practice leading to false positive inflation and low replicability in psychology (Simmons et al., 2011).

Optional stopping techniques have some drawbacks. Sample sizes from studies stopped early will be smaller, and so effect-size estimates will be less precise. In addition, studies that stop early will still 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. But drawbacks aside, we think that sequential analyses are a potent and underappreciated approach to research, in areas experiencing uncertainty about what effect sizes they should be aiming for.

Misunderstanding #9: "Power Must Always be 80%"

As noted above, the 80% power criterion suggested by Cohen (e.g., 1988) is now widely used, with its implication that a false positive is four times more important to avoid (5% risk given H_0) than a false negative (20% given H_1). More specifically, the 80% value represents an inflection point in the trade-off between cost and power. There is a near-linear relationship between sample size and power, until power hits .80. That is, a similar percentage increase in sample size is necessary moving from power of .50 to .60 to .70 to .80, roughly a 25% increase in sample size at each step. However, moving from .80 to .90 requires roughly a 33% increase, while gaining just .05 more power by moving from .90 to .95 requires a similar increase, about 33%.

We suggest that in psychology, 80% should be a bare minimum, based on its standing as an inflection point. However, if the increased resources can be justified, 90% represents a more rigorous standard, and 95%—exactly balancing false-negatives with false-positives—a strong ideal. 99% power is even stronger, but its advantages have to be weighed against the costs of getting there from 95%; to detect a population $\rho = .20$ at 95% takes N = 317, but at 99% takes N = 450, a 42% increase! It is important to justify trade-offs involved in adopting any particular power criterion, as with any particular value of α (Lakens, Adolfi, et al., 2018). The best practice is to report results for multiple reasonable power levels (e.g., 80%, 90%, and 95%).

Misunderstanding #10: "All Power Analyses Can be Easily Done with One Software Tool"

As the importance of power analysis has grown, tools for conducting it have flourished. One popular tool is the freely available and highly cited software, G*Power (Faul, Erdfelder,

Lang, & Buchner, 2007). However, just as there is not a one-size-fits-all solution for power analysis, no one software or analytic approach will be appropriate in all instances.

For example, non-analytic approaches that depend on simulation are *not* implemented in G*Power and similar software. These are often needed for accurate power analysis where analytic approaches, only involving formulas would, be too difficult to compute or have not yet been derived. Non-analytic approaches can be accessed in a variety of procedures written in statistical programs such as Mplus or R, as well as in some stand-alone applications. These techniques are particularly valuable for mediation, structural equation modeling, multilevel analysis, and other complex multivariate procedures.

Many of these non-analytic approaches use Monte Carlo techniques, in which 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, means and standard deviations can be input by assuming standardized data (SD = 1) and expressing mean differences in terms of these standard units (that is, to represent an effect size $d = 0.5$, or half a standard deviation, you might input one mean as -0.25 and the other as 0.25). Correlations among variables can be input 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 Monte Carlo approaches may at first seem intimidating, they can often be simplified in order to make them manageable.

A full review of power techniques is too detailed for the present article, but we have made available a critical review of tools for commonly used analyses in psychology in our Supplementary Materials. We take G*Power 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 describe some special considerations explained more fully in the Supplementary Materials, when using G*Power for regression and ANOVA:

- 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 Superpower (formerly ANOVApower) R packages allow calculation from unbiased effect sizes.
- Repeated measures ANOVA: This procedure is currently not documented in G*Power, but appears in its menus. It is important for users to do two non-obvious things: always select the option "effect sizes as in SPSS," from the Options window; and, for factorial repeated measures, enter "number of measures" using the numerator degrees of freedom of the ANOVA plus one (not the total number of measures in the design). If these steps are not followed, power will be greatly overestimated.

Table 1. Summary of Specific Power Analysis Methods Recommendations

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

- Factorial ANOVA: G*Power's sample size recommendations for factorial designs require some 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, will interaction effect sizes approach simple effect sizes.

Recommendations for Best Practices

In the sections above, we have responded to ten misunderstandings about power analysis and sample-size planning. We follow these with positive recommendations for best practices in three areas: planning future research, reporting power analysis in published research, and evaluating existing research on the basis of power. We also address the difficulties researchers may face in achieving a desirable level of power in research involving populations or methods that are more difficult than usual to work with.

Planning Future Research

Beyond deciding sample size in planning a study, *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. Researchers are encouraged to experiment with between- and within-subject variants of project ideas to assess the benefits that a within-subjects design might offer for any particular study. 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).

Arriving at an exact effect size may be difficult, but agreeing upon a reasonable range of effect sizes should not, starting from the likelihood that effect sizes from previous studies or meta-analyses are subject to publication bias and should at the very least not be exceeded. Without taking a principled approach to effect sizes and other decisions, as we advise, power analysis' usefulness will be lost. In a field where precedents for estimating effect sizes are currently uncertain due to publication bias, we suggest determining sample sizes and designs on the basis of resources typically accessible to research labs, then presenting for evaluation the minimum effect size that sample and design can detect at various desirable levels of power.

Reporting Power Analyses

Current writing guides for psychologists (e.g. the Journal Article Reporting Standards or JARS; APA Publication Manual, 7th ed., 2019; Appelbaum, Cooper, Kline, Mayo-Wilson, Nezu, & Rao, 2018) leave unclear how power analyses should be reported in manuscripts. We offer recommendations here.

As noted earlier in Misunderstanding #3, power within a single study may vary if multiple conclusions draw on statistical analyses with different tests, designs, and/or presumed effect sizes. In this case, one or more power *analyses* (plural) should be described, not in the Methods section, but in the Results section close to each type of analysis (following Sleegers,

2019). The *Participants* subsection of Methods need only specify which of these analyses, if any, the overall sample size was based on.

If sample size was decided *a priori* via power analysis, make sure to report the software and analysis option used, effect size (with units, e.g. d, f^2), rationale for choosing an effect size, target power (including, as we have suggested, values for 80%, 90% and 95% power), and any other parameters used in the power analysis. We also recommend full reporting of parameters and decisions if precision or sequential analysis are used.

However, *a priori* power analysis should not be reported 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, effect-size sensitivity analysis is the most useful and honest tool (Cohen, 1988). Even when sample size is planned via power analysis, missing or incomplete responses may reduce the amount of *usable* data, reducing achieved power and also making effect-size sensitivity analysis advisable. For example, an author who followed their plan to recruit 352 participants, with 80% power to detect an effect size $d = 0.30$, but only could keep 298, might state that the final analysis had "80% power to detect an effect of $d = 0.33$".

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 or how to use power in evaluation.

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. Other practices that inflate type I error, such as *undisclosed* optional stopping, also reduce confidence in the study's accuracy. A statement that all measures, manipulations, and even relevant studies are disclosed can give greater confidence in effect size estimates from research (Simmons et al., 2012). Preregistration and Registered Reports can also help ensure full disclosure of research practices.

To understand how power affects the evidence value of observed *p*-values, as mentioned under Misunderstandings #4 and #7, one must understand what error rates do—and do not—say about research. If $\alpha = .05$, many researchers and educators make the fallacious assumption 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) or more accurately, false positive risk (FPR; Colquhoun, 2019). An argument could then be made for attempting to restrict this risk to 5% or some other low number (e.g., Colquhoun, 2019).

Critically, the FPR depends on the frequency of false positives (determined by α) and true positives (determined by power), as well as the odds of the effect actually being true (prior odds). High power to detect a given effect size means the FPR is closer to an acceptable number. For example, in a study with a very low power of 10% given the population effect size, and uninformed prior odds of 1:1, the FPR is 33%, whereas an identical study with power of 80% has FPR of 5.9%. A middling power of 40% leads to an FPR of 11.1%, meaning that the α level needed to reach the same FPR as the study with 80% power would be closer to .025 than to .05. That is, *p*-values close to .05 are particularly untrustworthy in lower-powered studies, because

they are unlikely to reach the α level required to achieve an acceptable risk of false positives. (All calculations were facilitated by the online resources at Schönbrodt, 2019.)

In evaluating the power of a published article, it may be tempting to use the observed effect size and sample size in a *power determination* analysis. However, this method is flawed (see Misunderstanding #5). 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. Criteria for typical and minimal effect sizes are often not well established, and depend on the topic, method, and application of the research, as we have discussed. Because power analysis depends on difficult decisions about likely effect sizes, we are wary of suggesting a one-size-fits-all correction factor to apply to *p*-values under "low" power. However, we can suggest that if a study has considerably less than 80% power to detect the kind of effects that are typical or useful in a literature, then *p*-values in the .01 to .05 range should be viewed with caution as evidence for a proposition. The lower the power, the more we should reduce our willingness to accept *p*-values close to $\alpha = .05$ as evidence.

Power for Difficult Research Cases: Cautions and Solutions

Although it may be justified to be cautious about the results of studies low in power, excluding such results from publication and other forms of dissemination can limit a scientific field in undesirable ways. Because publication is a major metric of hiring, tenure, and promotion, these decisions will impact scholars' judgments about whether to pursue a particular type of research in the first place.An inflexible policy of rejecting "low-powered" research could thus 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 groups that are simply more difficult to study than relatively affluent Western citizens (WEIRD populations; Henrich, Heine, & Norenzayan, 2010).

Conversely, standards requiring high power to detect reasonable effects in a given field are most easily reached through samples such as undergraduate students, who can be recruited relatively easily, quickly, and in large numbers. But these samples are simply not appropriate or possible for some research questions. Additionally, prioritizing undergraduate samples can systematically exclude scholars from institutions with smaller participant pools, decreasing the diversity of perspectives in our field. Researchers have also 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). However, online samples are not going to be valid for all research contexts; for example, research involving immersive face-to-face social environments or tangible behavioral outcomes (Anderson et al., 2019).

As an example of the kind of research that strict sample size requirements might disadvantage, consider 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. Or, consider 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 between Whites and people of color in the United States, but additionally in how these ethnic disparities may be exacerbated in elderly populations.

In both cases, conducting research with 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. Further, 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, analyses targeting the kind of effect sizes detectable by larger studies are likely to show low power, despite their most assiduous efforts. In this case, a rigid decision to reject the work based on conventional power criteria 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 methodologically difficult studies with low power to detect effect sizes that are usual for the field of 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. For example, in research on a population where at most 100 individuals can be recruited at once, the smallest effect size detectable at 80% power is $\rho = .28$, meaning that many effects of a size typically found in social/personality studies would be underpowered in this research. In evaluating research, editors might consciously adopt different thresholds (e.g., a different power criterion, or higher α) for difficult methods or studies difficult-to-reach populations. Such publications would be allowed to be more tentative than publications addressing 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 involving a policy or health-related intervention. 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. Labs conducting similar work may benefit from collaborations pooling together resources and samples 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 effect sizes under publication bias, and aiding future meta-analyses.

Conclusion

Within social and personality 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 the statistical complexity surrounding a

multitude of different approaches. To remedy this, we have provided background on statistical power and other approaches, addressed ten potential misconceptions, provided a compendium of resources, and advocated for a combination of a priori and effect-size sensitivity approaches. With our overview of specific techniques and software, we have also further empowered researchers to conduct and evaluate research in line with sample-size considerations.

If there is one take-home message, it is that issues of power depend crucially on questions of meaningful effect size, which social and personality psychology have largely avoided tackling in theory and methodology development. The approximate nature of effect size criteria should be a caution against applying overly rigid "bright lines" to power statistics, and against repeating the mistaken ways in which the *p*-value has been treated as a live-or-die criterion of evidence (Wasserstein & Lazar, 2016). In emphasizing the essential role of effect size in power analysis, we challenge researchers and reviewers to reframe their evaluations of pending or completed research. Instead of asking "does this study have enough power?" we should ask "What effects does this study have acceptable power to detect?"

Author contributions: Authorship order was determined as follows: The first author convened the working group and took the lead in writing a first version, together with all authors. A second version incorporated major revisions, worked on by the first four authors. The second through fourth authors thus appear in alphabetical order, followed by the others in alphabetical order.

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 and in particular its past 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: A previously 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.

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

American Psychological Association. (2019). *Publication manual of the American Psychological Association* (7th 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.

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

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

Colquhoun, D. (2019). The false positive risk: a proposal concerning what to do about p-values. *The American Statistician*, *73* (sup1), 192-201.

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.

Forscher, P. S., Lai, C. K., Axt, J. R., Ebersole, C. R., Herman, M., Devine, P. G., & Nosek, B. A. (2019). A meta-analysis of procedures to change implicit measures. *Journal of Personality and Social Psychology*, *117*(3), 522-559.

Frick, R. W. (1998). A better stopping rule for conventional statistical tests. *Behavioral Research Methods, Instruments, & 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.

IBM Corp. (2017). IBM SPSS Statistics for Windows, Version 25.0. Armonk, NY: IBM Corp.

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

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., . . . Nosek, B. A. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology*, *45*(3), 142-152. http://dx.doi.org/10.1027/1864-9335/a000178

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.

Scheel, A. M., Schijen, M., & Lakens, D. (2020, February 5). An excess of positive results: Comparing the standard psychology literature with registered reports. https://doi.org/10.31234/osf.io/p6e9c

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

Shaffer, J. A., & Postlethwaite, B. E. (2012). A matter of context: A meta-analytic investigation of the relative validity of contextualized and noncontextualized personality measures. *Personnel Psychology*, *65*(3), 445-494.

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

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.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *The American Statistician*, *70*(2), 129-133.

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. L. (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/.

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

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/.

Wang, Y. A., & Rhemtulla, M. (in press). Power analysis for parameter estimation in structural equation modeling: A discussion and tutorial. *Advances in Methods and Practices in Psychological Science*.

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 generalized 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*.