Meta-Psychology, 2025, vol 9, MP.2020.2643
https://doi.org/10.15626/MP.2020.2643
Article type: Original Article
Published under the CC-BY4.0 license

Open data: Yes
Open materials: Yes
Open and reproducible analysis: Yes
Open reviews and editorial process: Yes
Preregistration: Yes

Edited by: Rickard Carlsson
Reviewed by: Ulrich Schimmack, Erin M Buchanan
Analysis reproduced by: Federico D'Atri
All supplementary files can be accessed at OSF:
https://doi.org/10.17605/OSF.IO/PZ296

A Multi-faceted Mess: A Review of Statistical Power Analysis in Psychology Journal Articles

Nataly Beribisky[1], Udi Alter[1], and Robert A. Cribbie[1]
[1]Quantitative Methods Program, Department of Psychology, York University, Toronto, Ontario, Canada

Many bodies recommend that a sample planning procedure, such as traditional NHST a priori power analysis, is conducted during the planning stages of a study. Power analysis allows the researcher to estimate how many participants are required in order to detect a minimally meaningful effect size at a specific level of power and Type I error rate. However, there are several drawbacks to the procedure that render it "a mess." Specifically, the identification of the minimally meaningful effect size is very challenging, the procedure is not precision oriented, and does not guide the researcher to collect as many participants as feasibly possible. In this study, we explore how these three theoretical issues are reflected in applied psychological research in order to better understand whether these issues are concerns in practice. To investigate how power analysis is currently used, this study reviewed the reporting of 443 power analyses in high impact Psychology journals in 2016 and 2017 using Google Scholar. It was found that researchers rarely use the minimally meaningful effect size as a rationale for the chosen effect in a power analysis. Further, precision-based approaches and collecting the maximum sample size feasible are almost never used in tandem with power analyses. In light of these findings, we offer that researchers should focus on tools beyond traditional power analysis when sample planning, such as collecting the maximum sample size feasible.

Keywords: power analysis, statistical power, Type II error rate, precision-based sample planning, minimally meaningful effect size

Introduction

Psychology researchers are continually critiqued for conducting studies that are underpowered (Maxwell, 2004). One of the consistently suggested recommendations to address the problem of low power is to conduct sample planning procedures like power analysis (Mistler, 2012; Association for Psychological Science, 2018; Wilkinson, 1999). In traditional null hypothesis significance testing (NHST) a priori power analysis, a researcher estimates the number of subjects required to detect a minimally meaningful effect size (MMES) at a given level of power ($1-\beta$, where β represents the Type II error rate), and Type I error rate (α) (Dienes, 2014; Sedlmeier and Gigerenzer, 1992). Accordingly, by specifying the desired power to detect an MMES, NHST-based power analysis is meant to decrease the likelihood of a study being underpowered. Unfortunately, in practice, when conducting traditional NHST a priori power analysis, researchers often do not consider what an appropriate MMES might be and instead rely on past experiences to come up with a target effect size. Further, upon seeing a large (and maybe infeasible) sample size, researchers may try different variations of the parameters (effect size, alpha level, power) in order to obtain a sample size that is attainable. Accordingly, although NHST-based power analysis is meant to help reduce power problems in psychological research, it sometimes complicates them further.

There are three reasons why traditional NHST-based power analysis is a problematic sample planning procedure. The first is the difficulty and feasibility of determining the MMES parameter. The MMES, or the smallest effect size that the researcher would find practically (e.g., clinically) important, is often challenging or impossible to estimate (Dienes, 2014). Quantifying it properly requires a thorough understanding of the measures, target population, and theories surrounding the study. However, specifying it appropriately is necessary in order to conduct power analyses correctly (Dienes, 2014). Second, unlike other sample planning procedures such as accuracy in parameter estimation (AIPE) (Kelley and Maxwell, 2012), traditional NHST-based power analysis does not consider precision but instead focuses on detecting the presence or absence of an effect. With no focus on confidence intervals, traditional NHST power analysis may be an outdated tool as

Psychology research moves towards an increased usage of estimation tools (Cumming, 2012). Third, traditional NHST-based power analysis does not support collecting the maximum sample size feasible (MSSF) for a given research study. The MSSF strategy weighs practicality constraints explicitly in the sample planning process, unlike traditional NHST power analysis.

The paper is organized in the following manner: 1) We begin by providing a brief introduction to sample size planning/power analysis; 2) We describe in detail each of the theoretical issues with sample size planning/power analysis raised in the previous paragraph; 3) We conducted a review of sample size planning/power analysis reporting in high-impact peer-reviewed Psychology journals to ascertain how this procedure is used; and lastly, 4) We provide recommendations to researchers regarding best practices for sample size planning/power analysis.

Types of Power Analyses

Statistical power is a function of sample size, Type I error rate, and the MMES. In traditional NHST a priori power analysis, a researcher estimates the required sample size from the other three parameters (power level, Type I error rate and MMES). For example, assume a researcher is looking to detect a difference in means with a two-group independent-samples t-test. Assume also that $1 - \beta = .95$, $\alpha = .05$ (two-tailed test), and within the context of the study, the researcher has reason to believe that the MMES is a Cohen's $d = .70$. Entering these into a power calculator, such as the `pwr.t.test` function within the `pwr` package in R (Champely, 2020), produces an estimate of the required sample size, which is a total sample of 108, or 54 per group.

Hypothetically, any three of these four parameters (sample size, power, α, MMES) could be used to solve for the fourth. For instance, a researcher might want to use a pre-existing dataset (with a fixed sample size) to find the minimally detectable effect size. Another researcher might want to determine how much power they had to detect an observed effect size for data that has already been collected. It is worth noting that when sample size is used as one of the input parameters (rather than as the estimated outcome), a procedure is necessarily no longer a sample planning tool. However, for the purposes of the review, we analyzed the reporting of all four variants of this procedure. The three outlined reasons power analysis is problematic (difficulty knowing the MMES, not precision-based, and not conducive to choosing the MSSF) also apply to these versions of power analysis. For clarity, we briefly outline the three alternative versions of power analysis below.

Sensitivity Power Analyses

Sensitivity power analyses are often used when researchers are working with existing datasets or otherwise constrained sample sizes. These types of power analyses have become increasingly common and even recommended as sample planning options by certain journals (Aberson, 2018; Elsevier, 2020). There are two types of sensitivity analyses: (a) effect size as the outcome, and (b) power level as the outcome. In sensitivity power analysis with the effect size as the outcome (also referred to as effect-size sensitivity analysis), researchers use a pre-determined sample size along with a specified α and power level to estimate what can be called the minimally detectable effect or minimal effect size (Giner-Sorolla et al., 2019). In sensitivity analysis with power as the outcome, a pre-determined sample size, an MMES, and a given level of α are used estimate the existing power in a given sample.

Post-Hoc Power Analyses

In post-hoc power analysis, researchers use an observed effect size, utilized sample size, and α to estimate the observed power within the study conducted. This type of power analysis has been noted to have many issues, in addition to the ones it shares with a priori and sensitivity analyses, such as the observed power being a one-to-one function with the p-value (Hoenig and Heisey, 2001). It has been noted to be especially problematic and its use is not recommended.

The underlying problems affecting all versions of this procedure are the same. The first issue, the difficulty and importance in choosing a correct MMES, may be the most critical.

Issue 1: The Importance and Necessity of MMES Specification

To avoid potentially under-powering a study, a researcher should specify an MMES (Algermissen and Mehler, 2018; Dienes, 2014; Zodpey, 2004). The MMES and the power level are directly related, such that the smaller the MMES, the lower the statistical power (holding sample size and α constant). Another way to say this is that the smaller the MMES, the larger the sample size must be in order to achieve an acceptable level of power (relative to the sample size required if the MMES is larger). Accordingly, specifying an effect (e.g., observed in a previous study) that is larger than an appropriate MMES can cause the study to be underpowered for detecting the MMES (Dienes, 2014). For example, let's say the MMES for a given study/context is $d = 0.80$, but the researcher mistakenly/inappropriately conducts a traditional power analysis using an effect size of d

= 1.00 (maybe based on results from a pilot study). The sample size required for a given power (let's say 0.90) will be lower for d = 1.00 than for d = 0.80, and thus if the researcher adopts the sample size required for achieving 90% power with d = 1.00, they will not have achieved 90% power for detecting the MMES of d = 0.80. In contrast, specifying an effect that is smaller than an appropriate MMES may suggest a sample size larger than what is necessary or feasible in the given study. For example, a researcher may expect a small effect size, but it would not be appropriate to use this small effect size as the MMES in a power analysis because it might not be meaningful.

If the observed effect size for the relationship of interest is expected to be larger in magnitude than the MMES, and power is the only consideration, then an effect larger than the MMES could be used in the power analysis (with the obvious limitation that if an effect is observed that is larger in magnitude than the MMES, but smaller in magnitude than the expected effect, there would be reduced power for such an effect). Additionally, even in this scenario, an MMES must be explicitly defined in order to know its relative position to any effect specified in a power analysis. Otherwise, there is no guarantee that, for instance, the specified effect is not smaller than the MMES, meaning that the corresponding power analysis would aim to detect a potentially unimportant effect. Accordingly, there is no way to conduct a power analysis correctly without clearly expressing the appropriate MMES; it must be specified a priori for both traditional a priori power analysis and sensitivity analysis with power as the outcome. Further, it should coincide with the effect obtained for sensitivity analysis with the effect as the outcome. Relatedly, another setting where the MMES is extremely useful is in equivalence testing where the MMES can be used to create an equivalence bound. In equivalence testing, any value that falls within the equivalence bound is considered negligible (for more information on equivalence testing see Wellek, 2010).

Specification of the MMES requires a thorough consideration of a study's context (Aberson, 2015); in other words, determining the smallest effect size that would be practically significant given the nature and measurement of the variables under study. The MMES cannot simply be defined as an observed effect from a small-sample pilot study or an observed effect from prior literature (Algermissen and Mehler, 2018). The MMES is also different than the smallest expected effect, since its interpretation is tied to importance or clinical value. For these reasons, it has been noted that the MMES "may be one of the hardest aspects of a theory's prediction to specify" (Dienes, 2014, p. 3). The considerable degree of subjectivity in choosing an MMES for power analysis can render the procedure quite difficult for both exploratory and confirmatory research, as in both instances choosing an MMES may require an unreasonable amount of guesswork about new phenomena. Practically, this parameter is often unknown and difficult to identify (Lipsey, 1990).

Examples within other disciplines, such as medical research, state that the MMES should be both realistic and important, and of substantial interest based on the phenomenon under study (Fayers et al., 2000). Thus, the selection of an MMES should be a multifaceted process involving a review of prior research in the area, which includes combining existing quantitative information about effects (the magnitude of which may also be overestimated: see Open Science Collaboration, 2015; Albers and Lakens, 2018) with opinions from stakeholders via panels and focus groups (Cook et al., 2018). However, depending on the nature of the research, ideal methods for selecting an appropriate MMES may not be available, which could lead to the selection of an arbitrary value for the effect size used in a traditional power analysis. To summarize, the selection of the MMES parameter is extremely challenging. Despite this, correct specification of an MMES is unavoidable if a researcher wishes to elicit meaningful information from a traditional power analysis.

Issue 2: Precision-Based Sample Planning

The specification and interpretation of confidence intervals provide more information than statistical significance alone when analyzing results (Cumming, 2012; Thompson, 2002). In the planning phase of a study, traditional power analysis aims to detect the presence or absence of an effect, based on statistical significance, rather than the precision or width around the effect size estimate. Less commonly utilized sample planning tools, such as precision-based sample planning (e.g., AIPE or accuracy in parameter estimation), allow researchers to estimate the sample size required to obtain a desired confidence interval width (Corty and Corty, 2011). The use of precision-based sample planning is not restricted to an NHST framework, making it useful in many more scenarios than traditional power analysis (e.g., Bayesian statistics and equivalence testing). Further, advancements in this area make use of procedures such as sequential analysis to identify the optimal sample size using a given stopping rule (see Kelley et al., 2018).

In order to demonstrate how precision-based sample planning may be used, we present an example of precision-based sample planning for a two-sample t-test. For all analyses, the procedure manipulates the standard confidence interval formula to solve for sample size. In this case, the approximate confidence interval formula for a two-sample *t*-test is,

$$\bar{X}_1 - \bar{X}_2 \pm 2 \cdot s_\mathrm{p} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

where s_p is the pooled standard deviation, 2 approximates the two-tailed t critical value (e.g., $t_{.975,\,\mathrm{df}\,=\,40} = 2.02$), and n_1 and n_2 represent the sample sizes in group one and group two, respectively. Assume $\alpha = .05$ (for a two-tailed test), a pooled standard deviation of 20, and a desired 95% confidence interval with a width of 4 points on a given outcome measure. To simplify, also assume that groups one and two are of equal size such that the resulting confidence interval is,

$$\bar{X}_1 - \bar{X}_2 \pm 2 \cdot s_\mathrm{p} \sqrt{\frac{2}{n}},$$

where $n = n_1 + n_2$. In order to determine the number of individuals required per group, half of the confidence interval width (for our example, the full width of 4 is divided by 2) is inserted into the formula. At this point, it is possible to get an estimate of the required sample size per group:

$$2 \approx 2 \cdot 20 \sqrt{\frac{2}{n}}$$

$$\frac{1}{20} \approx \sqrt{\frac{2}{n}}$$

$$\frac{1}{400} \approx \frac{2}{n}$$

$$n \approx 800 \text{ per group or 1600 total}$$

As can be noted from the example, the two pieces of information that the researcher specified in this scenario (other than α) are the confidence interval width and a measure of variability (in this case, pooled standard deviation). A measure of variability may be obtained from prior research and/or pilot studies, and, unlike the MMES, will not fluctuate based on a given study's aims.

To be clear, researchers conducting precision-based power analysis are not required to select extremely narrow confidence intervals that would necessitate a very large sample size. For example, in exploratory research, it may not be necessary to measure effects with high precision and therefore a wider interval would be appropriate. In contrast, for more confirmatory studies, higher precision (narrower intervals) may be desired, which will then require comparably larger sample sizes. This association is also reflected in equivalence testing (and other measurement settings), where the narrower the confidence interval, the more precise the measurement of the effect of interest.

It has been recommended that precision-based approaches be used in tandem with traditional NHST power analysis (Kelley et al., 2003). However, it is unknown how often precision-based sample planning is used in conjunction with, or instead of, traditional power analysis. This question will be addressed empirically in this paper.

Issue 3: Maximum Sample Size Feasible

We contend that using the MSSF, by explicitly outlining how a sample was chosen and what considerations were taken, is a sample planning strategy in and of itself. Obtaining the MSSF is quite different than obtaining the maximum sample size possible. In the former scenario, factors such as cost (e.g., volunteer, paid or course credit), accessibility (e.g., easy to access vs. rare population), and risk (e.g., minimal or more than minimal) are all thoroughly considered and described when estimating the sample size required for a given study. In contrast, in the latter scenario, participants could potentially be recruited with no concrete stopping point, without regard to cost, accessibility, or risk; this would qualify as irresponsible research practices. To summarize, even though it may be possible to select an impressive number of participants/units, it is important to consider feasibility factors when deciding on a final sample size.

For researchers that work with difficult to access populations, such as those studying giftedness or the physiology of individuals with rare disorders, small sample size is assumed to be "the rule rather than the exception" (Rost, 1991, p. 236). Often, this means that any NHST-based power analysis that a researcher conducts will result in an unattainable reference point for the number of participants required (unless the MMES is exceptionally large). As the most straightforward method of increasing power is simply to increase one's sample size, obtaining the MSSF within the context of the study is a useful way to approach participant recruitment planning.

To expand on the point made in the previous paragraph, a researcher working within a small population may actually be compelled to avoid running their study altogether for fear of being underpowered. Accordingly, a potential counter-argument against using the MSSF is that when there is a very limited pool of participants it might not be valuable to conduct the study at all. However, if a researcher has done all that they can to feasibly obtain the most participants practical, we argue that (with feasibility constraints clearly outlined in the sample planning section) their findings are still valuable (assuming other aspects of the research are satisfactory). This is especially true now that replication and meta-analyses are routinely conducted within the discipline.

In other words, we contend that it is much better to conduct valuable research with a small MSSF than to abandon the research altogether because of a small sample. Necessitating researchers to consistently have large sample sizes can cause either an abandonment of the study or downgrades in study quality; both of these consequences can lead to less original research and study design (e.g., changing features of a study to facilitate switching from in-person to online participant platforms for greater recruitment). Indeed, many have noted the importance of multi-site studies and meta-analyses, as opposed to simply relying upon a large sample size within a single study (e.g., Schauer and Hedges, 2020; Kelley and Rausch, 2006). Accordingly, instead of fixating on whether a single study has sufficient power to detect an MMES, explicitly defining what concerns came into sample planning decisions may allow for research to be integrated more efficiently into the literature.

Obtaining the MSSF also aligns with the notion that, when done correctly, power analyses really only allow researchers the opportunity to estimate, rather than calculate, a sample size for their study (Batterham and Atkinson, 2005). In actuality, the estimate may only serve to provide researchers with a reference point for whether their study will require, “tens, hundreds, or thousands of participants” (Williamson et al., 2000, p.10). Accordingly, as NHST-based power analysis is a procedure that estimates, rather than calculates, the sample size required, calls for power analysis to be the sole decision-maker for determining the precise sample size necessary (e.g., to avoid subjecting participants to protocols that may be higher risk) are unfounded.

There is another argument in support of the MSSF that looks at the problem from a different angle; if a researcher is feasibly and practically able to obtain more participants than the estimated sample size from a traditional power analysis, there is little reason to terminate enrolment at the power analysis estimate. Specifically, recruiting more participants beyond the estimated value would give a study greater likelihood to detect an existing effect and, even more important, greater precision (Tanaka, 1987). Specifically, larger sample sizes lead to a greater non-centrality parameter between the null and alternate distributions, while also minimizing the variability in each distribution (Kelley and Maxwell, 2012). Both of these features allow for the difference between null and alternative distributions to be amplified, leading to a higher likelihood of detecting a true effect (Kelley and Maxwell, 2012). Accordingly, researchers that terminate recruitment after reaching the sample size suggested from an a priori power analysis estimate (i.e., ignoring the MSSF) may lose the benefits that come from utilizing a larger sample if a greater sample can be feasibly and responsibly obtained. It is worth investigating, therefore, how often researchers use power analysis in conjunction with the strategy of obtaining the largest number of participants practical for the given study.

The Present Study: Power Analysis from Intention to Use

The present study aimed to explore the state of these three issues within the reporting of power analyses, by reviewing recently published articles in high impact peer-reviewed Psychology journals. First, we investigated the proportion of researchers that used power analyses as an a priori sample size planning tool, compared to alternative uses of the method, such as sensitivity and post-hoc power analyses. Second, we recorded how often the MMES, as recommended, was used as the effect size in the power analysis. Third, we looked for instances of researchers using precision-based sample planning to plan for the width of a confidence interval around an effect size, rather than planning for the presence of the effect itself. Finally, we looked at how often researchers stated they used the MSSF to select an appropriate sample size.

It is important to highlight the relationship between the theoretical discussion above and the empirical investigation of the reporting practices in power analysis/sample size planning. Although the theoretical issues associated with traditional power analysis/sample size planning methods that we discussed above alone render these practices a mess, and hence lead to the conclusion that these practices should be abandoned, the suggestion that researchers should abandon these traditional methods is indefensible if the power analyses conducted by Psychology researchers do not suffer from the issues raised above. For example, maybe we are wrong or outdated and in practice researchers have no issue selecting an appropriate MMES. Or, maybe researchers are frequently conducting precision-based sample planning. Further, maybe researchers typically ignore the recommendation to conduct a traditional power analysis and conduct the research with the MSSF. If the theoretical issues we highlight in the sections above do not actually exist in practice, then there is no need to modify recommendations regarding best practices for determining appropriate sample sizes for research (since researchers are already abiding by best practices). We know of no other study that has explored the reporting of power/sample size planning analyses within Psychology.

Method

Journal Articles

Journal articles published in 2016 and 2017 from 12 high impact Psychology journals were chosen for this analysis. We sampled journals from primary research areas within Psychology which had an impact factor greater than 1.80, based on the Journal Citation Reports (Clarivate Analytics, 2018). The journals and their impact factors are presented in Table 1.[1]

The search was completed in May 2018. Articles from these journals that employed a power analysis were located by entering search terms into *Google Scholar*. Specifically, articles containing either the term "power analysis" or "power analyses", that were published in the desired years and journals, were identified via a Google Scholar advanced search. From this initial collection, articles that did not conduct any type of power analysis were excluded. Two reviewers, working independently, found that from the 3,524 articles published in 2016-2017 within the selected journals, there were 623 articles that met our search criteria. After the reviewers removed articles that did not conduct a power analysis, 443 articles remained for review. The number of articles from each journal is presented in Table 1.

Measures

Each power analysis reported in an article was coded based on the research questions. Coding information thus addressed: (a) the type of power analysis conducted, (b) the reporting of the MMES, (c) the instances of precision-based sample planning, and (d) the collection of participants based upon the MSSF. In addition to these primary variables, we also collected supplementary information including psychological subfield (to allow for comparisons in reporting), frequency and magnitude of reporting of α and power levels, and the specific software used for conducting the power analyses.

Type of Power Analysis Conducted

To determine the relative frequency of the different forms of power analysis, we coded: (1) the type of power analysis reported (a priori, sensitivity, or post hoc); and, if the power analysis was a sensitivity analysis, (2) the outcome of interest (effect size or power level).

The Reporting of the MMES

To answer questions surrounding the use of an MMES in power analysis reporting in cases where the effect size is an input parameter (i.e., for a priori power analyses and sensitivity analyses with the power as the outcome), the following pieces of information were recorded:

1. Is any effect size reported within the power analysis?
2. Is any justification given for the effect size reported within the power analysis?
3. What is the justification presented for the effect size chosen (prior research, average effect size in sub-field, MMES)?
4. What was the scale and magnitude of the effect size adopted?

Precision-Based Sample Planning Use

To assess how often precision-based sample planning is conducted relative to traditional NHST a priori analyses, the instances of precision-based sample planning within the articles were recorded.

Meeting Sample Size Targets

For traditional NHST a priori power analyses, we assessed whether researchers met their estimated required sample size, by coding: (1) whether the researchers reported the estimated participants required from their analysis, (2) whether researchers used the MSSF, (3) how closely the number of participants enrolled in the study met the number of participants suggested by the power analysis, and (4) whether the researchers met this target after any exclusionary criteria and/or outlier removal.

If there was more than one power analysis within an article, only the first power analysis was coded to eliminate any nonindependence issues. Further, if there were multiple power conditions included within a single power analysis, only the first power condition was recorded. For example, if an article stated that a calculated sample size had 90% power to detect a Cohen's d of 0.90 and 80% power to detect a Cohen's d of 0.65, only the former power condition was recorded.

Procedure

After gathering the set of relevant articles, two coders read through these articles and coded the reported

[1] Note that we deviated from our original preregistration where we indicated that we would aim to collect data from 20 journals; we were able to obtain a sufficient sample of power analyses from only 12 journals.

Table 1

List of High Impact Peer-Reviewed Psychology Journals

Number of Articles	Journal	Impact Factor (2017)
8	Journal of Child Psychology and Psychiatry	6.486
56	Psychological Science	6.128
40	Journal of Personality and Social Psychology	5.733
10	Journal of Applied Psychology	4.643
9	Journal of Abnormal Psychology	4.642
39	Journal of Consulting and Clinical Psychology	4.537
131	Journal of Experimental Psychology	4.107
45	Computers in Human Behavior	3.536
4	Journal of Cognitive Neuroscience	3.468
19	Biological Psychology	2.891
64	Personality and Social Psychology Bulletin	2.498
18	Motivation and Emotion	1.837

power analyses. Initially, the two coders practiced coding a subset of the articles to ensure consistency between coders for each classification. In this initial stage, when coders found inconsistencies in their recording, the coders discussed the discrepancies and mutually agreed upon the best categorization. After this training stage, the two coders reached 96.78% agreement on a subset of 9 articles (based on the cells from the coding spreadsheet containing the same information). After the training stage, the coders worked independently. When classification questions arose during the coding process, they would be discussed by the coders until there was a mutually agreed solution on how best to categorize the results.

Results

The results below are presented either as frequencies and proportions or using histograms. As discussed above, in addition to the reported results, there were also tangential questions that were addressed and are reported in the supplementary information. These questions and analyses, along with the coding sheet, PRISMA checklist for the study, R files, and dataset (for all primary and secondary variables) can be found at https://osf.io/5jy9h. The preregistration of the study (conducted through the Open Science Framework) can be found at https://osf.io/tzhdj.

Type of Traditional NHST Power Analysis Conducted

Of the 443 power analyses conducted, 295 (66.59%) were a priori power analyses, 39 (8.80%) were post hoc power analyses, and 109 (24.60%) were sensitivity-based power analyses. The majority of sensitivity analyses had power, rather than effect size, as the outcome of interest. Specifically, within the sensitivity analyses, 81 (74.31%) had power as the outcome and 28 (25.69%) had effect size as the outcome.

The Reporting of the MMES

Out of the 376 a priori and sensitivity analyses with the power as the outcome recorded, 269 analyses (71.54%) included the utilized effect size within the power analysis reporting. Further, from the 376 power analyses, 168 (44.80%) provided some justification for the effect size chosen (of the 269 analyses that did report an effect size, 132 [49.07%] justified the effect size reported). It is worth noting, given the numbers above, that a small subset of the articles would provide justifications for the effect size selected while failing to state the effect magnitude itself.

Table 2 lists the justifications presented for the selection of the effect size chosen for the power analysis. Only two power analyses out of the 376 power analyses used the MMES as a justification for an effect's selection! Prior research (including prior meta-analyses) was the most common justification presented (n = 135, 35.90%) within the power analyses. Less frequently reported justifications included prior power analyses and pilot studies. Lastly, the justifications contained in the "Other" category of Table 2 included the average effect size in research, Cohen's small effect, the result of a simulation study, or a smallest expected difference between conditions.

Figures 1, 2, and 3 provide the frequencies of the three most common effect sizes used in traditional

Table 2

Justifications for the Effect Sizes Selected within Power Analyses for A Priori Power Analyses and the Sensitivity Analyses with Power as the Outcome

Justification	Raw Count	Percentage
Not included	208	55.20%
Included	168	44.80%
MMES	2	0.005%
Prior Research	135	34.67%
Pilot Study	18	4.80%
Prior Power-Analysis	1	0.003%
Other	12	3.20%
Total	376	100%

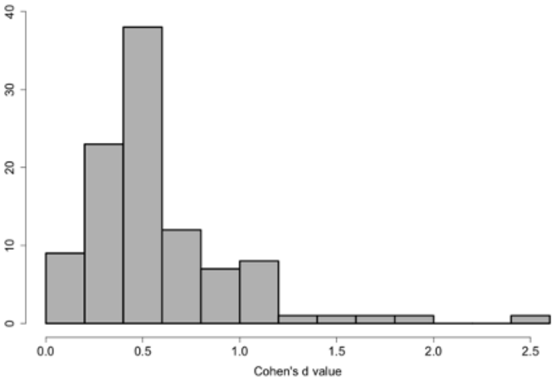

Figure 1

Reported Cohen's d Effect Sizes for A Priori Power Analyses and Sensitivity Analyses with Power as Outcome.

NHST a priori and sensitivity analyses with power as the outcome (Cohen's d, Pearson's ρ and Cohen's f). Cohen's d was, by far, the most commonly used effect size for power analyses, appearing in 102 of 269 analyses that reported any effect size (37.92%), compared to Cohen's f and Pearson's ρ which were recorded 52 times (19.33%) and 34 times (12.64%), respectively. The most frequently utilized effect size for Cohen's d was 0.50. Most commonly used Pearson's ρ values were around .20 and .30. Finally, the most common Cohen's f value reported was .25. Interestingly, most of these values fall at or near the cutoff for a medium effect size according to Cohen (1988). Other reported effect sizes included η^2 and R^2. A notable proportion of analyses would also report effect sizes without a unit (e.g., simply stating "a medium effect" or .50).

Instances of Precision-Based Sample Planning

Out of the 443 recorded power analyses, only one analysis (less than 1%) was precision-based.

Meeting Sample Size Targets

Recording the Number of Participants Required

For the 295 a priori power analyses, the number of participants required was reported in 90.20% of cases (266 analyses), with 257 (87.12%) requiring a total sample size of 500 participants or less. The range for the required total sample size was between 5 and 2600, with the median total sample size required being 84. The distribution of estimated total sample sizes required is presented in Figure 4. The left graph presents the distribution of participants required for all power analyses recorded. In order to present a magnified view of the majority of the distribution, the right graph is a histogram of participants required for power analyses that reported requiring a sample of 500 or less.

To assess how frequently researchers were able to obtain a sample size near the power analysis reference point, we adopted an estimation-based approach and recorded the proportion of instances where studies were able to obtain at least 90% of the sample size required during enrolment and analysis (the choice of 90% was made in order to avoid penalizing power analyses that were still within a reasonable range of the required sample size). Most power analyses met their sample size targets. Only 11 studies obtained a sample size that was less than 90% of the sample size required at the enrolment stage (before exclusionary criteria and outlier removal). However, 29 analyses did not provide enough

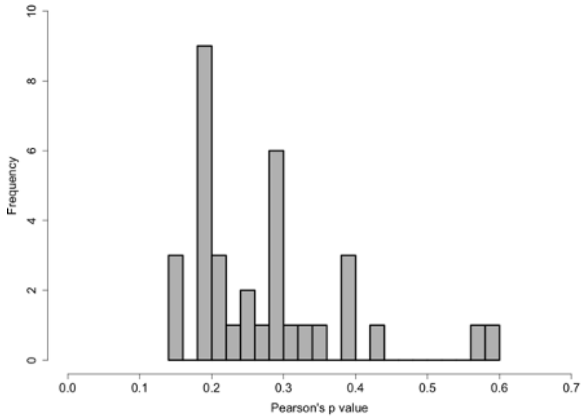

Figure 2

Reported Pearson's ρ Effect Sizes for A Priori Power Analyses and Sensitivity Analyses with Power as Outcome.

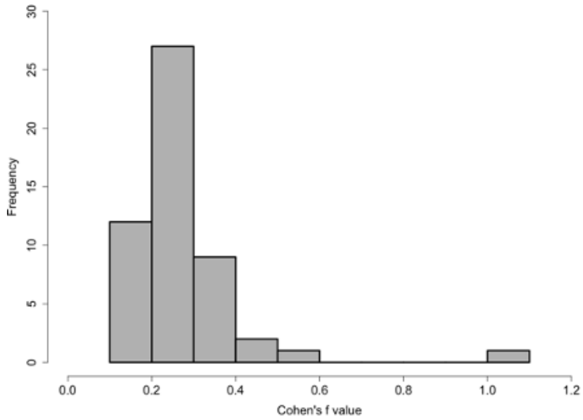

Figure 3

Reported Cohen's f Effect Sizes for A Priori Power Analyses and Sensitivity Analyses with Power as Outcome.

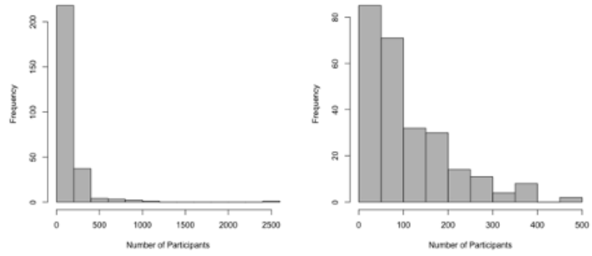

Figure 4

Number of Participants that are Estimated as an Outcome in A Priori Power Analyses. Note that N = 266 for left histogram and N = 257 for right histogram.

information to assess how or whether they met sample size targets. At the analysis stage (after exclusionary criteria and outlier removal), out of 295 a priori power analyses, 17 additional studies (28 total) were not able to obtain 90% of the sample size required. One hundred sixty-four studies (55.59%) reported having more participants than required for their power analysis. Further, 11 studies (3.73%) discussed the MSSF by describing their feasibility restrictions in participant recruitment in some way.

Discussion

Statistical power analysis is a heavily-relied upon sample planning tool, traditionally used by researchers to estimate how many participants are required to detect an MMES, at a given level of power and α. In this study, we have outlined three prominent issues with traditional power analysis. First, the specification of the MMES in power analysis is difficult to quantify in advance but necessary in order to conduct the procedure properly. Second, traditional power analysis does not consider confidence intervals or precision at the outset of a study like approaches such as AIPE. Finally, traditional power analysis does not explicitly require obtaining as many participants as feasibly possible, which is a practical sample planning strategy in its own right. Although these three issues provide evidence that traditional power analysis may not be a reasonable sample planning strategy, before this study it was unclear how these issues were reflected in reporting practices. More precisely, power analysis could be a valuable tool if researchers were consistently selecting the MMES as the effect size, conducting precision-based sample planning in conjunction with traditional power analyses, or selecting the MSSF.

To investigate whether these theoretical issues were reflected in practice we conducted a review of power analysis reporting in high impact Psychology journals. Generally, our findings show that researchers have difficulty reporting and providing a rationale for their power analyses. Accordingly, it may be worth modifying recommendations regarding best practices for sample planning.

Although the majority of traditional NHST power analyses recorded were a priori power analyses, a notable portion still analyzed power after the fact, and a large number of studies conducted sensitivity analyses with pre-existing data sets. Therefore, although traditional NHST power analyses were mainly presented as a sample planning tool, different variations of the same procedure were used to analyze the existing power to detect an effect, to find the smallest effect one may detect in a given sample, and finally to calculate how much power one had to detect an observed effect. These variations from a priori power analysis suggest that the procedure is being implemented in different time points of research and with goals beyond sample planning. It is worth reiterating that all variations of this procedure are still impacted by the same types of problems as traditional NHST a priori power analysis.

The MMES unfortunately was almost never recorded as a rationale for an effect size used within a power analysis. Without choosing the effect size in this way, a researcher risks wasting resources (if the chosen effect size is less than the MMES) or having reduced power for detecting the MMES (if the chosen effect size is larger than the MMES) (Lipsey, 1990). Our findings suggest that rather than being used to detect an MMES, contemporary power analyses largely aim to detect effects found in prior research, which are known to fluctuate in magnitude just as much as *p*-values (Gelman, 2019). This finding is consistent with other research that has found that effect sizes for power analyses have been largely based on Cohen's *d* cut-offs, prior literature or a prior study (Bakker et al., 2020). Accordingly, selecting an effect size in this way severely limits the value/applicability of the sample sizes determined via such a procedure. Although the difficulty of choosing an appropriate MMES cannot be overstated, quantifying an MMES is necessary for properly conducting all NHST-based power analyses.

Precision-based approaches were almost never used as a method of sample planning within the power analyses surveyed. Our results suggest that precision-based approaches are not common sample planning tools in Psychology. One of the advantages of a precision-based approach is a focus on confidence intervals, which allow researchers to move beyond the dichotomous thinking of whether or not an effect exists (that is often associated with interpreting *p*-values alone), and begin to answer questions related to the magnitude of effects, such as how much an intervention has caused a variable to change (Cumming, 2014). However, even for precision-based sample planning, there are many instances in which an effect must be specified to estimate the required sample size, even for designs such as bivariate correlations and chi-square contingency table tests (see Corty and Corty, 2011, for examples). As analyses grow more complicated (e.g., moving from *t*-tests to linear regression), the number of pieces of information that are required to estimate the sample size grows. This unfortunately leads to the same difficulty as traditional NHST power analysis and may result in the guessing of a number of parameters.

When recruiting participants, researchers appeared to use the sample size estimate determined by a power analysis as a specific requirement, rather than as a guide regarding what approximate sample size might be necessary to detect an MMES. This can be evidenced by the fact that less than 5% reported using the MSSF by mentioning feasibility constraints. Further, the median absolute difference between the number of participants obtained for the study and the number of participants suggested by the power analysis was only 8. Taken together, these findings imply that either: (a) researchers responsibly (and very accurately) obtain the sample size suggested by the power analysis, (b) researchers may be manipulating the parameters of their power analysis to match their already obtained sample size or to a sample size they can feasibly obtain, or (c) commonly used power analysis parameters in the field of Psychology ensure that the sample size required can easily be recruited. The results found in our review may have resulted from a combination of these factors.

Limitations

There are some limitations of this study. First, we focused our review on 12 high-impact Psychology journals, focusing on the years 2016 and 2017. Our results may not extend to articles published in other journals or years. Further, although we tested different variations to ensure our search criteria could capture as many power analyses as possible (e.g., using the search terms separately versus together), it may have missed analyses within articles that did not match our search terminology. In other words, our recording of power analyses relied quite heavily upon the precise wording used to describe these sorts of analyses within each article.

Precision-based approaches may have also been under-represented within our review, because the terminology for these types of analyses may not be reflected by our search criteria. However, it is worth noting that we conducted a post hoc search of precision-based terminology on Google Scholar and were not been able to locate any extensive evidence of its use in Psychology in 2016 or 2017. Specifically, searching "precision based sample size calculation", “precision-based power analysis”, “AIPE”, or "precision based sample calculation” in Google Scholar for 2016 and 2017 produced no further empirical articles from psychological journals that conduct precision-based sample planning.

Recommendations

Our review supports the notion that when using a statistical tool that has many theoretical challenges researchers inevitably run into difficulties implementing the tool. These challenges in implementation often make the power analysis procedure infeasible or render its results misleading at best.

Although power analyses that are precision-oriented are superior to traditional approaches of power analysis because the focus is placed on estimation rather than reject/not reject, they still require pre-specification of a confidence interval width, an estimate of the population variability, and often times the effect of interest itself. Further, like traditional power analysis approaches with a small MMES, high precision confidence intervals can require large sample sizes which can be infeasible for innovative research.

In contrast, the MSSF does not require the quantification of specific parameters and considers feasibility issues that affect sample size. We contend that many researchers may already do this informally, where even before sample planning begins they have an understanding of practical constraints on their sample size. We believe that responsibly outlining how many and how these feasibility concerns (e.g., access difficulties, financial cost, and risk) came into play when determining a sample size may provide more information on sample planning than a traditional NHST based power analysis. We recommend that researchers adopt the MSSF strategy and justify these constraints when describing how they selected their sample size. Although there are power analysis procedures that explicitly factor the idea of "cost" (see Guo and Luh, 2009), they are all nested within traditional NHST-based power analysis (and therefore face the same challenges). To conclude, our study highlights the multi-faceted mess that is traditional NHST-based power analysis. Some readers of this paper might conclude that it is necessary to accurately specify the MMES or more thoroughly outline their own traditional NHST-based power analysis in their paper. However, as the field of Psychology moves toward estimation as the primary goal, we argue that researchers should move away from exclusively focusing on the dichotomous detection of a single effect, and instead focus on replication and meta-analysis. More specifically, because it does not have a unique focus on statistical power in NHST, and it does not require guessing many difficult parameters, we argue that obtaining the maximum sample size feasible is a better approach to sample planning than traditional power analysis. In this way, through replication via multi-site studies and meta-analyses of existing literature, we can shift the focus from the statistical power of a single study to a cumulative literature of many findings.

Author Contact

Robert A. Cribbie, Quantitative Methods Program, Department of Psychology, York University, Toronto, ON, Canada.Email: cribbie@yorku.ca

ORCID IDs:
NB https://orcid.org/0000-0002-1081-0125
UA https://orcid.org/0000-0003-3133-839X
RC https://orcid.org/0000-0002-9247-497X

Conflict of Interest and Funding

The authors have no competing interests to declare. The authors thank SSHRC for their financial support.

Author Contributions

All authors developed and aided with writing the manuscript. NB and UA were the two coders of the articles.

Open Science Practices

This article earned the Preregistration+, Open Data, Open Materials and Open Code badge for preregistering the hypothesis and analysis before data collection, and for making the data, materials, and code openly available. It has been verified that the analysis reproduced the results presented in the article. The entire editorial process, including the open reviews, is available in the online supplement.

References

Aberson, C. L. (2015). Statistical power analysis [Advance online publication]. In R. A. Scott & S. M. Kosslyn (Eds.), *Emerging trends in the behavioral and social sciences*. Wiley.

Aberson, C. L. (2018). Improving scientific practices and increasing access. *Analyses of Social Issues and Public Policy*, *18*(1), 7–10. https://doi.org/10.1111/asap.12152

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

Algermissen, J., & Mehler, D. M. (2018). May the power be with you: Are there highly powered studies in neuroscience, and how can we get more of them? *Journal of Neurophysiology*, *119*(6), 2114–2117. https://doi.org/10.1152/jn.00765.2017

Association for Psychological Science. (2018). Submission guidelines. https://www.psychologicalscience.org/publications/psychological_science/ps-submissions

Bakker, M., et al. (2020). Recommendations in preregistrations and internal review board proposals promote formal power analyses but do not increase sample size. *PLOS One*, *15*(7), e0236079. https://doi.org/10.1371/journal.pone.0236079

Batterham, A. M., & Atkinson, G. (2005). How big does my sample need to be? a primer on the murky world of sample size estimation. *Physical Therapy in Sport*, *6*(3), 153–163. https://doi.org/10.1016/j.ptsp.2005.05.004

Champely, S. (2020). *pwr: Basic Functions for Power Analysis* [R package version 1.3-0]. https://CRAN.R-project.org/package=pwr

Cook, J. A., et al. (2018). Delta2 guidance on choosing the target difference and undertaking and reporting the sample size calculation for a randomised controlled trial. *BMJ*, *363*, k3750. https://doi.org/10.1136/bmj.k3750

Corty, E. W., & Corty, R. W. (2011). Setting sample size to ensure narrow confidence intervals for precise estimation of population values. *Nursing Research*, *60*(2), 148–153. https://doi.org/10.1097/NNR.0b013e318209785a

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge.

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

Dienes, Z. (2014). Using bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*, 781. https://doi.org/10.3389/fpsyg.2014.00781

Elsevier. (2020). Guide for authors. https://www.elsevier.com/journals/journal-of-experimental-social-psychology/0022-1031/guide-for-authors

Fayers, P. M., et al. (2000). Sample size calculation for clinical trials: The impact of clinician beliefs. *British Journal of Cancer*, *82*(1), 213–219. https://doi.org/10.1054/bjoc.1999.0902

Gelman, A. (2019). Don't calculate post-hoc power using observed estimate of effect size. *Annals of Surgery*, *269*(1). https://doi.org/10.1097/sla.0000000000002908

Giner-Sorolla, R., et al. (2019). Power to detect what? Considerations for planning and evaluating sample size [Preprint]. https://osf.io/jnmya/

Guo, J. H., & Luh, W. M. (2009). Optimum sample size allocation to minimize cost or maximize power for the two-sample trimmed mean test. *British Journal of Mathematical and Statistical Psychology*, *62*(2), 283–298. https://doi.org/10.1348/000711007X267289

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

Kelley, K., Darku, F. B., & Chattopadhyay, B. (2018). Accuracy in parameter estimation for a general class of effect sizes: A sequential approach. *Psychological Methods*, *23*(2), 226–243. https://doi.org/10.1037/met0000127

Kelley, K., & Maxwell, S. E. (2012). Sample size planning. In H. Cooper et al. (Eds.), *APA handbook of research methods in psychology, Vol. 1. Foundations, planning, measures, and psychometrics* (pp. 181–202). American Psychological Association. https://doi.org/10.1037/13619-012

Kelley, K., Maxwell, S. E., & Rausch, J. R. (2003). Obtaining power or obtaining precision: Delineating methods of sample-size planning. *Evaluation The Health Professions*, *26*(3), 258–287. https://doi.org/10.1177/0163278703255242

Kelley, K., & Rausch, J. R. (2006). Sample size planning for the standardized mean difference: Accuracy in parameter estimation via narrow confidence intervals. *Psychological Methods*, *11*(4), 363–385. https://doi.org/10.1037/1082-989X.11.4.363

Lipsey, M. W. (1990). *Design sensitivity: Statistical power for experimental research*. Sage.

Maxwell, J. A. (2004). Causal explanation, qualitative research, and scientific inquiry in education. *Educational Researcher*, *33*(2), 3–11. https://doi.org/10.3102/0013189X033002003

Mistler, S. (2012). Planning your analyses: Advice for avoiding analysis problems in your research. https://www.apa.org/science/about/psa/2012/11/planning-analyses

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*. https://doi.org/10.1126/science.aac4716

Rost, D. H. (1991). Effect strength vs. statistical significance: A warning against the danger of small samples. *European Journal of High Ability*, *2*(2), 236–243. https://doi.org/10.1080/0937445910020212

Schauer, J. M., & Hedges, L. V. (2020). Assessing heterogeneity and power in replications of psychological experiments [Advance online publication]. *Psychological Bulletin*. https://doi.org/10.1037/bul0000232

Sedlmeier, P., & Gigerenzer, G. (1992). Do studies of statistical power have an effect on the power of studies? In A. E. Kazdin (Ed.), *Methodological issues strategies in clinical research* (pp. 389–406). American Psychological Association. https://doi.org/10.1037/10109-032

Tanaka, J. (1987). "how big is big enough?": Sample size and goodness of fit in structural equation models with latent variables. *Child Development*, *58*(1), 134–146. https://doi.org/10.2307/1130296

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher*, *31*(3), 25–32. https://doi.org/10.3102/0013189X031003025

Wellek, S. (2010). *Testing statistical hypotheses of equivalence and noninferiority*. Chapman Hall/CRC.

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

Williamson, P., et al. (2000). Statistical review by research ethics committees. *Journal of the Royal Statistical Society A*, *163*(1), 5–13. https://doi.org/10.1111/1467-985X.00152

Zodpey, S. P. (2004). Sample size and power analysis in medical research. *Indian Journal of Dermatology, Venereology, and Leprology*, *70*(2), 123–128.