A MULTI-FACETED MESS: A REVIEW OF STATISTICAL POWER ANALYSIS IN PSYCHOLOGY JOURNAL ARTICLES

NATALY BERIBISKY

A THESIS SUBMITTED TO

THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF ARTS

GRADUATE PROGRAM IN PSYCHOLOGY

YORK UNIVERSITY

TORONTO, ONTARIO

June 2019

Abstract

The over-reliance on null hypothesis significance testing and its accompanying tools has recently been challenged. An example of such a tool is statistical power analysis, which is used to determine how many participants are required to detect a minimally meaningful effect in the population at given levels of power and Type I error rate. To investigate how power analysis is currently used, we review the reporting of 443 power analyses in high-impact psychology journals in 2016 and 2017. We found that many pieces of information required for power analyses are not reported, and selected effect sizes are often chosen based on an inappropriate rationale. Accordingly, we argue that power analysis forces researchers to compromise in the selection of the different pieces of information. We offer that researchers should focus on tools beyond traditional power analysis when sample planning, such as precision-based power analysis or collecting the largest sample size possible.

Table of Contents

List of Tables

List of Figures

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

The over-dependence on various statistical tools with the psychological toolbox has been recently challenged (Nuijten, 2016; Simmons, Nelson & Simonsohn, 2011). Disagreements about Psychology's reliance upon the null hypothesis significance testing (NHST) framework for data analysis have prompted proposal for the adoption of new procedures (Wagenmakers, 2008), a focus upon effect sizes (Cumming, 2014), and for some, even a complete abandonment of *p*-values (Trafimow, 2015). Despite many counter-movements to NHST, a systematic review of statistical reporting found that NHST remains a popular choice for researchers reporting statistical results (Counsell & Harlow, 2017).

Statistical power analysis is a sample planning tool that is heavily embedded within the NHST framework. Traditional *a priori* power analysis, an estimate of the sample size required to detect an effect at a given Type I error rate (α), is used only to detect an effect's presence, rather than to plan around the precision of an estimate itself. Despite this possible short-coming, *a priori* power analyses is still viewed as a best practice in psychological research. Specifically, many research bodies recommend the use of power analysis before data collection to avoid research being underpowered (Mistler, 2012; APS, 2018; Wilkinson and the APA task force, 1999).

However, beyond being grounded within NHST, an additional critique of power analyses is that the procedure is incredibly difficult to conduct and interpret correctly. To perform a power analysis the researcher must decide upon a minimally meaningful effect size (MMES) to use in their sample planning, which is usually unknown and difficult to estimate (Lipsey, 1990). Further, outputs of power analyses are often viewed as calculations rather than estimations

(Williamson, Hutton, Bliss, Campbell & Nicholson, 2000), creating a false sense of certainty around the output, which in turn leads to sample planning decisions that may be too stringent (either by being quick to state that the sample size is too small when it is slightly lower than the required sample size computed, or refusing to recruit a larger sample than the amount computed).

It is useful to review how power analyses are reported in the psychological literature in the context of current debates regarding the procedure's importance and feasibility. Specifically, poor power analysis reporting practices could be a symptom of an issue with the entire statistical procedure and trigger a need to reassess the role of power analysis as a viable sample planning procedure. Accordingly, to review these practices in depth, we conducted a systematic review of power analysis reporting in high-impact peer-reviewed psychology journals from 2016 and 2017. We review how often power analyses are conducted, whether researchers plan for precision around estimates or for the statistical significance of effects alone, whether the MMES is used within the analysis, and whether all necessary pieces of information are reported.

Traditional Power Analysis

Power is a function of a number of variables including sample size, effect size, and α. Power calculations are often used at different stages of research and for different reasons. For example, one researcher might want to determine appropriate *a priori* sample sizes, another researcher might want to use a pre-existing dataset (with a fixed sample size) to find the minimally detectable effect size or power level, or a third researcher might want to determine how much power they had to detect an observed effect size for data already collected.

To conduct a traditional *a priori* power analysis, a researcher estimates the minimum sample size that is necessary in order to detect an MMES at a predetermined level of power (1-β) and α (Dienes, 2014; Sedlmeier & Gigerenzer, 1989). For example, assume a researcher is

looking to detect a difference in means in a two-group independent-samples *t*-test. Assume also that 1- β = .80, α = .05 (two-tailed test), and the MMES is a Cohen's *d* = .10. Entering these into a power calculator such as the G*Power 3.1 software (Faul, Erdfelder, Buchner, & Lang, 2009), produces an estimate of the required sample size, which is a total sample of 3142, or 1571 per group.

Interpretational and Practical Challenges of Power Analysis

It has been noted that power analyses offer researchers the opportunity to estimate, rather than calculate, a sample size for their study (Batterham & Atkinson, 2015). The distinction between estimation and calculation is critical, as the term "sample size calculation" suggests that there is certainty about this estimate. 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).

Since sample sizes determined through power analysis should be interpreted as an estimate, rather than a fixed calculation, it is always better to collect more participants than computed by traditional *a priori* power analysis. 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 & 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 & Maxwell, 2012). Larger sample sizes also decrease standard error estimates and are more generalizable. Accordingly, researchers that terminate recruitment after reaching the sample size suggested from an *a priori* power analysis estimate may lose the benefits that come from utilizing a larger sample. Thus, if a researcher is feasibly able to obtain more participants than the estimated sample size, there is little reason to terminate

enrolment at the power analysis estimate, as recruiting more participants would only give their study greater power to detect an existing effect and more precision.

Treating a sample size computed by a power analysis as a calculation rather than an estimation may also create a belief that a sample size lower than the estimate will inevitably lead to a study that is under-powered and of little value. This type of ideology implicitly pushes statistical significance above other components of a study, such as descriptive statistics and the precision of effect size estimates. An over-reliance upon statistical significance has been convincingly challenged by many (e.g., Carver, 1978; Cumming, 2014; Fraley & Marks, 2007; McShane, Gal, Gelman, Robert, & Tackett, 2017). Further, these statements are particularly damaging to studies using rare or difficult to access populations, where sample sizes are almost guaranteed to be low. In these instances, a power analysis may discourage the researcher from running their study, though the experiment may still generate useful findings about the research question, either through a focus upon descriptive statistics or through its eventual incorporation into a meta-analysis.

To avoid potentially under-powering a study, a researcher must specify an MMES. This should be the smallest effect size that the researcher would find interesting or clinically interesting (Dienes, 2014). The MMES and the power level are inversely related, such that the smaller the MMES, the higher the power required to detect it (when holding sample size and α constant). Accordingly, specifying an effect (e.g., observed in a previous study) that is larger than an appropriate MMES can cause the corresponding study to be underpowered by failing to control Type II error rates and thereby potentially failing to detect an effect where one exists (Dienes, 2014). In contrast, specifying an effect that is smaller than an appropriate MMES may suggest a sample size larger than what is feasible in the given study.

However, specification of the MMES requires a thorough consideration of a study's context, variability in the dependent variables and the measures used in the experiment (Aberson, 2015). 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.

Guidelines in other disciplines, such as medical research, state that MMESs should be both realistic and important, of substantial interest based on the phenomenon under study (Fayers et al., 2000). The selection of an MMES should be a multi-faceted process involving a review of prior research in the area, which includes combining existing quantitative information about effects with opinions from stakeholders via panels and focus groups (Cook et al., 2018). However, ideal methods of selecting a minimally meaningful effect are often impossible to implement in practice, which may lead to the selection of an arbitrary value for the effect size.

Other Types of Traditional Power Analyses

Although traditional power analyses use an α, an MMES, and a specific level of power to estimate a sample size, researchers may also theoretically manipulate any three of these four variables to estimate the final one. Including sample size as an input in the power analysis formula results in other types of power analyses such as sensitivity power analysis and post-hoc power analysis.

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 required by certain psychology journals

(Giner-Sorolla, 2018). 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, 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. In sensitivity analysis with power as the outcome, a pre-determined sample size, a minimally meaningful effect, and a given level of alpha are used estimate the existing power in a given sample.

There are a couple of issues with sensitivity power analyses that should be noted. Specifically, sensitivity analyses face the same problem as traditional *a priori* power analyses, where their outcomes may be mistaken as calculations rather than estimations of a given variable. In the case of sensitivity analyses with an effect size as an outcome, there is arguably even less formal reasoning behind the minimally detectable effect size. By default, when determining this effect size researchers are not able to fully consider a study's context or variability, since the effect is simply estimated as a function of the other three variables.

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. Post-hoc power informs the researcher of what type of power they may expect if they replicated the study with the same sample size (and sample variance) and found the same effect size estimate.

There have been many arguments against the use of post-hoc power analysis. For instance, post hoc power has been argued to be noisy since it is based upon a noisy estimate of effect size (Gelman, in press). Consider a situation where a study's calculated effect size is similar in magnitude to the study's standard error. The effect size observed from the study, could potentially be attributed to error, and could have greatly fluctuated from the true effect-size.

Therefore, power calculated from the observed effect size will be a noisy estimate and an, "invitation to overconfidence" (Gelman, in press, p. 2).

Further arguments against the use of post hoc power analysis suggest that researchers may misinterpret the meaning behind observed power and that estimates of observed power are often biased (Gelman, in press; Hoenig & Heisey, 2001; Lenth 2001; Levine & Ensom, 2001; Yuan & Maxwell, 2005). Specifically, a common misinterpretation of post-hoc power analysis, known as the power approach paradox, states that high observed power for a non-statistically significant result is evidence of a true null hypothesis (Hoenig & Heisey, 2001). However, the reasoning behind this paradox has been demonstrated to be incorrect, since smaller *p*-values (leading to larger observed power) tend to be associated with more evidence that a null is false than larger *p*-values (which lead to smaller observed power) (Hoenig & Heisey, 2001). Further, only methods that test the hypothesis of no effect or a lack of association directly, such as equivalence testing, can be used as evidence to support a research hypothesis of minimal effect.

Precision-Based Power Analysis

Notably, 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 itself. Less commonly utilized sample planning tools, such as precision-based sample planning, allow researchers to specify a desired width of a confidence interval in order to estimate the sample size necessary to obtain that width of confidence interval at a given α. The use of precision based power analysis is also not solely restricted to an NHST framework, making it useful in many more scenarios than traditional power analysis.

A precision-based power analysis may occur as follows. Once again, assume a researcher is aiming to detect a mean difference in a two-group independent sample *t*-test, with α = .05 (for

a two-tailed test). Assume also that the researcher would like to have a 95% confidence interval that has a width of 4 points on a specific dependent measure. The researcher also estimates that the pooled standard deviation is about 30 (note that this estimate of variability can be obtained from prior literature more easily than the MMES, since the latter requires a knowledge of the centre of the distribution of the alternate hypothesis). By manipulating the confidence interval formula for a mean difference it is possible to find the number of participants required in each group.

$$\text{mean difference} \pm 2 \times s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Where, n_1 and n_2 are the sample sizes in each group and s_p is the pooled standard deviation. By inserting half of the desired confidence interval width into the formula, it is possible to obtain the estimated sample size that is required per group. Assuming the researcher would like groups of equal size,

$$\frac{4}{2} = 2 \times s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$1 = 30 \times \sqrt{\frac{2}{n}}$$

$$\frac{1}{30} = \sqrt{\frac{2}{n}}$$

$$\frac{1}{900} = \frac{2}{n}$$

$$1800 = n \text{ or } 3600 \text{ total participants}$$

It is worth noting that precision-based power analysis also does not avoid researcher subjectivity. Two pieces of information must be specified in addition to α: the pooled standard deviation of a measure of interest, which is obtainable through literature review (since it does not require a knowledge of the centre of the alternate hypothesis distribution), and the desired confidence interval width. Further, in contrast to traditional power analysis, procedures such as precision-based power analysis align more closely with some of the contemporary goals of statistical reporting by focusing on estimation rather than presence or absence of an effect alone.

The Present Study: Power Analysis from Intention to Use

The present study aimed to answer a constellation of questions surrounding the reporting of power analysis, by systematically reviewing recent published articles in high impact peer-reviewed psychology journals. First, we investigated the proportion of researchers that used the power analysis procedure as a sample planning tool compared to other stages of analysis, as recommended by major research bodies. To do this, we recorded the proportion of researchers conducting *a priori*, sensitivity, and post-hoc power analyses. Second, we recorded how often the minimally meaningful effect was used as a justification for the effect size selected for power analysis. Third, we looked for instances of researchers using precision-based power analysis to plan for the confidence interval around an effect size, rather than plan for the presence of the effect size itself. To gain an understanding of the specific values researchers use for their parameters within a power analyses, we recorded what pieces of information researchers provide in their power analysis reporting. This coding in particular allowed us to measure how often all necessary parameters are reported and which parameters (e.g., α) are most commonly adopted. Finally, to inform how regularly researchers use power analysis software with pre-specified parameters (such as G*Power), information about the statistical software used was also recorded.

Method

Journal Articles

Journal articles published in 2016 and 2017 from 12 high impact psychology journals were chosen for this analysis. These high impact journals were defined as those with an impact factor greater than 1.80, based on the Journal Citation Reports (Clarivate Analytics, 2018). Journals were categorized as belonging to one of the following sub-fields: social/personality, counselling/clinical/developmental, cognitive/neuropsych, and general psychology. There were three journals from each subfield, with each journal and subfield presented in Table 1.

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. Overall, from the 3,524 articles published in 2016-2017 within the journals, there were 623 search results, and 443 of these articles were ultimately utilized for this systematic review.

Measures

Each power analysis reported in an article was coded based on the research questions. Coding information thus addressed: (a) the stage at which researchers conducted the power analysis, (b) the reporting of the MMES, (c) the instances of precision-based power analysis, and (d) the completeness of the reporting.

Stage of research. 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 (for a priori and sensitivity analyses), 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 (minimally meaningful effect, prior research, meta-analysis, average effect size in sub-field)?

Precision-based power analysis use. To assess how often precision-based power analyses are conducted relative to traditional a priori analyses, the instances of precision-based power analyses within the articles were recorded.

Meeting sample size targets. For a priori power analyses, we assessed whether researchers met their estimated required sample size, by coding: (1) whether the number of estimated participants required was stated, (2) whether researchers used the maximum sample size available to them, (3) how closely the number of participants enrolled in the study met desired sample size requirements, and (4) how many participants were utilized in the analysis after any exclusionary criteria and/or outlier removal, and whether this number met the desired sample size.

The completeness of power analysis reporting. To assess whether all the required pieces of information were reported, as well as what specific parameters were used, the following information was obtained for each analysis:

(1) What statistical software was used for the analysis? (for a priori, post hoc, and sensitivity analyses)

(2) Was α reported? If so, what level was adopted? (for a priori, post hoc, and sensitivity analyses)

(3) Was the level of power reported? If so, what level was adopted? (for sensitivity and a priori power analyses)

(4) What was the scale and magnitude of the effect size adopted? (for a priori and sensitivity analyses)

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 using Google Scholar, two coders read through these articles and coded the power analyses reported. 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 for coding the required information.

Results

The results below are presented for all 443 power analyses together, without taking psychological sub-field into account. Although we looked at variations within sub-field, there were consistent indications of a lack of association between the sub-fields of psychology and any differences in power analysis reporting. The interested reader may view Appendix A for the specific statistical tests conducted to test for equivalence across sub-fields.

Stage of Research

Table 2 specifies the proportion of articles conducting each type of power analysis. Of the power analyses conducted, 66.59% were a priori power analyses, 8.80% were post hoc power analyses, and 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, about 75% had power as the outcome and about 25% had effect size as the outcome.

The Reporting of the MMES

Out of the 404 *a priori* and sensitivity analyses recorded, 295 analyses (73%) included the utilized effect size within the power analysis reporting. Further, from the 404 power analyses, 174 (43.07%) provided some justification for the effect size chosen (of the 295 analyses that did report an effect size, 138 [46.78%] 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 3 lists the justifications presented for the selection of the effect size chosen for the power analysis. Table 3 includes the selection results with and without sensitivity analyses with the effect as the output, as power analyses that have the effect as the output may conceptualize effect sizes differently than where it is an input (sensitivity analyses with power as an outcome

and *a priori* analyses). Only two power analyses out of the 404 sensitivity and *a priori* analyses used the MMES as a justification for an effect's selection. Prior research was the most common justification presented (n = 138) within the power analyses. Less frequently reported justifications included prior power analyses and pilot studies. Lastly, the least common justifications contained in the "Other" category of Table 3 included the average effect size in research, Cohen's small effect, the result of a simulation study, or a smallest expected difference between conditions.

Instances of Precision-Based Power Analyses

Out of the 443 recorded power analyses, only one analysis (less than 1%) was a precision-based power analysis. This small proportion of precision-based power analyses suggests that, so far, precision-based power analysis remains an under-utilized tool in psychological reporting.

Meeting Sample Size Targets

Recording of number of estimated participants required. For the 295 *a priori* power analyses, the number of participants required was reported in 90.20 % of cases (266 analyses). Of the 295 *a priori* analyses, 257 analyses or 87.12% required 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 Figures 1 and 2. Figure 1 presents the distribution of participants required for all power analyses recorded. In order to present a magnified view of the majority of the distribution, Figure 2 displays 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 reported obtaining a sample size that was less than 90% of the sample size required at the enrolment stage. However, 29 analyses did not provide enough information to assess how or whether they met sample size targets. At the analysis stage, out of 295 a priori power analyses, 28 studies reported not being 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, only 14 studies (4.75%) explicitly stated that they used the maximum sample size available to them.

The Completeness of Power Analysis Reporting

Statistical software used for the analysis. 56.21 % of studies (249 analyses in total) reported using software for their power analysis. G*Power software was by far the most common choice (n = 229 or 91.97%). Other less-commonly mentioned software included online power analysis calculators such as OpenEpi, Webpower and DanielSoper (4 analyses or 1.61%), coding packages such as R or Mplus (7 analyses or 2.81%), simulations with no listed software (3 analyses or 1.20%), and other sources such as using other power calculators (6 analyses or 2.41%).

Alpha level reporting. The α-level was reported in about half of the power analyses. Specifically, α was reported in 207 analyses or 46.73% of the time. Of the studies that reported

an α-level, an α of .05 was reported 95.17% of the time (197 analyses). Other α values and their respective frequencies are presented in Table 4.

Reporting of power levels. Arguably, sensitivity analyses that have power as the output may conceptualize power differently than the types of power analyses where power is an input (e.g., a priori power analyses and sensitivity analyses with the effect size as the outcome). For this reason, Table 5 presents the frequencies of power levels reported both with and without sensitivity analyses with power as outcome. When recording the frequencies, whenever power analyses presented their power as being, "equal to or greater than a certain power level" the power level referenced within the statement was the final number recorded (e.g., for a power level of ≥ .80, a power of .80 was recorded). Similarly, when analyses presented their power as, "greater than a certain power level" the next power level to the hundredth decimal place was recorded (e.g., for a power level of > .80, a power of .81 was recorded).

For the 404 *a priori* and sensitivity analyses listed in Table 5, power was reported in 354 studies or 87.62% of the time. When power was reported, the most commonly reported level of power was .80 (n = 201, 56.78%, for *a priori* and both types of sensitivity analyses). Other frequently adopted power levels reported were .90 (n = 25 or 7.06% for *a priori* and both types of sensitivity analyses) and .95 (n = 49 or 13.84% for *a priori* and both types of sensitivity analyses).

Effect sizes reported. Table 6 lists the frequencies of the three most common effect sizes used in *a priori* and sensitivity analyses (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 98 of 295 analyses that reported any effect size (33.22%), compared to Cohen's f and Pearson's ρ which were recorded 68 times (23.05%) and 33 times (11.86%), respectively. The most frequently utilized

effect size for Cohen's *d* was 0.50, a value labeled by Cohen as a "medium" effect size. Most commonly used Pearson's ρ values were around .20 and .30. Finally, the most common Cohen's *f* value reported was .25. 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).

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. At its best, power analysis may serve as a rough way to gauge how many participants a researcher should aim to recruit to detect an MMES. At its worst, power analysis unreasonably forces researchers to make wild guesses regarding important pieces of information (such as the MMES), encourages them to terminate recruitment even when additional participants are available, or encourages them not to run their studies at all if they are unable to reach the estimated required sample size.

In this study, we conducted a systematic review of power analysis reporting in high impact psychology journals from 2016 and 2017 to review the state of power analysis reporting. Generally, our findings suggest that there are many issues with the way power analyses are conveyed within the literature. These issues regarding power analysis reporting are approximately equivalent across the four pre-specified psychological sub-fields (social/personality, counselling/clinical/developmental, cognitive/neuropsych, and general psychology) and thus we discuss the results collapsed across sub-field.

Although the majority of power analyses recorded were *a priori* power analyses, a notable portion still analyzed power after the fact, and a large number of studies conduct

sensitivity analyses with pre-existing data sets. Therefore, although 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 stages of research with goals beyond sample planning. It is worth noting that all variations of this procedure are still impacted by the same types of problems as traditional *a priori* analysis.

Unfortunately, the minimally meaningful effect 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 may fail to detect the smallest effect that could be important within the context of their study, making the entire power analysis procedure inappropriate. This finding suggests that rather than being used to detect a specific effect of interest, 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, in press). Additionally, targeting a power analysis to find an effect size that is smaller than the MMES is a needless expense of both time and resources, since the power analysis will produce a sample size estimate larger than required. Similarly, setting the effect size higher than the MMES may cause a researcher to miss critical effects and result in reduced precision for the parameters estimated due to a smaller sample size. Although the difficulty of choosing an appropriate MMES cannot be overstated, quantifying an MMES is necessary for the proper conduct of all NHST-based power analyses. Accordingly, its lack of use may cast doubt on whether traditional power analyses are either feasible or useful.

Precision-based power analyses were almost never used as a method of sample planning within the power analyses surveyed. Our results suggest that precision-based power 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 of, "how strong" an effect or is, "how much" an intervention has caused a variable to change (Cumming, 2014).

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 just over half of studies collected more participants than required, less than 5% reported using the maximum sample size available, and that the vast majority (90.51%) of studies obtained at least 90% of the required sample. Taken together, these findings imply that either: (a) researchers responsibly obtain or can feasibly obtain the sample size required, (b) researchers may be manipulating the parameters 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 (and in each subfield) ensure that the sample size required can easily be recruited. The results found in our systematic review may have resulted from a combination of all of these factors.

However, if a researcher is unable to collect a sample within the neighbourhood of the target estimate, this does not mean a study should not be run or that the findings of the study should go into the file drawer. For researchers that work with difficult to access populations, such as those studying giftedness or the physiology of individuals who have rare disorders, small

sample size is assumed to be, “the rule rather than the exception” (Rost, 1991, p. 236). Ideally, if the study has not been run, there may be reasons to explore possible collaborations with other researchers in the relevant area. If the study has been run, then there should be a focus upon descriptive statistics rather than NHST results. Most importantly, if the study has been run, the reader should be cautious of interpreting its results on their own, but its findings should be published so that they may be meta-analyzed in the future.

Researchers did not consistently report all the information required for power analyses. Commonly, when specific parameters such as power, effect size, and α were provided, researchers tended to defer to the most commonly used standards in psychological research (i.e., a “medium” effect size, $\alpha = .05$, $1- \beta = .80$). Making the choice of these parameters arguably even less deliberate, some of these specific pieces of information may come readily-loaded into software such as G*Power, which most power analyses in our systematic review tended to use.

The state of reporting of power analyses would seem to suggest three possible options. The first possibility is that researchers are relying on these standards in order to obtain sample sizes that are feasible. This may be evidenced by choosing a medium effect size and $1- \beta = .80$ instead of trying to establish an MMES (which in many cases is likely lower than a medium effect size) or a higher power. Again, choosing a lower effect size and a higher power would lead to a higher recruitment target. Interestingly, most required sample sizes from the power analyses in the systematic review required 500 participants or less, meaning that larger sample sizes were rarely required. Another explanation to the reporting trends found in the review is that these specific values chosen for power analyses are encouraged or even required by journals or their fields of research. The final option is that certain specific values for power analyses have become so implicit within psychological statistics that they are either not mentioned or not critically

used. Values like these once again include a 1- β = .80 or a “medium” effect size. It is quite likely that all three of these possibilities affect power analysis reporting.

Limitations

There are some limitations of this study. First, we focused our systematic review on 12 high-impact Psychology journals, focusing on the years 2016 and 2017. Although these journals ranged across the various sub-fields of Psychology, our results may not extend to other journals in other areas and 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.

As mentioned, precision-based power analyses may have also been under-represented within our systematic review, because the terminology for these types of analyses may not be reflected by our search criteria. However, it is worth noting that we have done a search of precision-based power analysis terminology on Google Scholar and have not been able to locate any extensive evidence of its use in Psychology in 2016 or 2017. Specifically, searching "precision based sample size calculation" or "precision based sample calculation” in Google Scholar for 2016 and 2017 produced no journal articles published in psychology journals (but resulted in some articles from medical journals).

Recommendations

Our study provides evidence that power analyses are not consistently reported, and when they are reported they are missing important pieces of information. These findings are not surprising, given the amount of debate surrounding the utility of the power analysis procedure

itself. When using a statistical tool that has so many theoretical challenges, researchers inevitably run into difficulties when implementing it, often making the power analysis procedure unfeasible or improperly conducted. This can lead to uninformative, misleading, and potentially damaging results.

Power analyses that are precision-oriented may be a better way to estimate the required sample size, but they still require a pre-specification of a confidence interval width. However, we argue that it is much easier to specify a desired confidence interval width than it is to estimate an MMES, which constitutes a single point estimate. In other words, estimating the MMES requires knowing the center of the distribution associated with the alternate hypothesis, whereas specifying a desired confidence interval width only requires that the researcher be able to approximate the population variability. This variability can be estimated from past studies, unlike the MMES. Therefore, precision-based power analyses are not only theoretically superior to traditional NHST-based power analyses since they focus on the width of confidence intervals rather than reject/not reject decisions regarding the null hypothesis, but they are also easier to conduct.

In the unlikely scenario that a researcher knows the MMES, *a priori* power analysis may be used to ascertain whether they can achieve the appropriate range for a sample. Notably, we argue that a study should not be abandoned if researchers cannot collect participants within the reference point, since it may be meta-analyzed (e.g., combined with data from other labs around the world) in the future.

Finally, the best way to increase power within a study is not to do a power analysis, but simply to collect as many participants as possible and then replicate the study to test the validity of its findings. Conducting a power analysis does not automatically ensure high power, but when

sample planning is required, precision around an estimate is a better target than the dichotomous detection of the estimate itself. Conducting a power analysis without critically thinking about the inputs selected does not make the procedure useful, but instead contributes to the multi-faceted mess.

References

Aberson, C. L. (2015). Statistical power analysis. In R. A. Scott & S. M. Kosslyn (Eds.) *Emerging trends in the behavioral and social sciences.* Hoboken, NJ: Wiley.

Anderson, S. F., Kelley, K., & Maxwell, S. E. (2017). Sample-size planning for more accurate statistical power: A method adjusting sample effect sizes for publication bias and uncertainty. *Psychological science*, *28*(11), 1547-1562.

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J., Fiedler, K., ... & Perugini, M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, *27*(2), 108-119.

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

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.

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

Brysbaert, M. & Stevens, M. (2018) Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition, 1*(1), 1–20, doi: https://doi.org/10.5334/joc.10

Carver, R. (1978). The case against statistical significance testing. *Harvard Educational Review*, *48*(3), 378-399.

Cumming, G. (2014). The new statistics: Why and how. *Psychological science*, *25*(1), 7-29.

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

Fraley, R. C. & Marks, M. J. (2007). The null hypothesis significance testing debate and its implications for personality research. In R. W. Robins, R. C. Fraley, and R. Krueger (Eds.), *Handbook of Research Methods in Personality Psychology* (pp. 149–169). New York: Guilford Press.

Gelman, A. (in press). Post-hoc power using observed estimate of effect size is too noisy to be useful. *Annals of Surgery.*

Giner-Sorolla, R. (2018). Announcement of new policies for 2018 at JESP. Retrieved from https://www.journals.elsevier.com/journal-of-experimental-social-psychology/news/announcement-of-new-policies-for-2018-at-jesp

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.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science*, *23*(5), 524-532.

Kelley, K., & Maxwell, S. E. (2012). Sample size planning. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA handbook of research methods in psychology, Vol. 1. Foundations, planning, measures, and psychometrics* (pp. 181-202). Washington, DC, US: American Psychological Association. doi: 10.1037/13619-012

Lenth, R. V. (2001). Some practical guidelines for effective sample size determination. *The American Statistician*, *55*(3), 187-193.

Levine, M., & Ensom, M. H. (2001). Post hoc power analysis: an idea whose time has passed?. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy*, *21*(4), 405-409.

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

McHugh, M. L. (2013). The chi-square test of independence. *Biochemia medica: Biochemia medica*, *23*(2), 143-149.

McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2019). Abandon statistical significance. The American Statistician, 73(sup1), 235-245.

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

Rost, D. H. (1991). Effect strength vs. statistical significance: a warning against the danger of small samples: a comment on Gefferth and Herskovits's article “Leisure activities as predictors of giftedness”. *European Journal of High Ability, 2(*2), 236-243.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies?. *Psychological Bulletin*, *105*(2), 309-316.

Signorell, A. (2016). DescTools: Tools for descriptive statistics. *R package version 0.99*, 17.

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.

Shiskina, T., Farmus, L., & Cribbie, R. A. (2018). Testing for a lack of relationship among categorical variables. *The Quantitative Methods for Psychology, 14*(3), 167-179.

Wilkinson, L. (1999). Statistical methods in psychology journals: *Guidelines and explanations. American psychologist, 54*(8), 594-604.

Williamson, P., Hutton, J. L., Bliss, J., Blunt, J., Campbell, M. J., & Nicholson, R. (2000). Statistical review by research ethics committees. *Journal of the Royal Statistical Society A, 163*, 5–13.

Yuan, K. & Maxwell, S. E. (2005). On the post hoc power in testing mean differences. *Journal of Educational and Behavioral Statistics, 30*(2), 141–167.

Table 1

List of High Impact Peer-Reviewed Psychology Journals

Journal Number	Journal	Area	Impact Factor (2017)
1	Journal of Personality and Social Psychology	SP	5.733
2	Motivation and Emotion	SP	1.837
3	Personality and Social Psychology Bulletin	SP	2.498
4	Journal of Abnormal Psychology	CCD	4.642
5	Journal of Consulting and Clinical Psychology	CCD	4.537
6	Journal of Child Psychology and Psychiatry	CCD	6.486
7	Computers in Human Behavior	CN	3.536
8	Biological Psychology	CN	2.891
9	Journal of Cognitive Neuroscience	CN	3.468
10	Psychological Science	GEN	6.128
11	Journal of Applied Psychology	GEN	4.643
12	Journal of Experimental Psychology	GEN	4.107

Note. SP = social/personality, CCD = counselling/clinical/developmental, CN = cognitive/neuropsych, and GEN = general psychology.

Table 2

Raw Counts and Percentages for the Stage at Which Power Analysis was Conducted

	Raw Count	Percentage
Stage		
A Priori	295	66.59%
Sensitivity	109	24.60%
Effect as outcome	28	6.32%
Power as outcome	81	18.28%
Post Hoc	39	8.80%
Total	443	100.00%

Table 3

Justifications for the Effect Sizes Selected within Power Analyses

Justification	Raw Count		Percentage	
Not included	207	**230**	55.20%	**56.93%**
Included	168	**174**	44.80%	**43.07%**
Minimally Meaningful Effect	2	**2**	0.005%	**0.005%**
Prior Research	135	**138**	34.67%	**34.16%**
Pilot Study	18	**18**	4.80%	**4.46%**
Prior Power-Analysis	1	**4**	0.003%	**0.01%**
Other	12	**12**	3.20%	**2.97%**
Total	375	**404**		**100.00%**

Note. Under the raw count and percentage count columns, font without bold-type refers to power analyses that had effect size as an input (a priori and sensitivity analyses with power as outcome), whereas bold-type font refers to the power analyses that included both types of sensitivity analyses and a priori power analyses.

Table 4

Alpha Levels Utilized Within Power Analyses

Inclusion	α	Raw Count	Percentage
Not included		236	53.27%
Included		207	46.73%
	.001	3	0.007%
	.05	197	44.47%
	.01	2	0.005%
	.10	1	0.002%
	Other	4	0.009%
Total		443	100.00%

Table 5

Reported Power Levels for Sensitivity and A Priori Power Analyses

Inclusion	Power Level	Raw Count		Percentage	
Not included		42	**50**	13.00%	**12.38%**
Included		281	**354**	87.00%	**87.62%**
	.05 - .45	1	**4**	0.003%	**0.01%**
	.46 - .65	1	**6**	0.003%	**1.49%**
	.66 - .79	5	**17**	1.55%	**4.21%**
	.80	187	**201**	57.89%	**49.75%**
	.81 - .89	17	**25**	5.26%	**6.19%**
	.90	21	**25**	6.50%	**6.19%**
	.91 - .94	0	**4**	0.00%	**0.01%**
	.95	45	**49**	13.93%	**12.13%**
	≥ .96	4	**23**	1.23%	**5.69%**
Total		323	**404**	100%	**100%**

Note. Under the raw count and percentage count columns, font without bold-type refers to the power analyses that had power as an input (a priori and sensitivity analyses with effect as outcome), whereas bold-type font refers to the power analyses that included both types of sensitivity analyses and a priori power analyses.

Table 6

Reported Effect Sizes for Sensitivity and A Priori Power Analyses

Effect Size	Frequency	Effect Size	Frequency	Effect Size	Frequency
Cohen's *d*		Pearson's *ρ*		Cohen's *f*	
0.00	**0**	0.00	**0**	0.00	**0**
0.05	**0**	0.05	**0**	0.05	**0**
0.10	**0**	0.10	**1**	0.10	**2**
0.15	**2**	0.15	**3**	0.15	**8**
0.20	**10**	0.20	**10**	0.20	**6**
0.25	**1**	0.25	**3**	0.25	**23**
0.30	**6**	0.30	**7**	0.30	**8**
0.35	**9**	0.35	**2**	0.35	**3**
0.40	**9**	0.40	**4**	0.40	**11**
0.45	**3**	0.45	**1**	0.45	**3**
0.50	**28**	0.50	**0**	0.50	**2**
0.55	**5**	0.55	**1**	0.55	**1**
0.60	**6**	0.60	**1**	≥ 0.60	**1**
0.65	**3**	≥ 0.65	**0**		
0.70	**3**				
0.75	**1**				
0.80	**3**				
0.85	**3**				
0.90	**2**				
0.95	**3**				
1.00	**2**				
1.05	**2**				
1.10	**2**				
1.15	**0**				
≥ 1.20	**8**				
Total	**109**		**33**		**68**

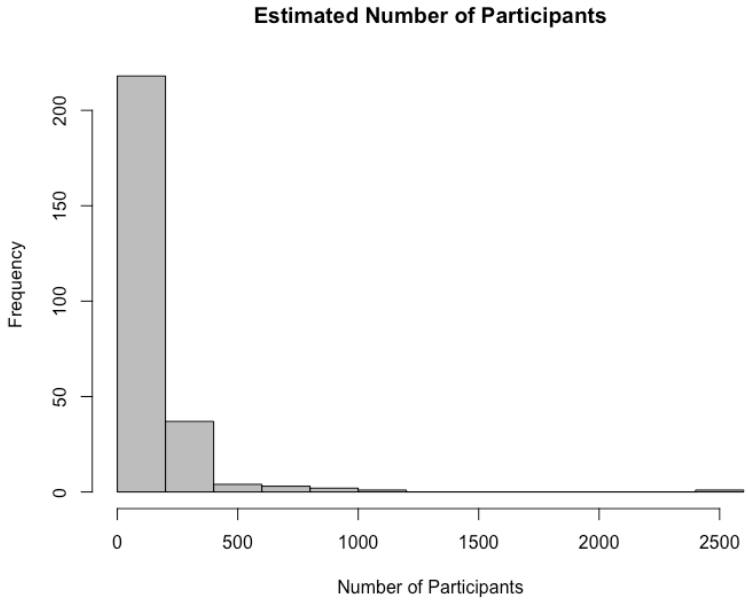

Figure 1. Number of participants that are estimated as an output in a priori power analyses (*N* = 266).

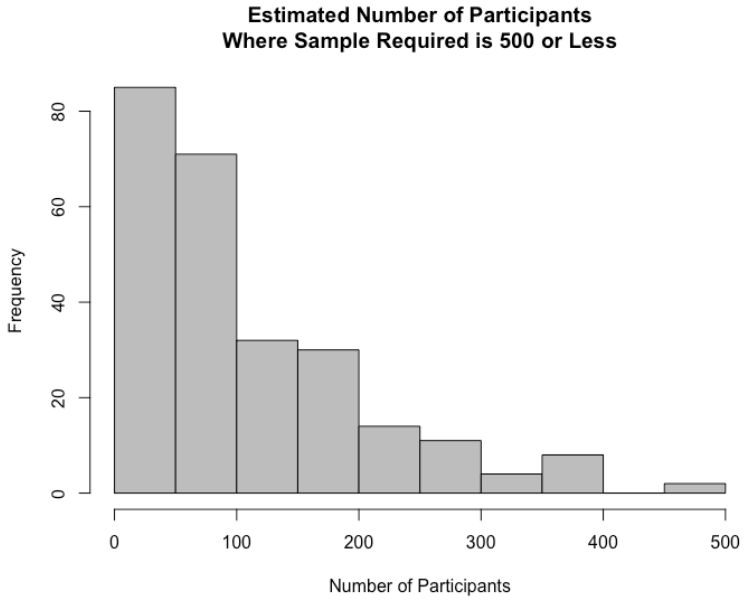

Figure 2. Number of participants that are estimated as an output in a priori power analyses (*N* = 257).

Appendix A: Subfield Reporting Consistency

To assess whether power analyses were being reported equivalently across the subfields of Psychology, an equivalence-testing approach for the chi-square test of independence was applied to how often parameters were reported across sub-field (cognitive/neuropsych, social/personality, counseling/clinical/developmental and general psychology). Specifically, the CramerV function in the DescTools package in R was utilized (Signorell et al., 2019). If the upper bound of the 90% confidence interval around Cramer's V was below .30, there was evidence of a lack of association between proportion of parameter reporting and sub-field (Shishkina, Farmus & Cribbie, 2018). Table A1 presents the proportions yes/no reporting of crucial pieces of power analysis information for a priori and sensitivity analyses (α, power, effect size, number of participants required) across psychological sub-field. For all parameters, there was no evidence of an association between parameter reporting and psychological sub-field.

Reported effect sizes, observed and desired power for post hoc power analyses. It was not possible to calculate a Cramer's V statistic for pieces of information unique to post hoc power analyses because their relatively low incidence prompted the expected frequencies of certain cells to be less than five observations (McHugh, 2013). However, across all sub-fields of psychology except social/personality, the effect sizes observed were reported more often than not (for the social/personality sub-field, observed effects were recorded for seven power analyses and were not recorded for eight). Across all four sub-fields, post hoc power was also usually reported. In contrast, desired power levels were not usually reported across sub-field.

Table A1

Proportion of Studies Reporting Parameters Across the Four Psychological Sub-Fields

	Proportion of Studies Reporting the Parameter Across Four Sub-Fields	Association Between Reporting and Sub-field? (Yes or No)	Cramer's V	90% Confidence Interval
Parameter				
α	38.0% to 67.2%	No	.20	[.10,.27]
Power	79.2% to 93.4%	No	.13	[.00, .20]
Reporting an effect size	66.7% to 77.4%	No	.08	[.00, .14]
Justifying an effect size	33.3% to 50.0%	No	.14	[.03, .22]
Reporting participants	81.5% to 98.0%	No	.18	[.04, .25]