UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

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

Methodology and Research Practice

Using and Understanding Power in Psychological Research: A Survey Study

Elizabeth Collins[1] [a], Roger Watt[1]

[1] Psychology Department, Faculty of Natural Sciences, University of Stirling, Stirling, UK

Keywords: estimation, effect size, sample size, survey, statistical power, power analysis

https://doi.org/10.1525/collabra.28250

Collabra: Psychology

Vol. 7, Issue 1, 2021

Statistical power is key to planning studies if understood and used correctly. Power is the probability of obtaining a statistically significant p-value, given a set alpha, sample size, and population effect size. The literature suggests that psychology studies are underpowered due to small sample sizes, and that researchers do not hold accurate intuitions about sensible sample sizes and associated levels of power. In this study, we surveyed 214 psychological researchers, and asked them about their experiences of using a priori power analysis, effect size estimation methods, post hoc power, and their understanding of what the term "power" actually means. Power analysis use was high, although participants reported difficulties with complex research designs, and effect size estimation. Participants also typically could not accurately define power. If psychological researchers are expected to compute a priori power analyses to plan their research, clearer educational material and guidelines should be made available.

Introduction

Statistical power is the probability of obtaining a statistically significant outcome for a test, given a particular alpha level, sample size and population effect size. The definition is often qualified by the requirement that the stated effect exists, although this extension is unnecessary when specifying the population effect size. Optimal power is widely accepted as 80%, or an 80% chance of a significant finding, based on Cohen's recommendations to balance high power with demands on the researcher to recruit enough participants (Cohen, 1992).

When research is underpowered, there is a heightened possibility of Type II errors. A Type II error is a false negative, where the null hypothesis should have been rejected but wasn't. A study with 80% power has a 20% chance of a Type II error, and as power decreases, such as when small samples are used, error rates increase. False negatives are often arbitrarily considered less problematic than Type I errors (Fiedler et al., 2012), and inattention to power means that potential new discoveries may be lost if null results are discarded.

Reviews suggest that psychological research is consistently underpowered, primarily due to insufficient sample sizes. Cohen calculated the average power of research in the 1960 volume of the Journal of Abnormal and Social Psychology to be 18%, 48% and 83% for small, medium and large effects, respectively (Cohen, 1962). As the Type II error rate is equal to $1 - power$, then Type II error rates in the literature reviewed by Cohen could be as high as 82% ($100\% - 18\%$) for research studying small effects. This is particularly important because the effects studied in psychology are often small, based on Cohen's original benchmark of $d = 0.2$ (De Boeck & Jeon, 2018). His research and subsequent textbook were some of the earliest works to encourage more consideration of power in research (Cohen, 1988). However, Cohen's work initially appeared to have little impact: indeed, in the same journal several years later, power was found to have decreased over time (Sedlmeier & Gigerenzer, 1989).

Contemporary reviews demonstrate that neither power or sample size have significantly improved across the discipline, even though bodies such as the American Psychological Association (APA) have now been encouraging researchers to consider statistical power for many years (APA, 2008). For instance, an examination of 2261 psychology papers by Szucs and Ioannidis (2017) found that the mean power for detecting small effects in psychology was 23%, 60% for medium effects, and 78% for large effects, while Stanley et al.'s (2018) review of 200 meta-analyses calculated a median overall statistical power of only 36%. Most recently, Nuijten et al. (2020) found that, in intelligence research, studies only had 11.9% power to detect small effects, along with a median sample size of just 60 participants. It appears that several decades after Cohen's first review, very few studies still come close to the widely accepted minimum power level of 80%.

a elizabeth.collins@stir.ac.uk

A Priori and Post Hoc Power

With regards to how the power of a study should actually be established, the recommended approach is to use an a priori power analysis to plan a project sample size, a process endorsed by Cohen (1988), the APA (2008) and many other organisations and individuals. This calculation uses a set alpha, intended power level, and estimated population effect size to identify the minimum sample size for a well-powered study. If the estimated population effect size is exactly correct, the given sample size then has an 80% probability of producing a statistically significant result.

Historically, the more controversial post hoc (or 'observed') power has also been used to evaluate power. This takes the form of a retrospective calculation of a study's power based on the measured effect size, alpha and actual sample size. Post hoc power analyses have traditionally been used to suggest that null results are actually Type II errors, attributed to a lack of power (Onwuegbuzie & Leech, 2004). However, using the measured sample effect size is highly unlikely to be a reliable reflection of the true population effect size due to sampling error, meaning that it should not be used in a power calculation (Gelman, 2019). In addition, using Fisher's z-transformation, we can demonstrate that post hoc power and p-values are directly related:

$$pw_{post-hoc} = 1 - normcdf\left(norminv\left(1-\frac{\alpha}{2}\right) - norminv\left(1-\frac{p_0}{2}\right)\right)$$

When p = .05, post hoc power will be 50%, regardless of the combination of sample size and sample effect size, or of the true power of the study (Lakens, 2014; Yuan & Maxwell, 2005). Null results (p > .05) will always result in low post hoc power, regardless of the actual power of the study, rendering post hoc power analyses uninformative in most circumstances.

Researchers and Power

Despite early encouragement from Cohen, a survey in the late 20th century found that only 36.1% of surveyed psychology and management academics used power analysis in any of their research (Mone et al., 1996). Another, more recent, survey of psychologists found that only 47% reported using power analysis for sample size planning; just a 10% increase since Mone et al.'s (1996) investigation two decades prior (Bakker et al., 2016). When evaluating actual behaviour instead of self-reported behaviour, power analysis use appears to remain much lower: for example, Tressoldi and Giofre (2015) found that only 2.9% of 853 psychology articles reported an a priori power analysis or discussed sample size. In addition, a review of reported power analyses revealed that they often lack detail, particularly regarding the process used to estimate effect sizes (Bakker et al., 2020).

Mone et al. (1996) also briefly examined barriers to power analysis use, with researchers reporting difficulties with software and an overall lack of knowledge about power. The research of Bakker et al. (2016) also suggests that an insufficient understanding of power is a barrier to power analysis use. In a brief knowledge test, three quarters of their sample could identify the correct definition of power when presented with a list of options. However, further testing found that most participants overestimated the power of studies investigating small effect sizes, and underestimated the sample sizes needed for studying typical effects in psychology, suggesting that psychologists have incorrect intuitions about power. Similarly, a brief study conducted by Vankov et al. (2014) found that two thirds of surveyed researchers held incorrect beliefs about sample size, and whether or not it should be increased to lead to successful future replications.

The Current Study

The literature to date has explored power and power analysis in published studies, and how psychologists evaluate power in hypothetical scenarios. Most recently, self-reported data suggests that up to half of researchers are using a priori power analyses. Our research seeks to investigate the use and understanding of power in psychological research, using a combination of quantitative and qualitative questions to capture different experiences and perspectives. Note that references to 'power analysis' throughout this work refer specifically to a priori power analyses, used to calculate suitable sample sizes.

The first objective of this research is to examine the use of power analysis in psychological research in 2020, using self-report data, to determine whether power analysis use has increased since the work of Bakker et al. (2016). This data is collected along with reasons for not using power analysis, capturing free-text responses to identify the variety of explanations that may influence the behaviour of psychological researchers. We also examine power analysis use in more detail by asking about effect size estimation, in order to judge whether effect size estimation is rigorous and as accurate as possible. This extends the more recent work of Bakker, who found that reported power analyses often lack detail about effect size estimation (Bakker et al., 2020).

The second objective of this work is to further examine knowledge of power in the psychology researcher population. In the present survey we ask participants to freely define the term 'statistical power', instead of presenting them with a multiple-choice measure, to capture all possible knowledge gaps and misconceptions. If researchers are encouraged to use power analyses, and consider power in their work, it is important to establish whether or not they understand what statistical power actually is, to know that they are calculating and evaluating it correctly.

Finally, we also investigate the prevalence of post hoc power analysis use within our sample, and ask researchers to explain in their own words why they have used it. Post hoc power is a flawed concept which typically does not measure the actual power of a study (e.g. Lakens, 2014), but it is unknown whether or not this is common knowledge in the psychology community. If post-hoc power analysis use is still high, or researchers indicate that they are still using it to (mistakenly) calculate study power, more education is required to ensure that power is well-understood and evaluated appropriately.

Method

This study received ethical approval from the General University Ethics Panel at the University of Stirling, and adhered to the Code of Human Research Ethics guidelines of the British Psychological Society (2014).

TOP Statement

Materials and data for this project are openly available at https://osf.io/ywk56/. We also confirm that we have reported our sample size determination and processes for data exclusion. This paper reports all measures except the final survey item, which asked participants to share any questions that they have about power. This resulted in nearly 100 questions and therefore will form a separate tutorial article at a later date.

Sample and Procedure

All self-identifying psychologists engaging with some degree of quantitative research were eligible to take part in the study, including doctoral students. There were no restrictions on location. Due to the exploratory nature of this project, an a priori power analysis was not a suitable approach for sample size planning. Instead, the intention was to capture a large convenience sample within a four week data collection window in 2020.

Study participation took the form of an online survey hosted on Qualtrics (2021). Participants were invited via Twitter, internal university mailing lists, and external JiscMail psychology lists. Informed consent was digitally obtained at the beginning of the survey.

Survey

Participants were asked first to explain their approaches to sample size planning in their quantitative studies. If Qualtrics did not identify the term "power analysis" in their description of sample size planning, they were then asked specifically if they had ever used power analysis in their research. Participants without any experience of power analysis were asked to explain why they had not ever used it, if they were happy to do so.

Subsequently, all participants were shown a short set of questions about the following: their experience of post-hoc power, their perceived importance of power, and how they would define power in their own words. Those participants *with* experience of power analysis were also asked about how often they use power analysis, their effect size estimation methods, and their software preferences. Brief demographic questions regarding participant job role, field, location, and engagement in any kind of open science behaviour were collected. As demographic data was only collected to establish the spread of the sample, and check for sampling bias towards open science experience due to the opportunity sampling approach, no further questions were deemed relevant to ask for the purposes of this study.

General Analysis

All analyses for this project were exploratory. Quantitative analysis took the form of descriptive quantitative analyses and explorations of demographic differences using chi-square tests. Some large tables of demographic differences, where no significant group differences were found, have been presented as a supplementary file (Additional Data S1) instead of in the main text. Quantitative analyses were computed using Jamovi (The Jamovi Project, 2021), with Figure 1 produced in R (R Core Team, 2020) using ggplot (Wickham, 2016).

Qualitative data was analysed using basic content analysis, utilising inductive or deductive methods depending on the research question. Basic content analysis is a process which codes, organises and counts qualitative data (Drisko & Maschi, 2016); in this case, the data is the free-text survey responses provided by participants. Note that inductive analysis is a bottom-up coding method, where codes are derived from the data and not from a pre-conceived list or set of expectations. Contrastingly, deductive coding is a top-down method, where a list of preconceived codes are applied to a data set.

Analysis of Definitions

To analyse definitions of power, first a deductive content analysis was used to categorise all responses that didn't write "I don't know" as either incorrect or 'shows understanding'. Incorrect definitions were characterised by describing other concepts, or making clear mistakes, such as "*the size/strength of the effect*", or "*the ability to detect an effect, given the null hypothesis is true*". The incorrect definitions were then analysed using a basic inductive content analysis to code and group mistakes made by participants.

All 'shows understanding' responses were then scored based on their inclusion of the three key elements of power as per the definition provided by Cumming, "*statistical power is the **probability** of **obtaining statistical significance** if the alternative hypothesis is true, that is, if **there really is a population effect of a stated size***" (2012, p. 322). Each definition received a point for using a term such as probability, another point for mentioning statistical significance or a similar term such as $p < .05$, and a third point for a mention of a *specified* effect or something equivalent such as "*given the alternative hypothesis is true*". For example, this definition would score three points: "*the probability of detecting a true effect of a given magnitude as significant at a given alpha level*". Scoring was deliberately strict with regards to giving points only when a definition mentioned a *specified* effect as opposed to a general effect, as power relates to a specified effect size. For example, "*detect the effect of interest*" or "*an effect of a given size*" would be acceptable, versus the more vague "*the chance of detecting an effect*".

The full analysis of definitions was completed by EC. A random 20% subset was analysed independently by RW to ensure high inter-rater reliability, both for categorising definitions and also for scoring them for mentioning the three key elements mentioned above. Cohen's kappa for categorising definitions was 0.988, and Cohen's kappa for scoring definitions was 0.920. Both of these kappa values correspond to "almost perfect" agreement as suggested by Landis and Koch (1977, p. 165).

Participants

256 participants began the survey, but 42 responses were removed for one of the following reasons: no progress past the consent page; being ineligible for participation (e.g. university undergraduates or not psychologists), or providing contradictory responses about using power analysis. This study consists of data from the remaining 214 participants, characteristics of which are displayed in Tables 1 and 2. Participants were predominantly European, along with 22 participants from the United States of America, and 11 from five other countries including Saudi Arabia and South Africa.

Results

The majority of participants in this sample indicated a belief that power is very, or somewhat important in psychological research (as shown in Table 3). Perceptions of the importance of statistical power did not differ by open science engagement (X^2 (8, n = 204) = 14.3, p = .074, V = 0.19) or job role (X^2 (24, n = 205) = 28.9, p = .225, V = 0.19). Demographic differences are presented in detail in the supplementary material file.

Part 1: A Priori Power Analysis Use

Self-reported use of a priori power analysis was high in the surveyed sample. One hundred and eighty four participants (86%) had experience of using power analysis for sample size planning, compared to 30 who had not. Of these 184 participants, 152 (71%) reported using it as a current method of sample size planning, while the other 32 participants did not report using it as a current method, but confirmed that they have previously used it at least once. Additionally, 90 of these 184 participants reported using power analysis alongside other sample size planning methods, such as convenience sampling, or following general rules of thumb for particular research designs.

The 30 participants with no experience of power analysis ranged from research assistants through to professors, with no significant differences between job roles (X^2 (12, n = 205) = 19.6, p = .075, V = 0.22). There was also no significance difference in power analysis use between those who did or did not report engaging with open science or psychological reform (X^2 (4, n = 204) = 7.04, p = .134, V = 0.13). Demographic details are presented in more depth in the supplementary materials.

Participants with experience of a priori power analysis (n = 184) were asked to estimate the frequency at which they use it, as a proportion of suitable (confirmatory hypothesis testing) studies. Eighty one participants reported using a priori power analysis 100% of the time, but the overall mean frequency was 79.1% (SD = 27.8), with a median of 90%, and mode of 100%. Estimated frequencies ranged from 9% to 100% of the time.

Software Preferences

Participants with experience of a priori power analysis indicated widespread use of G*Power (Faul et al., 2007), reported 128 times. The second most popular option was R (R Core Team, 2020) (n = 55), with the pwr (Champely, 2020) and simr (Green & MacLeod, 2016) packages mentioned most frequently. Eleven participants reported using 'online calculators' without additional detail, and other software choices, each mentioned fewer than five times, were: BrawStats, Excel, Jamovi, JASP, MATLAB, NQuery, PowerPlus, SAS, SPSS, and STATA.

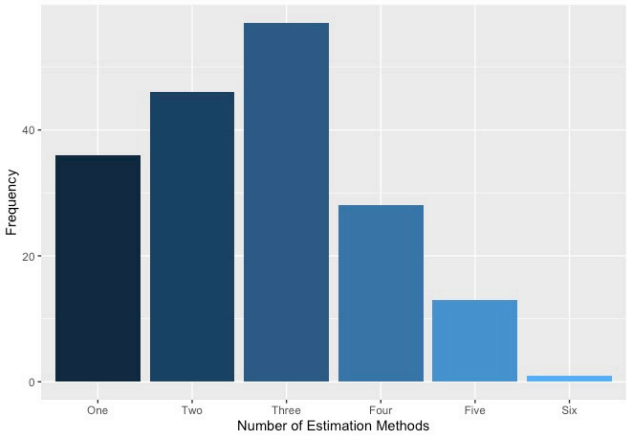

Figure 1. Graph illustrating the frequency of each number of estimation methods used by participants.

Effect Size Estimation

Methods of effect size estimation for a priori power analyses were varied, with many participants reportedly using multiple approaches. The number of times each method was selected is shown in Table 4, alongside the list of options which was presented to participants. The most frequently selected method was using an effect size from the results of other published literature, followed by using Cohen's recommendations or similar guidelines. The least popular listed option was asking for recommendations from other researchers (selected 35 times), and only 10 participants used an 'other' method.

Thirty-six participants (19.6%) reported only using one method of effect size estimation (as shown in the 'Exclusive Use' column of Table 4). The majority of participants reported using more than one method for effect size estimation, using a median of three approaches (as shown in Figure 1). No participants selected all seven options.

Not Using A Priori Power Analysis

Despite overall high use of power analysis in this sample, a large proportion of participants reported not using a priori power analyses for all suitable studies (n = 103), along with the 30 participants who reported not using it at all. Their explanations for not (or not always) using a priori power analyses are reported, with accompanying frequencies, in Table 5. It should be noted that 12 of these participants clarified that they were taking into account historic behaviour and *do* actually use an a priori power analysis for 100% of suitable recent and future studies, providing explanations such as "*[I] include studies I've done pre-replication cri-*

Table 1. Open Science Engagement, Job Role and Location Reported by Survey Participants (n=214)

Demographic Categories	Frequency	
	n	*%*
Open Science Engagement		
Yes	117	*54.67*
No	85	*39.72*
Prefer not to say	3	*1.40*
Missing	*10*	*4.67*
Job Role[a]		
Research or Teaching Assistant (no PhD)	7	*3.27*
MSc Student	2	*0.94*
PhD Student or equivalent trainee	102	*47.66*
Postdoctoral Researcher	23	*10.75*
Lecturer or Senior Lecturer	52	*24.30*
Professor	15	*7.01*
Other[b]	4	*1.87*
Prefer not to say	0	*0*
Missing	*9*	*4.21*
Location		
Australia	3	*1.40*
Belgium	1	*0.47*
Canada	3	*1.40*
Denmark	1	*0.47*
Finland	1	*0.47*
Germany	6	*2.80*
Ireland	3	*1.40*
The Netherlands	7	*3.27*
New Zealand	2	*0.94*
Saudi Arabia	1	*0.47*
South Africa	2	*0.94*
Sweden	4	*1.87*
United Kingdom	146	*68.22*
England	*81*	*37.85*
Northern Ireland	*1*	*0.47*
Scotland	*56*	*26.17*
Wales	*7*	*3.27*
"UK" [c]	*1*	*0.47*
United States of America	22	*10.28*
Missing	*12*	*5.61*

[a] Jobs were provided as a list of UK roles, with additional detail about cultural differences. Participants were asked to choose the job title that best applied to their position, to account for international roles. Note that in many UK institutions, 'professor' is often the most senior academic job position that can be held.
[b] Four other jobs were: assistant psychologist, data scientist, health improvement officer, and trainee clinical psychologist.
[c] One participant wrote 'UK' instead of providing a devolved nation.

sis".

Both groups of participants had several explanations in common, such as not using an a priori power analysis because it would suggest unrealistic large sample sizes, being negatively influenced by colleagues, and struggling with power analyses for complex study designs such as multilevel models. Several participants commented on using other rules and approaches to sample size planning such as "*[I] knew that as long as I met Tabachnick's and Fidell's rule then I'd be ok*", and the less mathematical "*if the sample size is larger than in comparable studies in peer reviewed journals, then I assume that I am safe*". Other explanations were tied

Table 2. Sub-Fields of Psychology Represented in the Sample (n=214)

Field	Frequency	
	n	*%*
Behavioural	2	*0.94*
Clinical	16	*7.48*
Cognition	35	*16.36*
Comparative	6	*2.80*
Counselling	3	*1.40*
Cyberpsychology	2	*0.94*
Developmental	15	*7.01*
Educational	4	*1.87*
Evolutionary	3	*1.40*
Experimental	3	*1.40*
Forensic	7	*3.27*
Health	34	*15.89*
Mental Health	3	*1.40*
Mathematical	5	*2.34*
Metascience	3	*1.40*
Neuropsychology	12	*5.61*
Occupational	2	*0.94*
Personality	3	*1.40*
Social	25	*11.68*
Other [a]	14	*7.48*
Missing	*17*	*7.94*

[a] Other fields represented in this sample were: affective psychology, applied psychology, autism research, biopsychology, consumer psychology, cross-cultural psychology, decision science, environmental psychology, legal psychology, moral psychology, music psychology, psycholinguistics, sexology, sports psychology.

Table 3. Response Frequencies for Importance of Power (Full Sample)

Response	Frequency	
	n	*%*
Very important	127	*59.35*
Somewhat important	66	*30.84*
Not very important	5	*2.34*
Not important at all	1	*0.47*
I don't know	7	*3.27*
Missing	*8*	*3.74*

directly to the calculation itself, most frequently reporting difficulty with effect size estimation, or commenting that power analysis is too difficult (or impossible) for complex statistical designs.

The "other" category in Table 5 represents a wide range of responses from participants not using power analysis in 100% of suitable studies, including not using power analyses when working with students, not using a power analysis for direct replications, and preferring to use sensitivity analyses. One participant offered a particularly critical perspective on the use of the power analysis for the sake of journal guidelines, as shown in the quote below:

> *"The poor understanding of power among co-authors and reviewers is a punishers* [sic] *for me to do power analyses well. I've had multiple situations where were people are satisfied with seeing "a" power analysis even though it's wrong. Doing it right can take a lot of effort, and honestly sometimes I wonder why I'm bothering*".

One other participant, who reported never using power

Table 4. Frequencies for Each Effect Size Estimation Method Including Number of Participants Exclusively Using Each Approach, with 'Other' Methods Reported by Participants.

	Method	Frequency	*Exclusive Use*
1	Use an effect size from the results of other published literature	122	*9*
2	Use the same effect size as a previous similar study reported in their methods	83	*2*
3	Use a small or medium effect size e.g. Cohen's recommendations	106	*16*
4	Use recommendations from other researchers	35	*2*
5	Use the smallest effect size of interest for my field or "meaningful" effect size for my field	79	*4*
6	Run a pilot study to calculate an effect size first	47	*0*
7	Other	10	*3*
	Relying on statisticians to decide	*2*	
	Using scaled-down estimates to account for publication bias	*2*	
	Taking into account sensitivity analyses	*3*	
	Relying on personal unpublished work	*1*	
	Using a personally meaningful effect size	*1*	
	No explanation given	*1*	

Table 5. Reasons Why Participants Don't, or Don't Always Use A Priori Power Analysis

Reason	Frequency	
	Power Analysis Use	
	Never	Not Always
Don't know enough about power	6	1
Power analysis is difficult to do	-	4
Unsure about effect size estimation	-	12
Don't have enough information to do a power analysis	-	4
Power analysis is too difficult for complex statistical designs	1	14
Produces unrealistic sample sizes	3	16
Influenced by colleagues	3	6
Influenced by time pressure	-	3
Not needed (no explanation)	3	-
Not needed (not applicable to work)	7	1
Not needed (access to large samples)	1	2
Use other rules and approaches to sample size planning	2	11
Rely on statisticians	1	1
Reflecting on historic behaviour	-	12
Choose not to	1	4
Other	1	6

analysis, also criticised the inherent relationship between p-values and power:

> *"It strikes me that Power analysis is a way of finding what would be a significant value (seeking a p-value)".*

Part 2: Experience of Post Hoc Power

All survey participants, regardless of a priori power analysis experience, were asked if they had ever used post hoc power analysis. They were then asked why they had used it if they responded *yes*. Frequencies of post hoc power analysis use are presented in Table 6, divided into experience of a priori power analysis (*yes* or *no*). Five participants who reported never using a priori power analysis reported that they had experience using post hoc power analysis.

Of the 97 participants with experience of post hoc power analysis, 86 provided one or more reasons explaining why they have used the calculation. The most common explanation was simply to check the actual power of a study (e.g. "*to prove that the research was well powered*"), which

Table 6. Experience of Post Hoc Power Analysis, Divided by Experience of A Priori Power Analysis (Yes or No)

Sample Group	Used Post Hoc Power Analysis? *Frequency*		
	Yes	**No**	**Missing**
Full Sample	97 *(45.3%)*	110 *(51.4%)*	7 *(3.3%)*
A Priori - Yes	92 *(50%)*	85 *(46.2%)*	7 *(3.8%)*
A Priori - No	5 *(16.7%)*	25 *(83.3%)*	0 -

Table 7. Explanations for Using Post Hoc Power Analysis

Reason	**Frequency**
Historic behaviour	11
For educational purposes	3
Personal curiosity	6
Required to do so (publishing or exams)	14
Check actual power	53
- general	*11*
- due to null results	*6*
- due to underrecruiting participants	*7*
- due to secondary data	*3*
- due to unexpectedly small effect sizes	*2*
- due to not calculating an a priori power analysis	*5*
- after changing study designs during research	*4*
- in order to demonstrate reliability of findings	*8*
- in order to plan larger future studies	*6*
Calculated for the purpose of a meta-analysis	2
Reproduce calculations when reviewing	3
Other	5

demonstrates that there are still widely-held misconceptions about post-hoc power analysis. A more detailed breakdown of explanations is presented in Table 7. The "other" category spans a variety of explanations, including "*when reading the lit other studies that seem rigorous do so*" and "*it is not always clear what the right power analysis is*".

Reassuringly, eleven participants explained their use of post hoc power analysis as being historic behaviour, with several explaining that they had used it before learning about the statistical issues associated with post hoc power. This more knowledgeable perspective was shared by the two participants who mentioned using post hoc power for educational purposes, for instance: *"to demonstrate (using a simulation) to students how crazily it bounces around with replication"*. One participant, who explained that they had used post hoc power to satisfy a reviewer, also commented that doing so went against their personal preferences and that they were aware that it is a nonsensical calculation.

It should also be noted that five participants, not included in Table 7, answered 'yes' to using post hoc power analysis, but their explanations indicated that they actually had used a sensitivity analysis, such as "*tested what the minimum effect I could have detected with my sample size is*".

Part 3: Defining the Concept of Statistical Power

All participants were asked to define power in their own words, or to write 'I don't know' if preferred. A content analysis of responses identified 57 as *incorrect*, 135 as *shows understanding*, and the remaining 13 were cases of participants stating "I don't know". These results are shown in Table 8, subdivided by a priori power analysis experience (*yes* or *no*). Of the 57 incorrect definitions, 40 were provided by participants with experience of power analysis.

Incorrect Definitions of Power

Several common mistakes emerged in the definitions of power given by participants, full details of which can be found in the supplementary materials. Some participants clearly defined other statistical concepts, such as incorrectly describing an effect size instead of power; or describing a power analysis instead of power itself. Three partici-

Table 8. Categorisation of Definitions of Power, Divided by Experience of A Priori Power Analysis (Yes or No)

Sample Group	Definition Category *Frequency*			
	Shows Understanding	Incorrect	I don't know	Missing
Full Sample	135 *(63.1%)*	57 *(26.6%)*	13 *(6.1%)*	9 *(4.2%)*
A Priori - Yes	125 *(67.9%)*	40 *(21.7%)*	10 *(5.4%)*	9 *(4.9%)*
A Priori - No	10 *(33.3%)*	17 *(56.7%)*	3 *(10%)*	0 -

Table 9. Scores for Definitions Rated as Shows Understanding

Score out of Three	Zero	One	Two	Three
Frequency	2 *(1%)*	51 *(38%)*	65 *(48%)*	17 *(13%)*

pants also confused power and Type I errors, defining power as "*the likelihood any significant effect is not due to chance*" or similar, while seven participants mistakenly described power as the Type II error rate. Many participants also made incorrect comments about power being a measure of meaningfulness, representativeness, or validity.

Definitions Rated as "Shows Understanding"

Scores out of three were calculated for all 135 definitions rated as 'shows understanding', as shown in Table 9. Most commonly, participants scored two out of three. Scoring was deliberately strict for mentioning a specified effect as opposed to a general effect, as power relates to a specified effect size.

Seventeen definitions scored three out of three. Examples scored this way include "*the probability of finding a significant effect according to null hypothesis significance testing, given a stated effect size*" and "*the probability that p will be <.05 assuming the alternate is true and a certain effect size, with a given n*".

The two definitions that scored zero but were not classed as incorrect were categorised this way because they indicated some understanding that power relates to the chance of identifying an exist if it exists, but did not quite mention any of the three key elements. For instance, one definition ("*I define power as whether or not my study has the power to detect an effect in the data, if there is an effect to be detected at all*") mentioned *an* effect instead of a specified effect, and uses binary 'whether or not' language instead of referencing probability or a sensible synonym. This resulted in a zero score but categorisation as 'shows understanding'.

Initial analysis identified 96 out of 135 definitions which directly used the word 'probability' or presented a definition in the format *1 – Type II error* rate, indirectly indicating probability. However, due to the prevalence of other similar terms such as ability and capability, the criteria 'mentions probability' was expanded to include the mention of associated terms, increasing the frequency to 126 out of 135. All definitions mentioning a similar term were scored as mentioning probability.

Sixty-six out of 135 definitions mentioned statistical significance, or provided similar descriptions such as correctly rejecting a false null hypothesis, or power being associated with a set alpha. Example definitions which were scored as mentioning statistical significance (or describing the same concept) include "*the chance of an effect to be detected (according to a set alpha) given the effect is true*" and "*the ability to detect a (statistically significant) effect*".

Only 37 out of 135 definitions mentioned a particular effect (as opposed to using general language about effects existing) and therefore were scored as correctly describing this third element of power. For instance, a definition such as "*the probability of detecting the effect you have predicted, assuming that is the true effect*" correctly refers to a specific effect size and was scored as mentioning this element; compared to "*the ability to detect a (statistically significant) effect*", which only refers to an unspecified effect. However, taking into account all mentions of 'an effect' and similarly vague terms, 101 participants made some reference to effects, and two more mentioned finding a '*significant difference*'. Further analysis looking for mentions of an effect 'truly existing', or other similar language, found that 75 definitions clearly stated that an effect needed to exist or was real, such as "*the ability to identify an effect if an effect truly exists in the population*". An additional 15 participants referenced a 'true alternative hypothesis' or 'false null hypothesis'.

Discussion

As psychological research is historically underpowered (e.g. Stanley et al., 2018), there are widespread efforts to encourage or enforce reflections on power in all appropriate published research, such as the APA recommending that all studies conduct and then report power analyses (APA, 2008; Appelbaum et al., 2018). However, recent research demonstrates that researchers lack intuitions about power (Bakker et al., 2016) and that guidelines do not necessarily result in power analyses being computed correctly (Bakker et al., 2020).

Within the current study, the majority of our participants

perceived power to be somewhat or very important in psychological research. We found much higher self-reported power analysis use than previous studies (e.g. Bakker et al., 2016). In this study, 152 participants (71%) explicitly mentioned using power analysis as a sample size planning method, and a further 32 (15%) confirmed they have previously used it at least once. However, similarly to Bakker et al., 90 participants reported using power analysis as one of multiple approaches to sample size planning, alongside other methods such as convenience sampling or using sample sizes similar to those in other published studies. Only 81 participants reported using power analysis for all suitable study planning.

Several barriers to using power analysis emerged from this study. Most importantly, participants reported difficulties with successfully calculating power. For instance, several struggled with power analysis for more complex study designs such as mixed models, which typically require programming skills that are difficult or time consuming. Less than 1/3 of participants reported using R for power analysis, suggesting that there is not yet widespread familiarity with using programming for power (which would enable more complex power calculations). This is an important line of future research to ensure that power analysis is an accessible procedure for all researchers, using any study design.

Other participants noted that effect size estimation is a difficult process, which may explain why participants do not consistently use just one approach to effect size estimation (as shown in Figure 1). Cohen (1992) provided guidelines to make estimation easier, which appear to still be popular – 106 participants reported using his, or similar, guidelines, 16 of whom use them exclusively. However, these guidelines have been criticised for lacking specificity and relevance to each field (e.g. Correll et al., 2020), and the high use found here indicates that this knowledge is not yet widespread. The only estimation approach used more frequently was to take effect sizes from the results of previous literature. This is also problematic, due to the likely overestimation of effect sizes in studies which have used small samples but found statistically significant p-values. This is principally attributed to uncorrected publication bias and questionable research practices (see Simmons et al., 2011; Smaldino & McElreath, 2016). If inflated effect sizes are used in future power analyses, suggested sample sizes will remain smaller than necessary, and Type II error rates are unlikely to decrease. Our findings appear to align with recent research by Bakker et al. (2020), who found that many reported power analyses mentioned relying on Cohen's guidelines, previous literature, or simply did not provide any detail at all.

Post Hoc Power Analysis

In our sample, 46% of participants have used post hoc power analyses, many of whom explained they had done so in order to calculate 'actual power'. This is only safe if the measured sample is a very close representation of the actual population of interest, which is highly unlikely due to sampling error. Post hoc power is a very poor estimate of actual study power, and will always be low when $p > .05$ (Lakens, 2014), as discussed in the Introduction to this paper. With nearly one quarter of our participants indicating a mistaken belief that post hoc power equals actual study power, we suggest that more education is certainly needed. More concerningly, several participants were asked to calculate post hoc power by reviewers, which should be discouraged, if not outright banned, by editors. It is apparent that many researchers in psychology have not yet discovered that post hoc power analyses are generally uninformative and should be avoided.

Conversely, several participants confirmed use of post hoc power analysis, but then went on to provide an explanation that clearly indicated they were describing a sensitivity analysis instead. This is another, more statistically acceptable, retrospective calculation which establishes the smallest effect size that could have been reliably detected, using the actual sample size of a study (Perugini et al., 2018). Given that the present study sampled many participants who report engaging with open science or psychological reform, where the use of sensitivity analyses is growing, we believe that a proportion of our participants may have interpreted post hoc power analysis as including sensitivity analysis.

Participants' Understanding of Power

More than 1/4 of participants incorrectly defined power when asked, demonstrating confusion with other concepts such as effect sizes and Type I/II errors. Of the 135 participants who offered a definition that demonstrated some understanding of power, just 66 participants mentioned statistical significance or a similar term, suggesting that there is a lack of awareness that power is a frequentist concept that is mathematically tied to null hypothesis significance testing. In addition, while more than 100 participants mentioned 'an effect' in some form, only 37 of these referenced power being related to a specific effect (as opposed to simply 'finding any effect'). Typically, participants indicated some awareness of power as a concept, but were unlikely to provide a clear and accurate definition, which suggests that understanding of power is somewhat limited. If researchers only have an insecure understanding of what power actually is, they cannot be expected to successfully calculate and report power analysis, or critically evaluate the power of their study as requested by the APA (APA, 2008).

Sample Limitations

The use of power analysis in sample size planning is highly likely to be overestimated in this sample due to the nature of convenience sampling and the demographic features of the sample. In comparison to the 47% of participants who reported using power analysis in Bakker et al. (2016), self-reported power analysis use is unexpectedly high. We attribute this, in part, to the high proportion of participants in this study who engage with some aspect of psychological reform or open science. These participants may be more likely to think critically about power and sample size, and adopt behaviours such as power analyses. The use of Twitter as a sampling approach is likely to explain the high proportion of these participants, as there is a strong community of open science-minded psychologists using Twitter. The sample also heavily features PhD students and early career researchers, who are more likely to have only

been involved in psychology since the replication crisis and subsequent statistical reform period, and hence have been exposed to discussions of power and sample size throughout the majority of their careers.

We also anticipate that asking participants to self-report behaviour is likely to result in a bias towards reporting power analysis use, particularly when compared to actual behaviour. For example, Tressoldi & Giofré (2015) found power analysis reporting to be as low as 3%, and it is unlikely that, just a few years later, true rates of power analysis use in the wider psychology community are as high as the 71% measured in the present study.

Conclusions

We believe that a larger and more representative sample would show lower use of power analysis, more difficulties with effect size estimation, and reduced knowledge about power. Further reviews of power analysis reporting in the literature are also needed, to enable ongoing comparisons between self-report and actual behaviour. It is also important to consider that there is also no guarantee that participants are using power analysis correctly even if use is increasing, which is a sentiment shared explicitly by one participant in this study who commented that any power analysis seems to be good enough to satisfy reviewers.

Our findings support the belief that a lack of understanding about power, what it is and how it is calculated, may be important barriers to successful adoption of power analysis use. Bakker et al. (2020) demonstrate that even when guidelines exist, power analysis reporting is insufficient and effect size estimation is unclear. We agree that guidelines are insufficient to improve the quality of psychological science. Clear tutorials, examples and templates should exist, particularly with regards to effect size estimation, which should be tailored to each research area. Journals should also consider supporting interactive power analysis web applications (e.g. Shiny apps), as point and click software options are more accessible than programming languages for many researchers. Web applications can be designed to make complex, model-based power analyses easier to calculate. Furthermore, psychologists, and indeed all researchers, should also be made more aware of the pitfalls of observed power and how it does not indicate the 'actual' power of a study.

Contributions

Contributed to conception and design: EC, RW
Contributed to acquisition of data: EC
Contributed to analysis and interpretation of data: EC, RW
Drafted and/or revised the article: EC, RW
Approved the submitted version for publication: EC, RW

Acknowledgements

We would like to thank our editor, Professor Chris Aberson, and our anonymous reviewer for their feedback on this manuscript. We also thank Dr Jordan Miller, for proof reading and giving feedback on both our initial and revised work.

Funding

This research forms part of a PhD funded by the Economic and Social Research Council.

Competing Interests

The authors declare no competing interests.

Supplemental Material

Additional Results S1: This file contains two tables which present the demographic differences for power analysis importance and power analysis use. This file also contains a third table detailing mistakes made in incorrect definitions of power.

Data Accessibility Statement

All data, study materials, and a list of analyses conducted are available on the Open Science Framework page associated with this paper: https://osf.io/ywk56/.

Submitted: May 07, 2021 PDT, Accepted: September 15, 2021 PDT

REFERENCES

APA. (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist*, *63*(9), 839–851. https://doi.org/10.1037/0003-066x.63.9.839

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *American Psychologist*, *73*(1), 3–25. https://doi.org/10.1037/amp0000191

Bakker, M., Hartgerink, C. H. J., Wicherts, J. M., & van der Maas, H. L. J. (2016). Researchers' intuitions about power in psychological research. *Psychological Science*, *27*(8), 1069–1077. https://doi.org/10.1177/0956797616647519

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

British Psychological Society. (2014). *Code of human research ethics*. British Psychological Society. https://www.bps.org.uk/news-and-policy/bps-code-human-research-ethics-2nd-edition-2014

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

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

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Lawrence Erlbaum Associates (Routledge). https://doi.org/10.4324/9780203771587

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Correll, J., Mellinger, C., McClelland, G. H., & Judd, C. M. (2020). Avoid Cohen's 'small', 'medium', and 'large' for power analysis. *Trends in Cognitive Sciences*, *24*(3), 200–207. https://doi.org/10.1016/j.tics.2019.12.009

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge. https://doi.org/10.4324/9780203807002

De Boeck, P., & Jeon, M. (2018). Perceived crisis and reforms: Issues, explanations, and remedies. *Psychological Bulletin*, *144*(7), 757–777. https://doi.org/10.1037/bul0000154

Drisko, J. W., & Maschi, T. (2016). *Content analysis*. Pocket Guides to Social Work R; Oxford University Press. https://doi.org/10.1093/acprof:oso/9780190215491.001.0001

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/bf03193146

Fiedler, K., Kutzner, F., & Krueger, J. I. (2012). The long way from α-error control to validity proper: Problems with a short-sighted false-positive debate. *Perspectives on Psychological Science*, *7*(6), 661–669. https://doi.org/10.1177/1745691612462587

Gelman, A. (2019). Don't calculate post-hoc power using observed estimate of effect size. *Annals of Surgery*, *269*(1), e9–e10. https://doi.org/10.1097/sla.0000000000002908

Green, P., & MacLeod, C. J. (2016). simr: an R package for power analysis of generalised linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493–498. https://doi.org/10.1111/2041-210x.12504

Lakens, D. (2014, December 19). *Observed power, and what to do if your editor asks for post-hoc power analyses*. http://daniellakens.blogspot.com/2014/12/observed-power-and-what-to-do-if-your.html

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

Mone, M. A., Mueller, G. C., & Mauland, W. (1996). The perceptions and usage of statistical power in applied psychology and management research. *Personnel Psychology*, *49*(1), 103–120. https://doi.org/10.1111/j.1744-6570.1996.tb01793.x

Nuijten, M. B., van Assen, M. A. L. M., Augusteijn, H. E. M., Crompvoets, E. A. V., & Wicherts, J. M. (2020). Effect sizes, power, and biases in intelligence research: A meta-meta-analysis. *Journal of Intelligence*, *8*(4), 36. https://doi.org/10.3390/jintelligence8040036

Onwuegbuzie, A. J., & Leech, N. L. (2004). Post hoc power: A concept whose time has come. *Understanding Statistics*, *3*(4), 201–230. https://doi.org/10.1207/s15328031us0304_1

Perugini, M., Gallucci, M., & Costantini, G. (2018). A practical primer to power analysis for simple experimental designs. *International Review of Social Psychology*, *31*(1). https://doi.org/10.5334/irsp.181

Qualtrics. (2021). [Software]. https://www.qualtrics.com

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Sedlmeier, P., & Gigerenzer, G. (1989). Do Studies of Statistical Power Have an Effect on the Power of Studies? *Psychological Bulletin*, *105*(2), 309–316. https://doi.org/10.1037/10109-032

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*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, *3*(9), 160384. https://doi.org/10.1098/rsos.160384

Stanley, T. D., Carter, E. C., & Doucouliagos, H. (2018). What meta-analyses reveal about the replicability of psychological research. *Psychological Bulletin*, *144*(12), 1325–1346. https://doi.org/10.1037/bul0000169

Szucs, D., & Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biology*, *15*(3), e2000797. https://doi.org/10.1371/journal.pbio.2000797

The Jamovi Project. (2021). *Jamovi (Version 1.6)* [Computer Software]. https://www.jamovi.org

Tressoldi, P. E., & Giofré, D. (2015). The pervasive avoidance of prospective statistical power: Major consequences and practical solutions. *Frontiers in Psychology*, *6*, 726. https://doi.org/10.3389/fpsyg.2015.00726

Vankov, I., Bowers, J., & Munafò, M. R. (2014). Article commentary: On the persistence of low power in psychological science. *Quarterly Journal of Experimental Psychology*, *67*(5), 1037–1040. https://doi.org/10.1080/17470218.2014.885986

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Yuan, K.-H., & Maxwell, S. (2005). On the post hoc power in testing mean differences. *Journal of Educational and Behavioral Statistics*, *30*(2), 141–167. https://doi.org/10.3102/10769986030002141

SUPPLEMENTARY MATERIALS

Peer Review History

Download: https://collabra.scholasticahq.com/article/28250-using-and-understanding-power-in-psychological-research-a-survey-study/attachment/72237.docx?auth_token=tsqGVMD4O9XZSCyV7Hpg

Supplementary Material

Download: https://collabra.scholasticahq.com/article/28250-using-and-understanding-power-in-psychological-research-a-survey-study/attachment/72238.docx?auth_token=tsqGVMD4O9XZSCyV7Hpg