


## STUDY PROTOCOL

### **An evaluation of reproducibility and errors in statistical power calculations performed using G\*Power**


Robert T. Thibault<sup>\*1,2,3</sup> ( 0000-0002-6561-3962)

Emmanuel A. Zavalis<sup>1,4</sup> ( 0000-0001-6205-1362)

Mario Malički<sup>1</sup> ( 0000-0003-0698-1930)

Steven Goodman<sup>1</sup> ( 0000-0002-3872-5723)

Marcus R. Munafò<sup>2,3</sup> ( 0000-0002-4049-993X)

Hugo Pedder<sup>5</sup> ( 0000-0002-7813-3749)

Projected contributor roles according to the Contributor Roles Taxonomy (CRediT) are detailed in Appendix B

\*Address correspondence to [robert.thibault@stanford.edu](mailto:robert.thibault@stanford.edu)

<sup>1</sup>Meta-Research Innovation Center at Stanford (METRICS), Stanford University. <sup>2</sup>School of Psychological Science, University of Bristol. <sup>3</sup>MRC Integrative Epidemiology Unit at the University of Bristol. <sup>4</sup>Department of Learning, Informatics, Management and Ethics, Karolinska Institutet. <sup>5</sup>Department of Population Health Sciences, Bristol Medical School, University of Bristol

## Background

Power calculations can be used to identify the sample size required to test a specific hypothesis with a predetermined level of certainty of rejecting the null hypothesis, if the assumptions used in the power calculation are correct. When used effectively, power calculations can help constrain the occurrence of both false positives and false negatives. On the other hand, when used without a clear understanding, they can mislead us and increase our confidence in false findings.

Despite the substantial attention low statistical power has received in the life and health sciences over the past decade, this issue continues to undermine research efforts (e.g., Border et al., 2019; Button et al., 2013; Lakens, 2022; Marek et al., 2022; Smaldino & McElreath, 2016). Several initiatives encourage the use of appropriate sample sizes. For example, the CONSORT statement (Moher et al., 2012: item 7a), STROBE statement (von Elm et al., 2008: item 10), ARRIVE guidelines (Percie du Sert et al., 2020: item 2b) transparency checklist for social and behavioural research (Aczel et al., 2020: item 4 of 12), discipline specific reporting statements (e.g., Ros et al., 2020: item 1b), preregistration templates (OSF, 2016), journal reporting checklists (e.g., Nature Publishing Group, 2019), and the Experimental Design Assistant (Percie Du Sert et al., 2017) all ask researchers to justify their chosen sample size. At least two studies analyzed the impacts of these policies at journals (Carter et al., 2017; The NPQIP Collaborative Group, 2019). They found that, after the journals requested sample size justifications, more articles in those journals commented on sample size, but formal sample size calculations remained uncommon (e.g., a power or precision calculation).

Of the published studies that do use power calculations, many may be irreproducible or contain errors (e.g., Charles et al., 2009; Clark et al., 2013; Rutterford et al., 2015). Researchers can perform power calculations “by hand” using statistical formulae, with programming languages such as R or Python, or with graphical user interfaces such as G\*Power or dedicated websites. Anecdotally, one of us (RTT) has reviewed manuscripts that include irreproducible or erroneous power calculations performed in G\*Power, and has also been a co-author on a paper with such an error (explained in Thibault & Pedder, 2022). G\*Power has provided a much-needed platform for non-statisticians to conduct power calculations for over a decade and likely helped raise the importance of statistical power. And yet, because this type of graphical user interface can be used without a thorough understanding, it can lead to erroneous calculations. Given that thousands of published articles use G\*Power, the present study sets out to estimate the prevalence of irreproducible and erroneous power calculations performed in G\*Power.

This question is of particular interest to us because, if irreproducibility and errors are common, they could be remedied by further tailoring the G\*Power software to its user base. For example, to encourage reproducibility, the software could prompt users to take

a screenshot of their calculation in the GUI and include this image as supplementary material linked to their publication. If these issues exist and are not remedied, then G\*Power users may have false confidence that they are conducting appropriately powered research when they are not. As more organizations create policies that require researchers—many of whom are unfamiliar with sample size calculations—to perform said calculations, we imagine the use of such graphical user interfaces will continue to increase.

## **Study objectives**

This study is descriptive; we have no hypotheses, but we have specific objectives. These objectives are further operationalized in the methods section of this protocol.

1. Estimate the number of published articles that use G\*Power for a power calculation.
2. Assess the reproducibility of power calculations that use G\*Power in published articles.
3. Estimate the frequency of error-free power calculations that use G\*Power in published articles, including the appropriate use of the options G\*Power provides for ANOVAs (see Figure 1).

## **Sample Size**

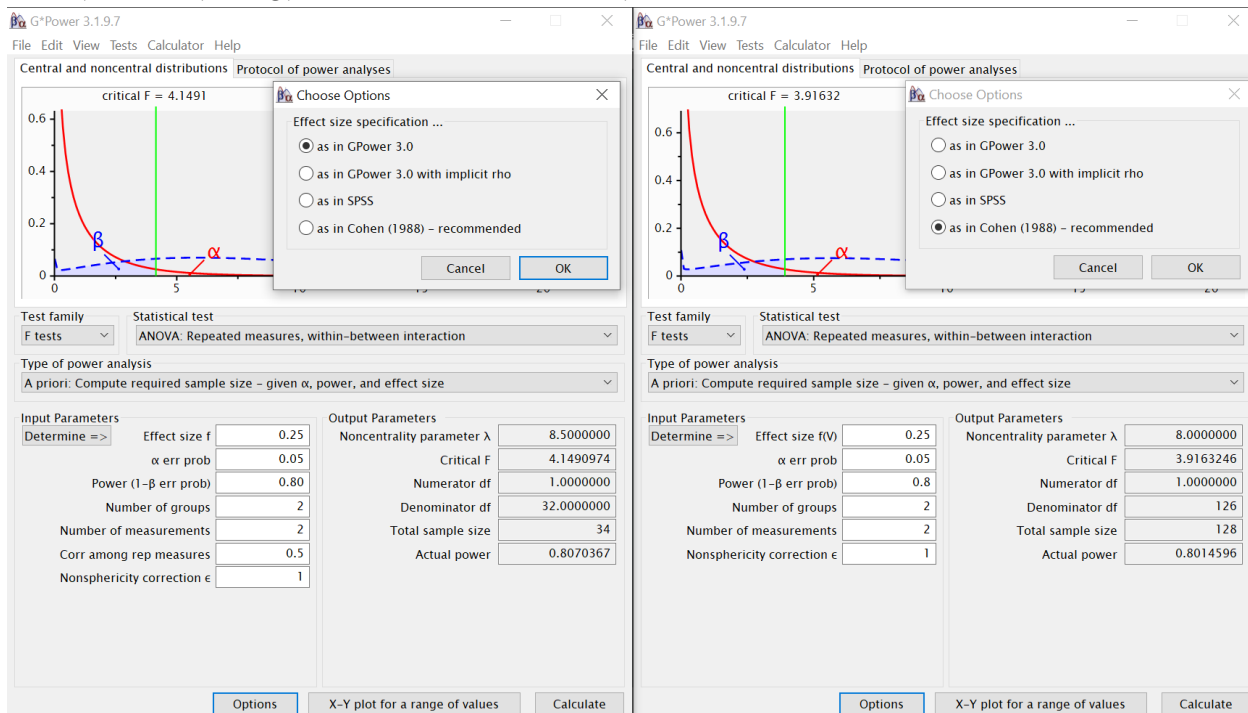
Rather than test a hypothesis, we aim to estimate the prevalence of reproducible power calculations and error-free power calculations that use G\*Power. Thus, we use a precision calculation rather than a power calculation to inform our sample size (Rothman & Greenland, 2018). We perform a precision analysis using Monte Carlo sampling for 95% confidence intervals (outlined in the protocol code available at [osf.io/dsv4m](https://osf.io/dsv4m)). This analysis returns very similar sample sizes to the equation  $n = (z^2 * p(1 - p)) / MOE^2$ , which is often used for precision analyses. This more common equation, however, relies on a normal approximation to the binomial distribution, which doesn't hold true for small sample sizes or proportions near 0 or 1.

The majority of our survey questions are binary and thus, we can use the same precision calculation for all these variables (e.g., could we reproduce the calculation, was the calculation error-free). As we will be producing estimates rather than testing hypotheses, we will not make adjustments to account for multiple comparisons. We aim to balance the time it takes to code articles with the precision of our results, which need not be highly precise for the purposes of this study. For example, if we discover that 20-40% (95% CI)

of power calculations are reproducible, this remains a problem worth addressing regardless of whether the true value is 20% or 40%. Thus, we will use a 95% confidence interval with a maximum width of 20% between the lower bound of the confidence interval to the upper bound, and corresponding to the most conservative expected proportion of 0.50, which equates to a minimum sample size of 95 articles.

One of our questions of interest within Objective 3—whether power calculations for ANOVAs clearly select the appropriate option related to their analysis (see Figure 1)—applies to only a subset of the articles we will survey. Thus, we have run an additional precision analysis for this question. We expect that about 20% of relevant articles will clearly select the appropriate option related to their analysis. This expected proportion of 0.20 equates to a sample size of 60 articles. With a sample size of 60 articles, the 95% CI could have a width of up to 26% (i.e.  $\pm 13\%$ , if the sampled proportion turns out to be 0.50), which we deem acceptable.

**Figure 1. Depiction of the G\*Power options for an ANOVA.** If users are unaware of the default option, they may run the power calculations in such a way that they account for the correlations between repeated measures a second time, and in turn substantially—but erroneously—increase power. This issue is explained in further detail by Kieslich, 2020; Lakens, 2013; and Thibault & Pedder, 2022.



## Methods

### Sample

To identify published articles that used G\*Power to run a power calculation we will use the search query (*GPower* OR “*G Power*”) AND (“2017/01/01”[*Publication Date*] : “3000”[*Publication Date*]) in PubMed Central (PMC). Note, searches in PMC convert the characters “\*” and “-” (when in the middle of the word and within quotation marks) to a space. This makes the terms “*G Power*”, “*G\*Power*”, and “*G-Power*” equivalent in the query. We will search PMC because, unlike PubMed and most other databases, PMC only indexes open access articles and search queries scan the full text article rather than only the title, abstract, and keywords. Because of this difference in search capability, our query returns 21,767 results in PMC but only 110 results in PubMed (as of 16 May 2022).

We will restrict our survey to articles published after 2017 because research practices regarding power calculations and open access publishing have changed over time and we would like to take a snapshot of recent practices. We do not restrict the search to an even shorter timeline because we would like to get an idea regarding the magnitude of G\*Power usage, and not only for a single year. A different date range could also be justifiable.

We will download the full list of PMCID IDs returned from our search query. We will then randomly order the list of PMCID IDs using the *sample\_n()* function in R with seed number 1313 and sample from that list until we reach our minimum sample size of 95 articles. We will complete the extraction form on 95 articles that perform a power calculation using G\*Power that solves for sample size for a study reported in that article. We will count the number of articles that perform power calculations using G\*Power that solve for variables other than sample size (e.g., for power or effect size), but we will not code these calculations for reproducibility or errors. We will exclude articles that discuss G\*Power without using it to perform a power calculation, as well as tutorials on how to use G\*Power.

We will then continue to sample articles, selecting only articles that contain a power analysis that solves for sample size for an ANOVA, until we reach 60 articles that meet this criteria. Thus, we will sample more than 95 articles in total. However, we will report all results in terms of the 95 articles, except for results concerning ANOVAs, which we will report for the 60 articles that include an ANOVA power calculation.

### ***Data Extraction***

We created an extraction form specifically designed for this project (available in Appendix A). This form includes three overarching sections. In this form, we first check if the power calculation calculates the sample size needed, the statistical power achieved, or the effect size detectable. Second, we check whether the article transparently reported key elements of the power calculation—including alpha, power, sample size, effect size, and the

statistical test—and whether we can reproduce the calculation using G\*Power. Third, we check for errors in the power calculations, including whether the power calculation matched the statistical analysis used in the results section of the article, whether the appropriate ANOVA option was selected, and other visible errors (e.g., inputting a non-standardized effect size, stating that Cohen's  $d = 0.2$  is a conventionally large effect size, claiming a two-tailed test when a one-tailed test was run). We have not preemptively declared every type of issue we will code as an error. We will define additional types of errors based on open ended responses to the extraction form.

If an article has more than one power calculation that solves for sample size, we will record that there are multiple power calculations, but only use the extraction form on the primary power calculation or the first listed power calculation if none is demarcated as primary.

Two investigators (EAZ & RTT) will independently code each article and resolve coding differences through discussion. If necessary, an additional investigator will arbitrate (HP). Investigators have the option to select that they don't have the statistical expertise to code an article. If they select this option, a statistician from our team (HP) will code this article. In supplementary tables, we will report interrater agreement for each non-open-ended coding item with Cohen's kappa and percentage agreement.

### ***Objective 1. Estimate the number of published articles that use G\*Power***

We will extrapolate from our sample of articles indexed in PMC to estimate the total number of articles published since 2017 and indexed in PubMed that use G\*Power for a power calculation. To do so, we will: (1) measure the percentage of articles from our sample that conduct a power calculation (e.g., if we need to sample 100 articles to identify 95 studies that include a power calculation, this value will be 95%); (2) use this percentage to estimate the total number of PMC articles that include a power calculation (i.e.,  $95\% * \sim 21,767$  articles; we will include 95% CIs); and (3) expand this number from PMC articles to all PubMed articles by multiplying it by the number of times more articles in PubMed compared to PMC (which is  $\sim 2.24$ ).

We will perform this analysis in three different ways, to estimate the total number of articles that use G\*Power for: (1) any power calculation, (2) a power calculation that solves for sample size, (3) a power calculation for an ANOVA that solves for sample size. We will round all estimates to the nearest thousand to avoid misrepresenting greater precision than we truly have. Exact numbers will be provided in the Supplementary Material.

We report estimates separately for the number of articles that use G\*Power for any power calculation from those that use G\*Power to solve for sample size. If an *a priori* power calculation that solves for sample size is conducted incorrectly, it can lead to research waste. For example, an error in this type of power calculation could lead to conducting a study with 30% power, which has little inferential value and can lead to a noisy literature.

To demonstrate how we plan to present the results for Objective 1, we have attached code to this protocol and simulated some of the data ([osf.io/dsv4m](https://osf.io/dsv4m)). The output from this **\*simulated\*** data is shown here:

*We sampled 120 articles, of which 110 included a power calculation for a study in that article and 95 of these articles performed power calculations to solve for sample size. n articles included a power calculation that solved for power and n for effect size (see Supplementary Table 1 for all counts). n (%) articles we surveyed included human participants and n (%) included non-human animals. n (%) were protocols. Sampled articles were published in 2017 (n = n), 2018 (n = n), 2019 (n = n), 2020 (n = n), 2021 (n = n), and 2022 (n = n). The median Journal Impact Factor of included articles was n (IQR n-n).*

*We estimate that between 35000 and 42000 articles indexed by PubMed and published since 2017 use G\*Power for a sample size calculation and that between 9000 and 16000 do so for a sample size calculation for an ANOVA (see Table 1 for additional details).*

*To calculate the total number of articles using GPower in PubMed versus PMC, we simply multiplied the estimates by 2.24, which is the number of articles indexed in PubMed from 2017 onwards divided by the number of articles indexed in PMC from 2017 onwards. This calculation assumes that all PMC articles are indexed in PubMed. If we want to take a conservative estimate and assume that 50% fewer articles that are indexed in PubMed, but not indexed in PMC, use GPower, then we would need to multiply the PMC estimates by 1.62 (i.e.,  $1 + (2.24 - 1) * 0.50$ ) or the PubMed estimates by 0.72 (i.e.,  $1.62 / 2.24$ ). For the rest of this article we will assume that the frequency of use of GPower in PMC and PubMed is the same.*

Table 1. Estimates of the number of published articles that use GPower

	Any power calculation	Sample size calculation	ANOVA sample size calculation
PubMed Central	20000 (19000 - 21000)	17000 (16000 - 19000)	5000 (4000 - 7000)
PubMed	45000 (42000 - 47000)	39000 (35000 - 42000)	12000 (9000 - 16000)

The table is divided into articles that use G\*Power for: any power calculation related to any statistical test (Any power calculation), a power calculation for any statistical test that solves for sample size (Sample size calculation), and a power calculation for an ANOVA that solves for sample size (ANOVA sample size calculation). The total number of articles in each database since 2017 is: PubMed Central 3,246,604; PubMed 7,264,911.

**Supplementary Table 1. Type of power calculations**

Solves for	Count (percent with 95% CI)
Sample size	
Power	
Effect size	
Other	
Unsure	

## ***Objective 2. Assessing the reproducibility of power calculations in G\*Power***

We will present our results regarding reproducibility of power calculations in the format of Table 2 (see below).

**Table 2. Complete reporting and reproducibility of G\*Power calculations.**

Item	% Reported (95% CI)	Total number of articles (indexed in PubMed and published since 2017) (95% CIs)
Alpha 0.05 Other significance level		



<b>Power</b> 80% 95% Other power level		
<b>Effect size</b> $d$ $f$ Other		
<b>Statistical test</b> t-test ANOVA Other		
<b>Sample size</b>		
<b>Version</b>		
<b>Reproducible*</b> Reproducible with assumptions** Reproducible without assumptions		
<b>Justification for effect size provided</b> Previous published research Pilot data Conventions (e.g., a “medium” effect size of $d = 0.5$ ) Smallest effect size of interest*** No justification provided		
<b>Adjusted for multiple comparisons</b>		

\*Our first precision calculation ( $n=95$ ) is for this result.

\*\*For example, in cases where the number of tails on a t-test or the type of ANOVA is not specified. The coding form contains an open-ended response box where coders will outline the assumptions they needed to make.

\*\*\*This includes Minimal Clinically Important Difference (MCIDs)

Below, we include a template text that reports all the items in Table 2. A screenshot of the calculation in the G\*Power GUI would report all these items except for the justification and multiple comparisons.

*“We performed a power calculation using G\*Power **3.1.9.7** for a two-tailed **independent sample t-test** with **alpha = .0167** (Bonferonni corrected for 3 comparisons), and **power = .80**. Powering for a **medium effect size** of Cohen’s  **$d = 0.5$** , requires a sample size of **64 participants per group**”.*

In the manuscript, we may change some of the sub-categories in Table 2. For example, if we find other effect sizes are common (say,  $r$ ), we may include numbers for those effect sizes in the table. We will only count effect size as reported if both the *type* of effect size and the *value* are reported (e.g., Cohen's  $d = 0.5$ ). For effect sizes that are used in 10 or more publications, we will include the median and range. For sample size, we will also include the median sample size and interquartile range. To calculate the percentage of power calculations that adjusted for multiple comparisons, we will only consider studies where adjusting for multiple comparisons could be reasonably performed (i.e., more than one analysis was performed).

**Objective 3. Estimate the frequency of error-free power calculations that use G\*Power**

We will present our results regarding errors in the format of Table 3 (see below). Our extraction form includes three distinct questions regarding errors: (1) whether the default ANOVA option was used incorrectly, (2) whether the power calculation matches a statistical analysis used in the article, and (3) for any other errors, followed by an open text response. After collecting data, we will group the responses to this last question into distinct bins of types of errors.

**Table 3. Errors in G\*Power calculations**

	<b>Numerator/denominator (Percentage with 95% CI)</b>	<b>Total number of articles</b> (indexed in PubMed and published since 2017) (95% CIs)
<b>Reproducible and error-free</b> Appropriate ANOVA option selected* Power calculation clearly matches analysis conducted		
<b>Contains any type of error</b> Inappropriate ANOVA option selected Power calculation does not match an analysis performed Error type #3 Error type #4 Other type of error		

<b>Too unclear to code errors</b>		

\*Our second precision calculation (n=60) is for this result.

We will also present the results excluding the articles that were too unclear to code. For example, we will present the number of articles that clearly use the appropriate ANOVA option divided by the number of articles where we can clearly code whether or not they used the appropriate ANOVA option. These results will have wider confidence intervals than we input for our precision calculations because they will have a smaller denominator due to excluding articles that were coded as unclear.

Some errors will be quantifiable in terms of how they impact the effect size detectable. Where possible, we will quantify the impact of errors. For example, if a researchers uses G\*Power's default option for a within-between ANOVA when intending to use Cohen's method, G\*Power will recommend a sample size of 34 instead of 128 (outlined in Figure 1). Thus, the researchers may only reach 34/128 (27%) of their desired sample size and only have power to detect an effect size twice as large as desired. Some errors will not be quantifiable. For example, if the statistical test used in the power calculation has no relation to an analysis performed in the manuscript.

We will also include supplementary tables with summary data for questions not fully presented in the main text. For example, the multiple comparisons line of Table 2 will only include responses where correcting for multiple comparisons was possible. The supplementary tables will present the summary counts for every response option.

## Statistical Analysis

Unless otherwise stated we will estimate 95% CIs for all proportions and numbers using Monte Carlo sampling as outline in the protocol code.

We will calculate values for the rightmost column of Table 2 and Table 3 by multiplying our estimates in the middle column of the tables by the the percentage of published articles that use G\*Power for sample size calculations. 95% CIs for multplied proportions will be estimated using Monte Carlo sampling to avoid making normal approximations to the binomial distribution (outlined in protocol code). The resulting percentage and 95% CI will then be used to calculate the estimated number of articles for each item in the tables following the approach described in Objective 1.

## Protocol Code

This protocol comes with code that provides a precise planned implementation of certain aspects of the study, available at [osf.io/dsv4m](https://osf.io/dsv4m).

## **Data Management**

Upon manuscript submission, the data will be stored as open data on the Stanford Data Repository ([www.sdr.stanford.edu](http://www.sdr.stanford.edu)). Open data are made available, free of charge, to anyone interested in the project, or who wishes to conduct their own analyses of the data.

## **Publication Policy**

The findings from this research study may be published in an appropriate scientific journal (and made available open access), and/or presented at an appropriate meeting. The results of this study may be published as part of a larger manuscript on sample size calculations or published in smaller units on a platform such as *science-octopus.org*. The data and code will be made available for sharing via the Stanford Data Repository.

## **Funding Source**

Robert T. Thibault is supported by a general support grant awarded to METRICS from the Laura and John Arnold Foundation and postdoctoral fellowships from the Fonds de recherche du Québec – Santé and the Canadian Institutes of Health Research. The Meta-Research Innovation Center at Stanford (METRICS) is supported by a grant from the Laura and John Arnold Foundation. Marcus Munafò and Robert Thibault are part of the MRC Integrative Epidemiology Unit (MC\_UU\_00011/7). Emmanuel Zavalis is supported by a grant from the Carl Erik Levin Foundation. Mario Malički is funded by the Stanford Program on Research Rigor & Reproducibility (SPORR). Hugo Pedder is funded by the National Institute for Health and Social Care Excellence, UK, and the Medical research council, UK (Grant number: MR/M005232/1). The funders have no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

## **Conflicts of Interest**

The researchers declare no conflicts of interest

## **Protocol registration date**

This protocol was registered on 31 May 2022. Before uploading this protocol, RTT and EAZ iteratively piloted and refined the extraction form on ~5 sets of ~5 articles (~25 total).

## **References**

- Aczel, B., Szaszi, B., Sarafoglou, A., Kekecs, Z., Kucharský, Š., Benjamin, D., Chambers, C. D., Fisher, A., Gelman, A., Gernsbacher, M. A., Ioannidis, J. P., Johnson, E., Jonas, K., Kousta, S., Lilienfeld, S. O., Lindsay, D. S., Morey, C. C., Munafò, M., Newell, B. R., ... Wagenmakers, E.-J. (2020). A consensus-based transparency checklist. *Nature Human Behaviour*, 4(1), 4–6. <https://doi.org/10.1038/s41562-019-0772-6>
- Border, R., Johnson, E. C., Evans, L. M., Smolen, A., Berley, N., Sullivan, P. F., & Keller, M. C. (2019). No Support for Historical Candidate Gene or Candidate Gene-by-Interaction Hypotheses for Major Depression Across Multiple Large Samples. *The American Journal of Psychiatry*, 176(5), 376–387. <https://doi.org/10.1176/appi.ajp.2018.18070881>
- Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14(5), 365–376. <https://doi.org/10.1038/nrn3475>
- Carter, A., Tilling, K., & Munafò, M. R. (2017). A systematic review of sample size and power in leading neuroscience journals. *BioRxiv*, 217596. <https://doi.org/10.1101/217596>
- Charles, P., Giraudeau, B., Dechartres, A., Baron, G., & Ravaud, P. (2009). Reporting of sample size calculation in randomised controlled trials: Review. *BMJ*, 338, b1732. <https://doi.org/10.1136/bmj.b1732>
- Clark, T., Berger, U., & Mansmann, U. (2013). Sample size determinations in original research protocols for randomised clinical trials submitted to UK research ethics committees: Review. *BMJ*, 346, f1135. <https://doi.org/10.1136/bmj.f1135>

- Kieslich, P. J. (2020). *Cohen's f in repeated measures ANOVAs*. <https://osf.io/gevp6/>
- Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, 4. <https://www.frontiersin.org/article/10.3389/fpsyg.2013.00863>
- Lakens, D. (2022). Sample Size Justification. *Collabra: Psychology*, 8(1), 33267. <https://doi.org/10.1525/collabra.33267>
- Marek, S., Tervo-Clemmens, B., Calabro, F. J., Montez, D. F., Kay, B. P., Hatoum, A. S., Donohue, M. R., Foran, W., Miller, R. L., Hendrickson, T. J., Malone, S. M., Kandala, S., Feczko, E., Miranda-Dominguez, O., Graham, A. M., Earl, E. A., Perrone, A. J., Cordova, M., Doyle, O., ... Dosenbach, N. U. F. (2022). Reproducible brain-wide association studies require thousands of individuals. *Nature*, 603(7902), 654–660. <https://doi.org/10.1038/s41586-022-04492-9>
- Moher, D., Hopewell, S., Schulz, K. F., Montori, V., Gøtzsche, P. C., Devereaux, P. J., Elbourne, D., Egger, M., & Altman, D. G. (2012). CONSORT 2010 explanation and elaboration: Updated guidelines for reporting parallel group randomised trials. *International Journal of Surgery*, 10(1), 28–55. <https://doi.org/10.1016/j.ijssu.2011.10.001>
- Nature Publishing Group. (2019). *Reporting Checklist*. [https://www.nature.com/documents/Reporting\\_checklist\\_new.doc](https://www.nature.com/documents/Reporting_checklist_new.doc)
- OSF. (2016). *Templates of OSF Registration Forms*. <https://osf.io/zab38/>
- Percie Du Sert, N., Bamsey, I., Bate, S. T., Berdoy, M., Clark, R. A., Cuthill, I. C., Fry, D., Karp, N. A., Macleod, M., Moon, L., Stanford, S. C., & Lings, B. (2017). The Experimental Design Assistant. *Nature Methods*, 14(11), 1024–1025. <https://doi.org/10.1038/nmeth.4462>

- Percie du Sert, N., Hurst, V., Ahluwalia, A., Alam, S., Avey, M. T., Baker, M., Browne, W. J., Clark, A., Cuthill, I. C., Dirnagl, U., Emerson, M., Garner, P., Holgate, S. T., Howells, D. W., Karp, N. A., Lazic, S. E., Lidster, K., MacCallum, C. J., Macleod, M., ... Würbel, H. (2020). The ARRIVE guidelines 2.0: Updated guidelines for reporting animal research\*. *Journal of Cerebral Blood Flow & Metabolism*, 40(9), 1769–1777. <https://doi.org/10.1177/0271678X20943823>
- Ros, T., Enriquez-Geppert, S., Zotev, V., Young, K. D., Wood, G., Whitfield-Gabrieli, S., Wan, F., Vuilleumier, P., Vialatte, F., Van De Ville, D., Todder, D., Surmeli, T., Sulzer, J. S., Strehl, U., Sterman, M. B., Steiner, N. J., Sorger, B., Soekadar, S. R., Sitaram, R., ... Thibault, R. T. (2020). Consensus on the reporting and experimental design of clinical and cognitive-behavioural neurofeedback studies (CRED-nf checklist). *Brain*, 143(6), 1674–1685. <https://doi.org/10.1093/brain/awaa009>
- Rothman, K. J., & Greenland, S. (2018). Planning Study Size Based on Precision Rather Than Power. *Epidemiology*, 29(5), 599–603. <https://doi.org/10.1097/EDE.0000000000000876>
- Rutterford, C., Taljaard, M., Dixon, S., Copas, A., & Eldridge, S. (2015). Reporting and methodological quality of sample size calculations in cluster randomized trials could be improved: A review. *Journal of Clinical Epidemiology*, 68(6), 716–723. <https://doi.org/10.1016/j.jclinepi.2014.10.006>
- Smaldino, P. E., & McElreath, R. (2016). *The natural selection of bad science*. <https://doi.org/10.1098/rsos.160384>
- The NPQIP Collaborative Group. (2019). Did a change in Nature journals' editorial policy for life sciences research improve reporting? *BMJ Open Science*, 3(1),

e000035. <https://doi.org/10.1136/bmjos-2017-000035>

Thibault, R. T., & Pedder, H. (2022). Excess significance and power miscalculations in neurofeedback research. *NeuroImage. Clinical*, 103008.

<https://doi.org/10.1016/j.nicl.2022.103008>

von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gøtzsche, P. C., & Vandenbroucke, J. P. (2008). The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: Guidelines for reporting observational studies. *Journal of Clinical Epidemiology*, 61(4), 344–349.

<https://doi.org/10.1016/j.jclinepi.2007.11.008>



## Appendix A. The survey

[The Qualtrics version is available here.](#)

Start of Block: Triage

Coder initials

☐ RTT

☐ EZ

☐ HP

☐ Other \_\_\_\_\_

Enter the article's **PMCID**.

\_\_\_\_\_

Is this a study protocol?

☐ Yes

☐ No

Does this article **meet our inclusion criteria**? Included articles must use GPower to perform a power calculation. The power calculation can solve for sample size, power, effect size, or any other relevant parameter and may be conducted before or after the study. If the article discusses GPower, but does not use it to conduct a power calculation for a study **presented within that article**, the article should be excluded (e.g., articles that run power calculations only for future studies should be excluded).

☐ Include

☐ Exclude

End of Block: Triage

Start of Block: Power Calc Type

What **type of power calculation(s)** does this article have? Select multiple options if there are multiple power calculations that solve for different variables.

- ☐ Solves for **sample size** (often called *a priori*)
- ☐ Solves for **power** (often called *post hoc*)
- ☐ Solves for **effect size** (often called *sensitivity*)
- ☐ Other \_\_\_\_\_
- ☐ Unsure

Does the article contain **more than one power calculation** that solves for sample size?

- ☐ Yes
- ☐ No

**Complete this form for only one power calculation that solves for sample size.** If there is one power calculation that is unambiguously identified, or likely to be, the primary power calculation, select this one. If there's no indication that one power calculation is more prominent than another one, select the power calculation that solves for sample size that first appears in the article. If there is only one power calculation that solves for sample size, select that one.

Copy-paste text from the article that describes this power calculation. This text should be **VERBATIM**. Do not put text in this box that doesn't come directly from the article. You do not need to include " " quotation marks.

---

---

---

---

---

Do you feel **your statistical knowledge is adequate** to fill out this form for this specific power calculation?

- ☐ Yes
- ☐ Probably
- ☐ No (If you select this option, the survey will terminate and we will be sure to get a stats savvy coder to be the other coder for this power calculation)

End of Block: Power Calc Type

Start of Block: Power Calc Detail

Does the power calculation include the **version of GPower** used? If yes, enter the version number.

- ☐ Yes \_\_\_\_\_
- ☐ No

Does the power calculation report the **power or beta** (i.e., 1 - power) used? If yes, enter the power value (e.g., 0.8).

- ☐ Yes \_\_\_\_\_

☐ No

Does the power calculation report **alpha or the p-value** used? If yes, enter the value (e.g., 0.05).

☐ Yes \_\_\_\_\_

☐ No

Does the power calculation report the **sample size**? If yes, enter the **TOTAL sample size** (e.g., if the power calculation has 2 groups with 30 participants each, then enter 60). Enter the sample size used in the power calculation (rather than the actual sample size used). Some articles may add participants to their sample size because of attrition. In this case, enter the sample size output from the power calculation (i.e., what's expected after attrition).

☐ Yes \_\_\_\_\_

☐ No

Does the power calculation report the **type of effect size**?

☐ d

☐ f

☐ r

☐ w

- ☐  $f^2$
- ☐ ratios (e.g. odds ratio or risk ratio)
- ☐ Other: Standardized \_\_\_\_\_
- ☐ Other: **Non**-standardized (e.g., 5 points on a questionnaire scale).
- \_\_\_\_\_
- ☐ No

Does the power calculation report the **value for the effect size**? If yes, then enter the value (e.g. 0.5). Only enter a number here. We will know what type of effect size it is based on your response to the previous question.

- ☐ Yes \_\_\_\_\_
- ☐ No

Does the power calculation report the **statistical test**?

- ☐ t-test
- ☐ ANOVA
- ☐ correlation (e.g., Pearson's or Spearman's)
- ☐ test of proportions (e.g., Fisher's, or chi-squared)
- ☐ other regression (please describe)
- \_\_\_\_\_

☐ other non-regression (please describe)

---

☐ the statistical test is NOT reported

Are there **any other elements of the power calculation** that should be reported, but are not reported? For example, the expected proportions in Fisher's test or the number of predictors in a regression. You don't need to reply to this question.

---

---

---

---

---

Can you **reproduce** the power calculation in GPower (within a few minutes)? If yes, take a screenshot of GPower to show the reproduction and save the screenshot filename as the PMCID.

☐ Yes, based solely on the information in the article or its supplementary material

☐ Yes, but I've had to make some assumptions. (please list the assumptions you made)

---

☐ No

Does the article mention a **justification or basis for the effect size** chosen in their power calculation (you may select multiple options)?

☐ Previously published research

☐ Their own pilot data (that is not published)

- ☐ Conventions of small, medium, and large effect sizes (e.g., "we powered for a medium effect size of Cohen's  $d = 0.5$ ")
- ☐ An effect size of interest or minimal clinically importance difference (MCID)
- ☐ No justification or basis is reported
- ☐ This is a power calculation run after the study that uses the effect size from this study (i.e., a post-hoc power calculation)
- ☐ Other (please describe) \_\_\_\_\_

Does the power calculation **mention accounting for multiple comparisons?**

- ☐ **Yes**
- ☐ **No**, and the **article contains multiple analyses** with no clear indication of a sole primary analysis for which this power calculation is for. (Note, if they are powering for an ANOVA and reporting multiple effects--e.g., two main effects and an interaction for a 2x2 ANOVA--then there should be a correction for multiple comparisons).
- ☐ **No**, and **there is no reason to account for multiple comparisons** (e.g., there is only one analysis, or this analysis is clearly demarcated as the primary analysis)
- ☐ Unsure

Justify your response to the previous question regarding **multiple comparisons**. Or enter "NTA" for "nothing to add".

---



---



---



---

Is this power calculation for a within-between **ANOVA** interaction or within ANOVA main effect?

- ☐ **Yes**, and the researchers **selected a non-default option or accounted for the default option.** (e.g., by adjusting the value of  $f$  to account for the within subjects correlation)
- ☐ **Yes**, but the researchers **use the default option without accounting for it** (e.g., powering for a "medium" effect size by entering  $f = 0.25$ ).
- ☐ **Yes**, but I cannot reasonably assume which option they used.
- ☐ **No**
- ☐ **Unsure** whether the power calculation was for an ANOVA or another type of statistical test.

Does the **power calculation match a statistical analysis used in the article.** For example, whereas a paired t-test could be used in the power calculation, the article may use an independent samples t-test. The power calculation could also be completely unrelated to the actual analyses conducted (e.g., powered for an ANOVA, but Fisher's tests conducted).

- ☐ Yes
- ☐ No
- ☐ Unsure (there's not enough information to reasonably code yes or no).
- ☐ The article is a protocol

Justify your response to the previous question regarding whether the **power calculation matches a statistical analysis in the article.** Or enter "NTA" for "nothing to add".

---

---



---

---

---

Is there any other reason (beyond the two previous questions on ANOVAs and matching) why the power calculation contains an error or is inappropriate? For example, the effect size entered could be non-standardized, incorrectly calculated, incorrectly identified in relation to Cohen's indices without justification (e.g., calling  $f = 0.25$  a small effect size), or performed after the study is complete. We do not have an exhaustive list of all possible errors, so please also code "Yes" for this question if you come across an error that we haven't described.

- ☐ Yes
- ☐ Likely, but I cannot be certain based on the information provided.
- ☐ No
- ☐ Unsure (there's not enough information to reasonably code yes or no).

Justify your response to the previous question regarding whether the power calculation is **inappropriate for any other reason**. Or enter "NTA" for "nothing to add". If you responded "Yes" or "Likely..." to the previous question, you must explain why in this box.

---

---

---

---

---

If an error is present, **can the impact of the error be quantified**? For example, if the default ANOVA option was mistakenly chosen, what is the difference in sample size between what the article calculated and the same calculation with a different option selected? If no error is present in the calculation, you may leave this box blank.

---

---

---

---

---

Provide any additional comments you might have about this specific power calculation.  
(optional)

---

---

---

---

---

End of Block: Power Calc Detail

Start of Block: Article meta-data

Who are the "**participants**" in the study the power calculation was written for? In other words, what type of study is this?

- ☐ Humans
- ☐ Non-human animals (this includes both in vivo and in vitro studies)
- ☐ Other \_\_\_\_\_

Enter the name of the **journal** where the article was published.

---

Enter the **year** the article was published.

---

Enter the **Journal Impact Factor** for 2 YEARS PRIOR to the year of publication (i.e., if the article is published in 2021, enter the JIF for 2019. This is the JIF that would have been available the year of publication. At the moment, JIFs have only been calculated for 2020 and earlier years).

Use the Journal Citation Report from <https://jcr.clarivate.com/jcr/home>. If the journal is not listed in that database, or does not have an impact factor for the year the article was published, enter "NA" into this box.

---

Provide any additional comments you may have about **this survey form or this study in general** (optional).

---

---

---

---

---

End of Block: Article meta-data

## **Appendix B: Contributor Roles Taxonomy (CRediT) \*projected\* roles**

Conceptualization:	R.T.T.
Data curation:	E.A.Z., R.T.T.
Formal analysis:	E.A.Z., R.T.T.
Funding acquisition:	R.T.T.
Investigation:	E.A.Z., H.P., R.T.T.
Methodology:	R.T.T., E.A.Z., H.P., Mario.M., Marcus.M., S.G.
Project administration:	R.T.T.
Software:	E.A.Z., R.T.T.
Supervision:	R.T.T.
Validation:	E.A.Z., R.T.T.
Visualization:	E.A.Z., R.T.T.
Writing - original draft:	R.T.T., E.A.Z.
Writing - review & editing:	R.T.T., E.A.Z., H.P., Mario.M, Marcus.M., S.G.