# Chapter 7. Systematic Review and Meta-analysis of Statistical power surveys of Psychological Science

## Abstract:

This study uses mixed effects meta-analyses to analyze 46 power surveys, including over 8,000 individual studies published from 1932-2014, to estimate the average statistical power of psychology research at Cohen’s standardized effect size benchmarks and to show how this value has changed over time. The average statistical power of psychology research over this time period at Cohen’s (1988) benchmarks is extremely low for ‘small’ effects, .23 (95% CIs [.17, .29]), somewhat low for ‘medium’ effects, .62 (95% CI [.54, .70]), and only acceptably high for ‘large’ effects, .80 (95% CI [.68, .92]). The average statistical power of published psychology research has seen little to no change over time. A secondary analysis of surveys which assessed how often power analyses are reported in psychology research suggests that power analysis reporting rates have increased slightly over time but remain low. Finally, methods for avoiding the negative impacts of low powered research are outlined.

**Keywords:** *Meta-analysis, psychology, statistical power, research planning, AIPE*

## **5.1 Introduction**

Statistical power describes the probability of a statistical test finding statistically significant results given that a specific alternative hypothesis holds true. Cohen’s first power survey (1962) showed that articles published in a 1960 issue of the Journal of Abnormal and Social Psychology had a mean power of .48 to detect a ‘medium’ effect size equivalent to 0.5 Cohen’s d. This suggests that the average psychology study should fail to reach statistical significance even when studying a true “medium” effect on more than 50% of occasions. Since the publication of Cohen’s 1962 article over 40 power surveys have been performed, studies which systematically assess the statistical power of bodies of psychology research. The current study brings those papers together to estimate the statistical power of psychology research at Cohen’s effect size benchmarks and show how the statistical power of psychology research has changed over time.

If studies in a body of literature have a low average statistical power, several major negative outcomes follow. Firstly, low power studies often produce estimates of effects that are so imprecise as to not allow researchers to make meaningful inferences (Cohen, 1962, 1988), wasting research funds, participant, and researcher time. Secondly, in the presence of publication and reporting biases toward statistically significant results, low average power leads to effect size exaggeration among published studies, and an increased false positive error rates among significant reported results (Marjan Bakker, van Dijk, & Wicherts, 2012; DeCoster, Sparks, Sparks, Sparks, & Sparks, 2015; Ioannidis, 2008). In recent years the low average power of psychology research has been pointed to as one of the driving factors of the “replication crisis” in psychology (Maxwell, Lau, & Howard, 2015). Finally, low power research can be self-reinforcing. If researchers base the sample sizes they use in their studies on previous low power research, or if they base sample size decisions on published (and, on average, exaggerated) effect sizes, their own research will often have a lower than desirable level of statistical power to detect likely effect sizes [cite pub bias paper].

Hundreds of articles have been published since the 1960s discussing the issue of low statistical power in psychology [history chapter], and numerous tools have been developed to make power analysis an easy and routine part of research planning, from Cohen’s own textbooks and publications (e.g., 1988; 1992) to statistical power analysis computer programs (e.g., Faul, Erdfelder, Lang, & Buchner, 2007). However, it is unclear whether the relative ease of use of these tools, as well as other changes in the way that research is planned and performed, have led to any change in the average power of psychology research over time. Given this large and growing body of work and the importance of avoiding the negative impacts of low statistical power on research literatures, it seems essential to begin to assess whether these efforts have had any impact on the statistical power of psychology research.

In order to address this question, we use a systematic review and meta-analysis of power surveys in psychology to estimate the average power of this literature and to show whether this value has changed over time. Power surveys are articles which have examine a set of published studies and calculate their power to detect Cohen’s “small”, “medium” and “large” effect size benchmarks (see Table 2 for a list of Cohen’s effect size benchmarks for different analyses) given the particular statistical analyses that are used and the sample size included in each statistical analysis.

Given that many of the included power surveys suggest that power analysis should be performed as part of research planning – along with the American Psychological Association and CONSORT reporting guidelines (APA Publications Communications Board Working Group on Journal Article Reporting Standards, 2008; Moher et al., 2010; Moher, Schulz, & Altman, 2001; Wilkinson, 1999) – a related and crucial question is whether researchers are following these instructions and performing and reporting power analyses more often. In order to address this question, Supplementary Materials 5 reports a meta-analysis of examinations of the proportion of articles which report a power analysis in order to assess whether there has been any change in how often power analyses are reported over time.

## 5.2 Method

### 5.2.1 Research Design

The design and hypotheses, along with a detailed analysis plan for the secondary analysis, were preregistered after an initial pilot sample of 17 articles had been collected, but before any analysis or summary statistics had been calculated. The pre-registration and pilot data are available from <https://osf.io/n6jfd/>, see table 1 for a list of deviations from the pre-registered protocol.

Table 1.

*Deviations from preregistered protocol.*

|  |  |
| --- | --- |
| Deviation | Explanation |
| Missing means and SDs were estimated or imputed when missing | Means and variances were imputed as large numbers of studies had some missing data. Analyses were also run without data imputation as was preregistered (see supplementary material 4). |
| Meta-analysis estimated means not medians | Mean levels of power were reported more often than medians, in 45 compared to 47 articles, and as the standard error of means is smaller than that of medians all else being equal. |
| Restricted maximum likelihood estimation was used | Restricted maximum likelihood estimation was used, no estimation method was preregistered. |
| Empirical Bayes estimates and 95% credible intervals for random effects were estimated | No method of examining random effects estimates was preregistered, although empirical Bayes estimates and 95% credible intervals for random effects of area of psychology were calculated following Morris, 1983 and Robinson, 1991 |
| Random effects were included for area of research and original study in both primary and secondary analyses | No method of accounting for non-independence between articles was preregistered. The mixed effects meta-analyses reported here include random effects for study, area of research as well as each study’s estimate. The preregistered models were also performed and are reported as sensitivity analyses, see supplementary materials 4 for model output for the primary analysis and supplementary materials 5 for model output for the secondary analysis. |
| No analysis was performed examining sample size as an outcome | No analysis was performed with sample size as an outcome as few articles (7) reported the average sample sizes of the investigated areas of research. |
| New fields of psychology research were included: “Sport and exercise psychology” and “communication research” | “Sport and exercise psychology” and “communication research” are distinct areas of research not listed as subfields in the preregistration |

### 5.2.3 Record identification

See figure [PRISMA] for a PRISMA flow diagram of article identification, screening, eligibility analysis and inclusion. The sampling strategy was designed to return all reviews of the statistical power of bodies of research in psychological research (broadly defined, including educational, occupational, management, clinical, psychiatry, and neuroscience research). Power surveys were included if they systematically calculated the statistical power of statistical tests in a body of published research articles using effect sizes equivalent to Cohen’s (1988) benchmarks estimates for “small”, “medium” and “large” effects (see table 2). Articles which analysed the power of fewer than six articles were excluded to exclude articles which were not literature surveys but rather criticisms of a small body of “underpowered” research. Only articles with full texts available in English were included.

Table 2. Effect size benchmarks following Cohen (1977, 1988, 1992)

|  |  |  |  |
| --- | --- | --- | --- |
|  | Effect size benchmark | | |
| Type of Test (effect size unit) | Small | Medium | Large |
| *t* test on means (*d*) | .2 | .5 | .8 |
| *t* test on correlations (r) | .1 | .3 | .5 |
| F test ANOVA (f) | .1 | .25 | .4 |
| F test for multiple correlation or regression (f2) | .02 | .15 | .35 |
| Chi-square test (w) | .1 | .3 | .5 |

Note. Cohen (1962) used slightly different estimates for small and large benchmarks (e.g., for *t* tests for mean differences *d* = .25 and 1 respectively) although the medium benchmarks has remained the same.

On the 11th of September 2017 the PsycInfo and Medline databases for all records including the words “power\*” “sampl\*” in their title and "power analysis", "statistical Power" or "sample size" in the main text, identifying an initial 1988 articles. After de-duplication, 1526 articles remained in the database. This database is available from <https://osf.io/t6jf8/>. Hand searches of all identified applicable articles’ reference lists were performed to attempt to identify any papers detailing power surveys that may have been missed by these search criteria, identifying an additional 18 articles. One additional article (Szucs & Ioannidis, 2017) was identified through a Google Scholar search of “power survey psychology”. See supplementary materials 1 for a list of articles included, and supplementary materials 2 for search criteria. All of the articles identified in this literature search which reported the proportion of examined articles which reported a power analysis were included in the secondary analysis, along with two additional articles identified through reference list searches of the articles included in the secondary analysis.

#### 5.2.3.2 Abstract and full text screening

1432 articles were excluded during abstract screening as they did not report examinations of the power of a body of psychology research (e.g., they discussed social power dynamics not statistical power, provided power analysis advice but did not examine a body of literature, etc). After screening of abstracts, 92 records remained and were subjected to full text screening. During full text screening, 46 articles were excluded leaving a total of 46 articles which gave mean or median power estimates for at least one of the small, medium or large effect size benchmarks. See Figure 1 for exclusion reasons at the full-text eligibility assessment phase.

## Screening

Full-text articles excluded  
Reasons: 3 records were of areas of research outside of the current scope, 4 were not available in English, 4 full texts were not available, three included duplicate data (i.e., data included in other papers sampled), 15 did not calculate power for a benchmark level of power or report average sample sizes, 1 did not specify how many articles were included in their analysis, 2 included < 6 articles

(total *n* excluded = 46)

Records excluded from this study as they did not estimate the statistical power of studies  
(n = 1419)

Full-text articles assessed for eligibility in the primary quantitative synthesis  
(*n* = 92)

Studies included in quantitative synthesis of average power of bodies of research  
(*n* = 46)

*Figure 1.* Prisma flow diagram of article identification, screening and selection. Note that this diagram does note show the selection process for the secondary analysis. The secondary analysis included 17 studies, 15 of which were identified during data eligibility screening for the primary analysis, and two which were identified through reference list searches conducted during data extraction for the secondary analysis.

## Identification

## Eligibility

## Included

Records identified through hand search of included articles reference lists (*n* = 18)

Records identified through database searching of PsycINFO (Ovid interface, *n* = 916), Web of Science  
Core Collection (*n* = 1072), and Google Scholar search (*n* = 1, total *N* = 1989)

Records after duplicates removed  
(*n* = 1511)

Record abstracts screened (*n* = 1511)

### 5.2.4 Data extraction

The articles included in the primary analysis were examined in randomized order to avoid systematic order effects. When additional power surveys were identified during data extraction by reference list searches, they were put aside until the current round of data extraction was complete, at which time all newly identified articles were assessed in random order. See <https://osf.io/7ncke/> for data, and Supplementary Materials 3 for the codebook as well as a full list of datapoints extracted from articles for both the primary and secondary analyses. See Supplementary Materials 1 for a list of all included studies along with the sample size included in each study.

### 5.2.5 Missing data handling and imputation

There were a total of 53 year ranges (henceforth “cases”) for which mean or median power estimates were given for at least one of the small, medium or large effect size benchmarks in the 46 included articles. In eleven of these cases, no means were given for at least one of Cohen’s 1988 benchmark effect sizes (including Cohen, 1962, which used different “small” and “medium” benchmark effect sizes), although medians and interquartile ranges were provided. In total of 24 cases no variances or standard deviations were given for at least one of the power estimates. Several methods were used to estimate these missing values. For nine articles (including one which only provided a frequency table at the small benchmark (Cashen & Geiger, 2004)) variances were estimated using the cumulative frequency tables reported in the original articles as

being the frequency of occurrences within each bin, being the mid interval value (e.g., for the bin .1 - .19, the mid interval value would be .145), being the total number of values included and being the estimated mean value calculated as

An r script with the data extracted from the frequency tables and the working for these estimates can be found at <https://osf.io/7ncke/>. This method was also used to estimate the means of two articles (Haase, 1974; Woolley, 1983) which did not provide means or variances (also included in the count above), but which did provide frequency tables. In order to validate this mean estimation methods, the difference between the estimated means and the reported means was calculated for all papers for which variances were estimated using frequency tables. The mean absolute difference between the 22 estimated means and their reported values was just .022.

For six cases at the small and the medium effect size benchmarks and five at the large benchmark, missing means were estimated from reported medians and standard deviations using Wan, Wang, Liu, and Tong (2014)’s method (equation C3) using reported the medians and interquartile values (using the R package varameta; Grey, 2019). In order to validate this approach, the means for all articles which reported medians, quartiles as well as means were calculated (18 articles reporting 52 estimated means), which lead to a mean absolute error of .04.

The range and interquartile ranges of power at the small and medium benchmarks was extracted from plots in one article (Smith, Hardy, & Gammell, 2011) for use in Wan and colleagues (2014) estimator using R’s ‘locator’ function (Poisot, 2010), see <https://osf.io/7f2q9/> for the code used. In order to validate the accuracy of this extraction method, median power levels for the medium and small effect size benchmarks for each year were also extracted and compared to the estimates provided in the paper’s text; all six extracted values were within 0.005 of the values reported the text.

Two power surveys had medians and quartiles which were identical at the large effect size benchmark (.99) which would lead to an estimated variance of zero using the Wan and colleagues’ (2014) method. For these two articles, and three remaining articles which did not report variances, standard deviations, interquartile ranges, or frequency tables, we imputed the variances using the mean variance of all other studies in the reported analyses. To assess the sensitivity of the presented results to this imputation rule all models were run using alternative variance imputation rules including multiple imputation, using the median, minimum, and using the maximum of the other studies’ variances instead of the mean. None of the alternative imputation decisions substantially altered any of the reported results (see “Sensitivity Analyses” below for more detail). At the end of data estimation and imputation there were 51 cases from 44 articles with means and variances at the small benchmark, 53 cases from 46 articles at the medium benchmark, and 49 cases from 42 articles at the large benchmark.

### 5.2.6 Primary analysis

All data-analysis was conducted using R 3.5.0 (R Development Core Team, 2018) , and meta-analyses were performed using the metafor package (version 2.0.0; Viechtbauer, 2010). Data and code for the performed analyses are available at <https://osf.io/as7md/>.

At each benchmark level of power (small, medium, and large) a mixed effects meta-regression was performed. Random effects were included for each estimate , survey (), and area of psychology research , to account for non-independence of sub-studies within articles (e.g., when a power survey reported multiple power estimates for different year ranges), and when studies covered the same fields of research (e.g., clinical psychology). The year studied in each power survey was included as a fixed effect.

The variable *year* was mean centered, making the overall intercept interpretable as the estimated mean power at the mean examined year included in this study (1985). When a study covered a range of years, the mean year of the range of studies included in each set was entered as a predictor in the meta-regression. Sampling variances were calculated as the sample variance divided by the number of articles included in each power survey. All analyses used restricted maximum likelihood estimation. Mean statistical power was treated as a simple mean despite the fact that it is bounded between 0 and 1, a reasonable approximation as each power estimate included in this study is a mean of a set of individual power estimates, not a proportion or fraction. An additional, non-preregistered exploratory analysis was performed to obtain empirical Bayes estimates and 95% credible intervals for the random effect of area of psychology (using the methods outlined in Morris, 1983).

### 5.2.7 Sensitivity analyses

To investigate whether the results are sensitive to data imputation and estimation methods, analyses were also performed excluding including any studies for which any data had to be estimated or imputed, using different data imputation rules (i.e., the median, minimum and maximum variance imputation instead of mean imputation, and without random effects for estimate, study or field of research). A further sensitivity analysis was performed using multiple imputation to estimate variances using predictive mean matching (imputing variances for all articles where variances or standard deviations were not directly reported in the paper). None of these changes altered the effect of year by more than .002, altered the intercept parameter by more than .04, changed any parameters statistical significance at the .05 level, or provided results which would lead to substantially different conclusions being drawn. See Supplementary Materials 3 for coefficient values produced under these different imputation rules.

Leave one out cross validation was used to assess whether any individual article has a large impact on the model coefficient values. No included articles changed the estimated effect of time by more than .004. Intercept estimates did not change by more than .018 in the small or medium benchmark, but the removal of Woods et al., (2006) at the large effect size benchmark increased the intercept parameter by .05. As preregistered, this article has been left in for the results reported below.

Because power is bounded between 0.05 and 1 for all included studies, studies which found estimated mean powers that are close to either bound are expected to have lower variances due to range restriction. This means that the typical inverse variance weighting approach to meta-analysis (Hedges, 1992) may systematically overweight these studies. In order to account for this issue, meta-analyses were also run weighting by the number of articles included in papers following Hunter and Schmidt (2004). This analysis showed little difference in parameter estimates from the inverse-variance weighting approach, with the estimated effect of year changing by less than .002, and the intercept estimate changing by -.001, -.002, and -.04 at the small, medium and large effect size benchmarks respectively. Although this approach avoids the overweighting studies which had mean powers near either bound, it is a less efficient estimator and as such the main results presented below use inverse variance weighting (Marín-Martínez & Sánchez-Meca, 2009).

## 5.3 Results

### 5.3.1.2 Primary analysis results

The mixed effects meta-regression intercept parameter suggests the mean power of psychology across this time period is .23 (95% CIs [.17, .29]) for ‘small’ effects, .62 (95% CIs [.54, .70]) for ‘medium’ effects and .80 (95% CIs [.68, .92]) for ‘large’ effects following Cohen’s effect size benchmarks. The estimated effect of time is negligible at all three benchmarks, 0.000 (95% CIs [-0.003, 0.003]), .001 (95% CIs [-0.002, 0.004]) and -.001 (95% CIs [-0.002, 0.001]) at the small, medium and large benchmarks respectively, with these values representing the estimated change in power per year. There is substantial heterogeneity across fields of research, with estimated standard deviations in subfield means of .05, .08, and .15 at the small medium and large benchmarks respectively. See Tables [BLUPS] – [BLUPL] for empirical Bayes estimates of the random effects for each included subfield. In interpreting the empirical Bayes estimates of random effects for area of psychology at each benchmark, it’s important to emphasize the degree of imprecision in these estimates. This data doesn’t provide strong evidence to claim that any particular subfield has substantially higher or lower power than any other. See Tables [Meta-regression primary small]-[Meta-regression primary large] for full model output, variance estimates and QE tests for excess heterogeneity.

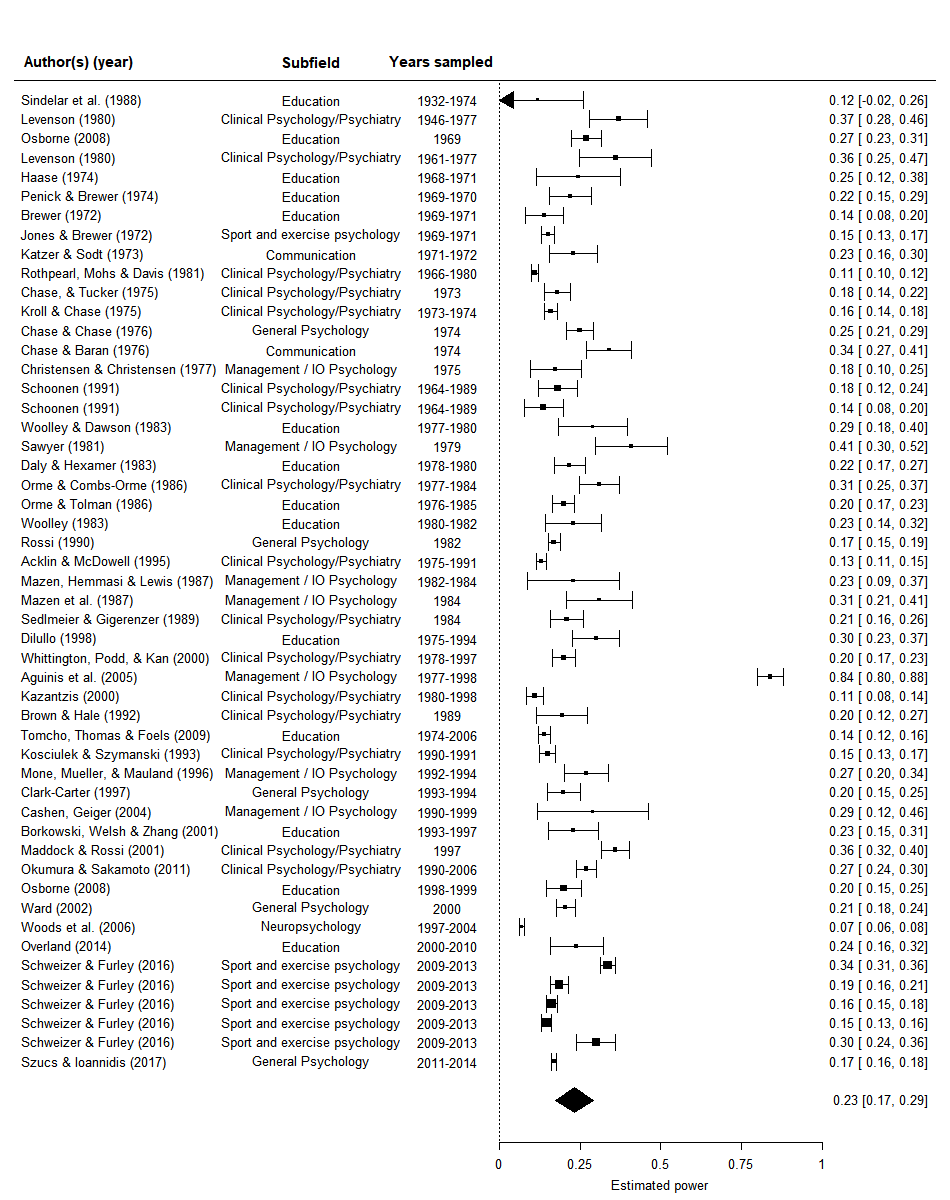


Figure [ForestMedSmall]. Forest plot of studies of the power of the power of psychology research literatures at Cohen’s (1988) small effect size. The polygon depicts reports the model intercept. NAs included for studies that did not estimate power at this effect size benchmark but did for the medium or large benchmarks.

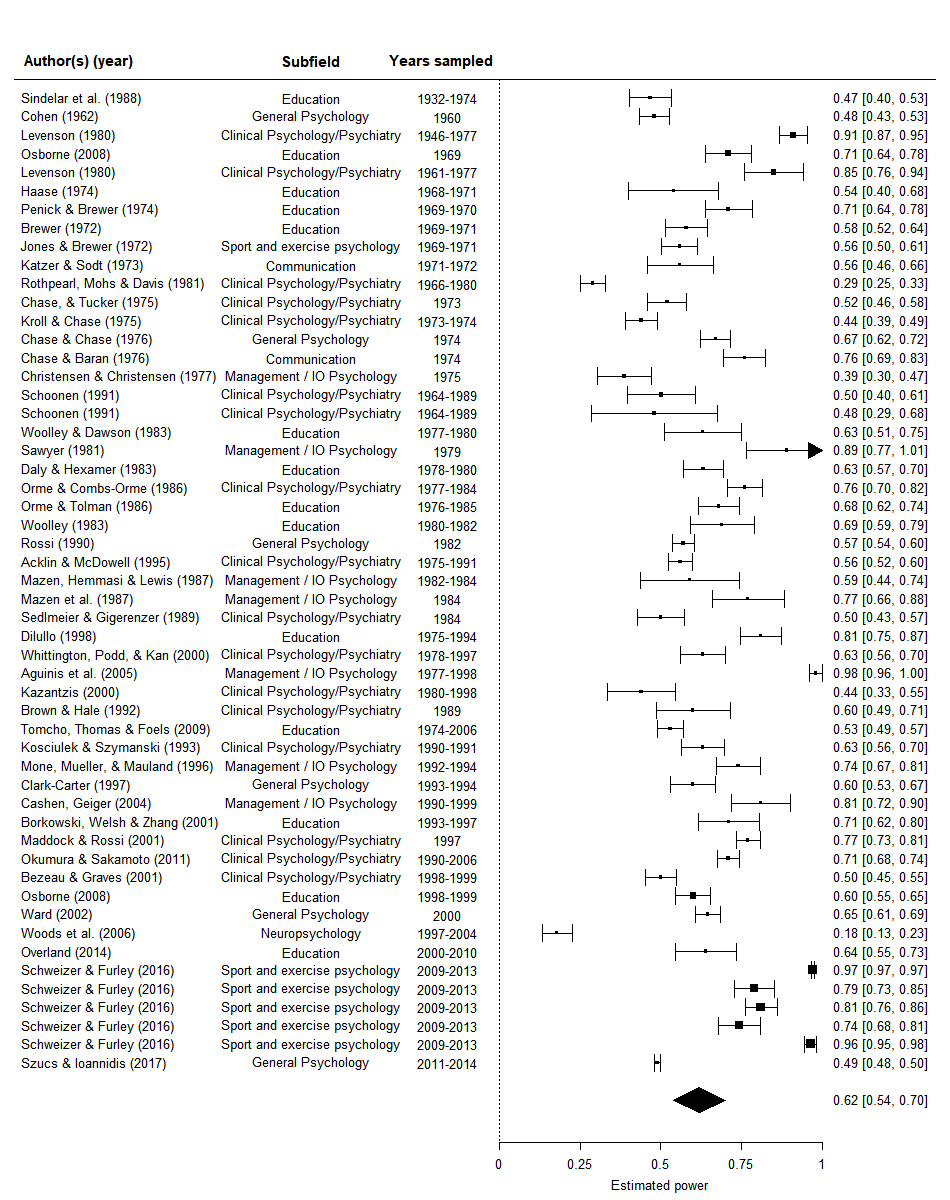


Figure [ForestMedMean]. Forest plot of studies of the power of the power of psychology research literatures at a Cohen’s (1988) medium effect size. The polygon depicts reports the model intercept.

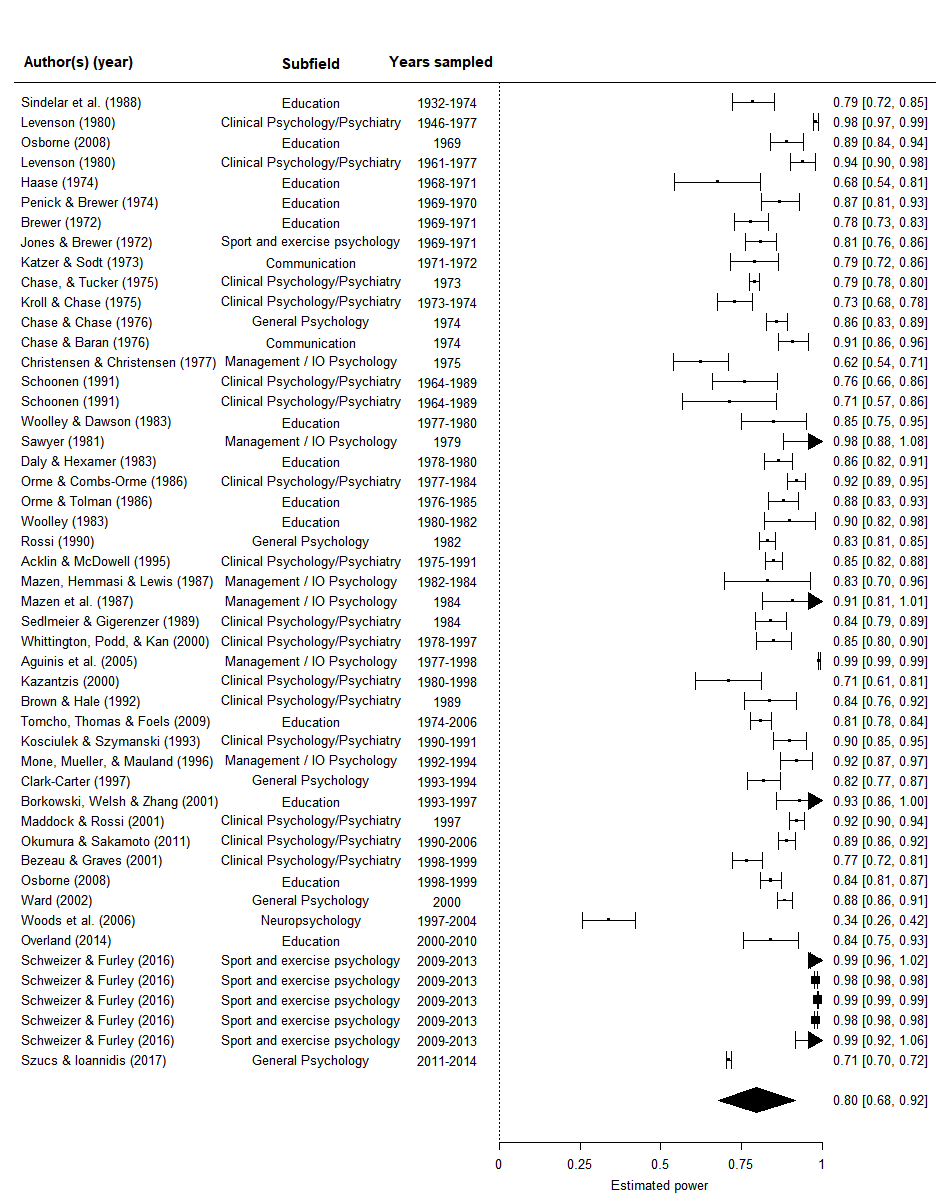


Figure [ForestLargeMean]. Forest plot of studies of the power of the power of psychology research literatures at Cohen’s (1988) large effect size. The polygon depicts the model intercept. NAs included for studies that did not estimate power at this effect size benchmark but did for the small or medium benchmarks.

Table [Meta-regression primary small].

*Meta-regression of the power of psychology studies at a small effect size, including the year studied in each power survey as a moderator.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Coefficient | *b* | *b*  95% CI  [LL, UL] | *p* | Random effects and variance |
| Estimated power | 0.233 | [0.174, 0.292] | < .001 |  |
| Year | -0.000 | [-0.003, 0.003] | .99 |  |
|  |  |  |  | Effect σ2 = .005, n = 51 |
|  |  |  |  | Article σ2 = 0.008, n = 44 |
|  |  |  |  | Subfield σ2 = 0.003, n = 7 |
|  |  |  |  | QE(50) = 2435.76, *p* <.001 |

*Note.* *LL* and *UL* indicate the lower and upper limits of a confidence interval, respectively.

Table [Meta-regression primary medium].

*Meta-regression of the power of psychology studies at a medium effect size, including the year studied in each power survey as a moderator.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Coefficient | *b* | *b*  95% CI  [LL, UL] | *p* | Random effects and variance |
| Estimated power | 0.620 | [0.539, 0.701] | < .001 |  |
| Year | 0.001 | [-0.002, 0.004] | .47 |  |
|  |  |  |  | Effect σ2 =.007, n = 53 |
|  |  |  |  | Article σ2 = .015, n = 46 |
|  |  |  |  | Subfield σ2 = .007, n = 7 |
|  |  |  |  | QE(51) = 10826.44, *p* < .001 |

*Note.* *LL* and *UL* indicate the lower and upper limits of a confidence interval, respectively.

Table [Meta-regression primary large].

*Meta-regression of the power of psychology studies at a large effect size, including the year studied in each power survey as a moderator.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Coefficient | *b* | *b*  95% CI  [LL, UL] | *p* | Random effects and variance |
| Estimated power | 0.797 | [0.677, 0.916] | < .001 |  |
| Year | -0.001 | [-0.002, 0.001] | .24 |  |
|  |  |  |  | Effect σ2 = 0.000, n = 49 |
|  |  |  |  | Article σ2 = 0.007, n = 42 |
|  |  |  |  | Subfield σ2 = 0.023, n = 7 |
|  |  |  |  | QE(47) = 6667.43, *p* < .001 |

*Note.* *LL* and *UL* indicate the lower and upper limits of a confidence interval, respectively.

Table [BLUPS]. Empirical Bayes estimates and 95% credible intervals for random effects of area of psychology included in the current study at the large benchmarks, which are equivalent to 95% confidence intervals assuming that the studies are a random sample from a population with normally distributed average effect size differences.

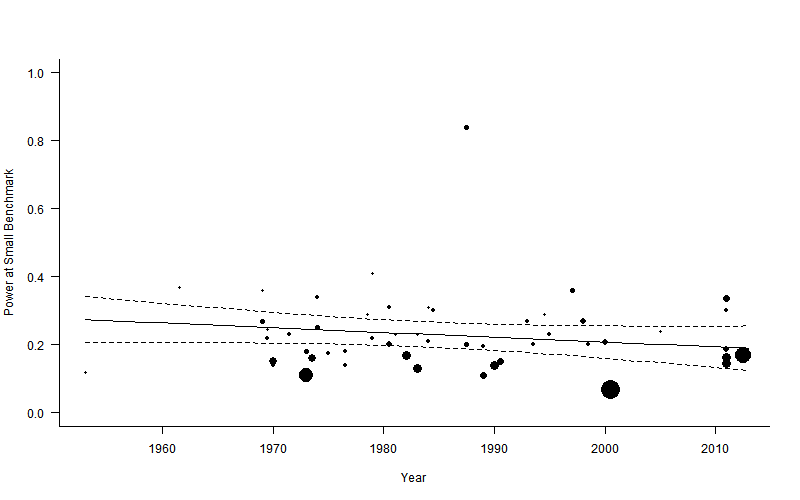
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  | 95% Credible interval | |
| Area of research | Estimate | se | LB | UB |
| Clinical Psychology/Psychiatry | -0.019 | 0.035 | -0.087 | 0.048 |
| Communication | 0.016 | 0.047 | -0.076 | 0.108 |
| Education | -0.011 | 0.036 | -0.081 | 0.058 |
| General Psychology | -0.018 | 0.04 | -0.098 | 0.061 |
| Management / IO Psychology | 0.079 | 0.04 | 0.002 | 0.157 |
| Neuropsychology | -0.032 | 0.05 | -0.129 | 0.065 |
| Sport and exercise psychology | -0.014 | 0.045 | -0.102 | 0.074 |

Table [BLUPM]. Empirical Bayes estimates and 95% credible intervals for random effects of area of psychology included in the current study at the medium effect size benchmark.

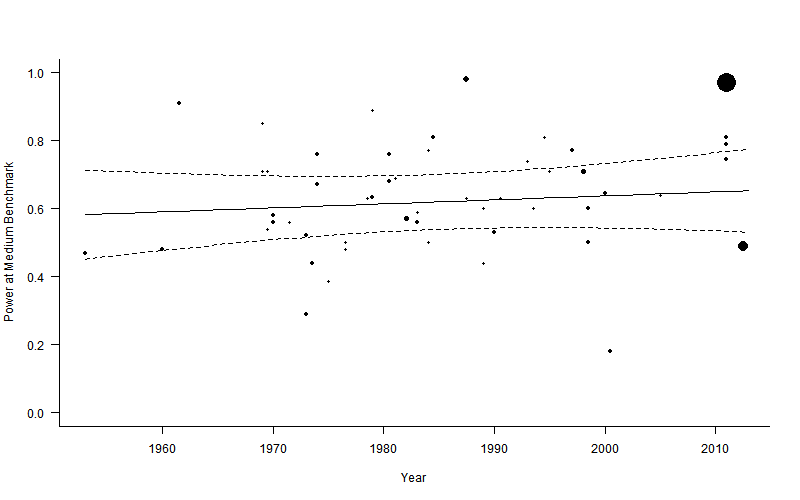
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  | 95% Credible interval | |
| Area of research | Estimate | se | LB | UB |
| Clinical Psychology/Psychiatry | -0.026 | 0.049 | -0.122 | 0.069 |
| Communication | 0.021 | 0.067 | -0.111 | 0.153 |
| Education | 0.018 | 0.05 | -0.08 | 0.116 |
| General Psychology | -0.03 | 0.056 | -0.139 | 0.079 |
| Management / IO Psychology | 0.08 | 0.055 | -0.028 | 0.188 |
| Neuropsychology | -0.105 | 0.073 | -0.248 | 0.037 |
| Sport and exercise psychology | 0.043 | 0.065 | -0.084 | 0.171 |

Table [BLUPL]. Empirical Bayes estimates and 95% credible intervals for random effects of area of psychology included in the current study at the large effect size benchmark.

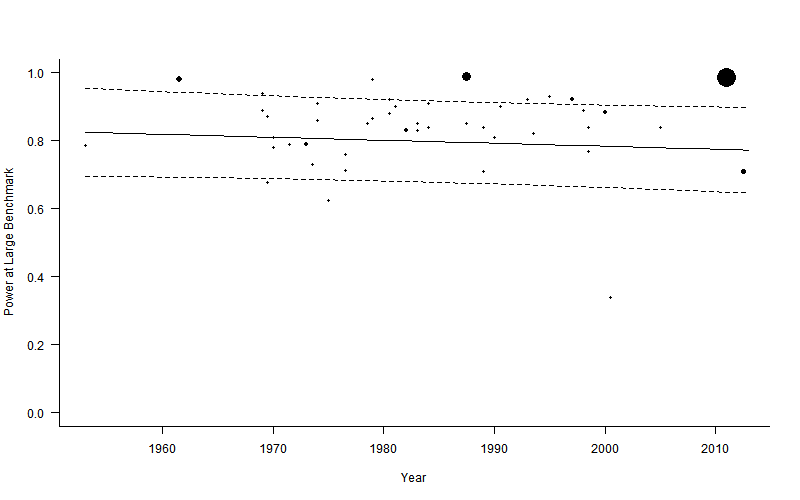
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  | 95% credible interval | |
| Area of research | Estimate | se | LB | UB |
| Clinical Psychology/Psychiatry | 0.044 | 0.064 | -0.081 | 0.169 |
| Communication | 0.039 | 0.078 | -0.114 | 0.193 |
| Education | 0.038 | 0.065 | -0.088 | 0.165 |
| General Psychology | 0.028 | 0.068 | -0.105 | 0.162 |
| Management / IO Psychology | 0.079 | 0.068 | -0.054 | 0.213 |
| Neuropsychology | -0.323 | 0.091 | -0.502 | -0.144 |
| Sport and exercise psychology | 0.094 | 0.077 | -0.057 | 0.245 |



**Figure [scatter small].** Scatter plot of statistical power to detect a small effect over time.Dotted lines are 95% confidence intervals, and the solid line is the estimated proportion power of psychology by year, point sizes reflect the relative weighting of articles.



**Figure [scatter medium].** Scatter plot of statistical power to detect a medium effect over time.Dotted lines are 95% confidence intervals, and the solid line is the estimated proportion power of psychology by year, point sizes reflect the relative weighting of articles.



**Figure [scatter large].** Scatter plot of statistical power to detect a medium effect over time. Dotted lines are 95% confidence intervals, and the solid line is the estimated proportion power of psychology by year, point sizes reflect the relative weighting of articles.

### 5.2.8 Bias assessment

Although the included articles do not use traditional significance testing to assess their primary outcome, it is still possible that smaller articles which find more “alarming” results are more likely to be published. In order to assess for signs of publication bias we performed an analogue to Egger’s Test (Egger, Smith, Schneider, & Minder, 1997) by including the number of articles which were surveyed in each study as a moderator. The number of included articles was used instead of sampling variances as the sampling variances are expected to be associated with outcome scores, again due to the fact that estimated mean power levels towards either bound (1 or .05) are expected to have reduced variances. This test showed that sample size was not a significant predictor of average estimated statistical power at any of the benchmarks with parameter estimates for the small, medium, and large effects of b = 0.000, 95% CI [-0.0001, 0.0001], *p* = .45, medium b = 0.000, 95% CI [-0.0001, 0.0001], *p* = .45, and large b = -0.000, 95% CI [-0.0001, 0.0000], *p* = .44, giving no obvious indication of publication bias in this sample.

**5.3 Discussion**

This analysis suggests that there has been little to no change in the statistical power of psychology research over the previous half century. The reporting of statistical power analysis appears to have increased slightly over time, but is still uncommon. These results are unexpected given the large number of papers that have been published arguing for power analysis to be performed as a part of research planning over the last 50 years (e.g., Bezeau & Graves, 2001; Cohen, 1962; Rossi, 1990; Sedlmeier & Gigerenzer, 1989), the increasing availability of user friendly power analysis tools (e.g., Cohen, 1988; Faul et al., 2007), as well as technological innovations (e.g., internet or Amazon Turk studies) and larger undergraduate cohorts that could make larger sample research more tractable at least in many areas of psychological research.

Given that the average effect size seen in the psychology literature has been estimated to be around or even slightly below Cohen’s ‘medium’ effect size (e.g., Bosco, Aguinis, Singh, Field, & Pierce, 2015; Gignac & Szodorai, 2016; Quintana, 2017) and have remained quite stable or even decreased slightly over time [effect size chapter], this suggests that the average psychological study should fail to find significant results as much as 40% of occasions assuming that the effect under study is in fact present. Despite this, over 90% of psychology papers report statistically significant findings (Fanelli, 2010). This suggests that either a large proportion of performed research is going unreported, or that a large amount of research is presented as having found statistically significant findings achieved in some part through p-hacking, HARKing or through the exploitation of researcher degrees of freedom (M Bakker, van Assen, Crompvoets, Ong, & Soderberg, 2017; LeBel, McCarthy, Earp, Elson, & Vanpaemel, 2018; Wicherts et al., 2016).

Given the evidence regarding how poor our intuitions about the likely power and precision of research (e.g., Marjan Bakker, Hartgerink, Wicherts, & van der Maas, 2016; Obrecht, Chapman, & Gelman, 2007; Tversky & Kahneman, 1971), formal sample size planning should play a major role in helping researchers plan their studies. Formal sample size planning (e.g., planning for adequate levels of power, narrow confidence or credible intervals, convincing evidence via Bayes factors, etc.) is an important tool for researchers who wish to ensure that they are not wasting their participant’s time, their own time and limited research funding on experiments which are unlikely to allow them to draw accurate inferences. A variety of research planning packages and programs are freely available and should enable researchers to plan for most statistical analyses (e.g., the R package "SIMSEM" for structural equation modeling; Beaujean, 2014; G\*Power for the most common analyses such as ANOVA, regression or chi-square analysis; Faul et al., 2007; for advice on planning for sufficiently convincing Bayes factors see Schönbrodt & Wagenmakers, 2017; PINT 2.2 for two level hierarchical modeling; Snijders & Bosker, 1993; and "PANGEA" for more complex ANOVA designs; Westfall, 2015). More complex analyses may require consultation with a statistician (Van Meter & Charnigo, 2014).

Editors and reviewers can play a role in supporting the routine performance and reporting of a priori power analysis by requiring a statement of justification for the included sample size following the formal reporting guidelines that have already been established (APA Publications Communications Board Working Group on Journal Article Reporting Standards, 2008; Moher et al., 2010; Moher et al., 2001; Wilkinson, 1999). Requiring the accurate justification of sample sizes as a routine part of research reporting (e.g., stating that the sample size was chosen due to practical constraints such as in the current study, identified through formal sample size planning such as AIPE or power analysis, or even stating that no sample size planning occurred when this is the case) could help establish a norm for these issues to be considered during research planning.

This advice – that researchers should consider the statistical power of their analyses during research planning and that editors should request or even require the reporting of power analyses – is the suggested remedy in almost all of the statistical power reviews included in the current analysis. It has apparently failed to influence the practices of working scientists. It is hard to imagine that saying it again here will result in anything different.

The recent development and rapid uptake of new research, publication and reporting initiatives give some reason for optimism. Preregistration of confirmatory analysis plans can help allow researchers to avoid unwittingly altering their analysis plans and increasing the probability of obtaining a false positive finding (Simmons, Nelson, & Simonsohn, 2011). The use of preprint servers allows researchers to disseminate findings outside of the traditional publication system, subverting publication bias and allowing small scale studies or research to be available for future meta-analysts, helping to avoid effect size inflation. Largescale mutli-lab collaborative efforts like the Psychological Science Accelerator (Moshontz et al., 2018) and the Many Labs projects (Klein et al., 2018) facilitate extremely large scale research, allowing for extremely high powered research even when effect sizes may be small. However, these initiatives still make up an extremely small part of the scientific literature. For research consumers this means we must accept that the research literature likely provides exaggerated effect size estimates ([chapter pub bias] (Stanley, Carter, & Doucouliagos, 2018)) and has a higher false positive error rate than it otherwise would.

### **Limitations**

In interpreting these findings, it’s important to keep in mind that the individual articles in these power surveys are not a random sample from the psychological research literature, and it is difficult to predict whether the sampling choices will tend to underestimate or overestimate the average power of psychological research. It is possible that power surveys are more likely to be performed when a particular area of research is underpowered, which could lead to this analysis underestimating the average statistical power of psychology. This issue only holds for a subset of the included studies, with the other included studies either using convenience samples (e.g., Szucs & Ioannidis, 2017), samples chosen to be broadly representative of a subfield (e.g., Orme & Combs-Orme, 1986), or samples selected to represent high-impact journals in a subfield (e.g., Cashen & Geiger, 2004; Rossi, 1990, a strategy which could upwardly bias estiamtes). Secondly, the included power surveys assume that alpha is set at .05, meaning that these results may overestimate power as they ignore alpha corrections for multiple comparisons which lead to lower power. Power surveys also almost uniformly target tests for which power can be easily estimated or defined, ignoring more sophisticated analyses (e.g., SEM, factor analysis, or multilevel models). This means that this study may underestimate the average power of psychological research if larger studies tend to use these more sophisticated techniques. However, given that simple significance testing is rarely a primary concern for these statistical techniques their exclusion from this analysis may not be unreasonable.

### **Conclusion**

Statistical power to detect small to medium effects appears to be substantially lower than recommended standards and power analysis is rarely reported in psychology research. Statistical power does not appear to have increased over the last 60 years, despite continued criticism of this fact, the advocacy for the use of formal sample size planning techniques the increasing ease of use of. Research consumers should be aware that the average power of psychological science is lower than would be ideal for ‘small’ or ‘medium’ effects, and only acceptably high for ‘large’ effects. Research consumers should make sure they read and interpret the published literature with these facts in mind, take steps to avoid performing underpowered research, and ensure that the results of their analyses are available to future meta-analysts regardless of the statistical significance of their results.

## References

APA Publications Communications Board Working Group on Journal Article Reporting Standards. (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *The American Psychologist, 63*, 839-851. doi:10.1037/0003-066X.63.9.839

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*, 1069-1077. doi:10.1177/0956797616647519

Bakker, M., van Assen, M., Crompvoets, E., Ong, H., & Soderberg, C. (2017). *Restriction of opportunistic use of researcher degrees of freedom in pre-registrations on the Open Science Framework: The human Fallibility of scientists.*

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The Rules of the Game Called Psychological Science. *Perspectives on Psychological Science, 7*, 543-554. doi:10.1177/1745691612459060

Beaujean, A. A. (2014). Sample size determination for regression models using monte carlo methods in r. *Practical assessment, research & evaluation, 19*, 1.

Bezeau, S., & Graves, R. (2001). Statistical power and effect sizes of clinical neuropsychology research. *Journal of Clinical and Experimental Neuropsychology, 23*, 399-406. doi:10.1076/jcen.23.3.399.1181

Bland, J. M. (2009). The tyranny of power: Is there a better way to calculate sample size? *BMJ, 339*. doi:10.1136/bmj.b3985

Bosco, F. A., Aguinis, H., Singh, K., Field, J. G., & Pierce, C. A. (2015). Correlational effect size benchmarks. *Journal of Applied Psychology, 100*, 431-449. doi:10.1037/a0038047

Cashen, L. H., & Geiger, S. W. (2004). Statistical Power and the Testing of Null Hypotheses: A Review of Contemporary Management Research and Recommendations for Future Studies. *Organizational Research Methods, 7*, 151-167. doi:10.1177/1094428104263676

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

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, New Jersey: Erlbaum.

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

DeCoster, J., Sparks, E. A., Sparks, J. C., Sparks, G. G., & Sparks, C. W. (2015). Opportunistic biases: Their origins, effects, and an integrated solution. *American Psychologist, 70*, 499-514. doi:10.1037/a0039191

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ, 315*, 629.

Fanelli, D. (2010). “Positive” Results Increase Down the Hierarchy of the Sciences. *PLOS ONE, 5*, e10068. doi:10.1371/journal.pone.0010068

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*, 175-191. doi:10.3758/bf03193146

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences, 102*, 74-78. doi:<https://doi.org/10.1016/j.paid.2016.06.069>

Grey, C. (2019). verameta. R package

version 0.0.0.9. Retrieved from <https://github.com/softloud/varameta/>

Haase, R. F. (1974). Power analysis of research in counselor education. *Counselor Education and Supervision, 14*, 124-132.

Hedges, L. V. (1992). Modeling Publication Selection Effects in Meta-Analysis. *Statistical Science, 7*, 246-255.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*  Retrieved from <http://methods.sagepub.com/book/methods-of-meta-analysis-3e> doi:10.4135/9781483398105

Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology, 19*, 640-648. doi:10.1097/EDE.0b013e31818131e7

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., . . . Nosek, B. A. (2018). Many Labs 2: Investigating Variation in Replicability Across Samples and Settings. *Advances In Methods and Practices in Psychological Science, 1*, 443-490. doi:10.1177/2515245918810225

LeBel, E. P., McCarthy, R. J., Earp, B. D., Elson, M., & Vanpaemel, W. (2018). A Unified Framework to Quantify the Credibility of Scientific Findings. *Advances In Methods and Practices in Psychological Science*, 2515245918787489. doi:10.1177/2515245918787489

Marín-Martínez, F., & Sánchez-Meca, J. (2009). Weighting by Inverse Variance or by Sample Size in Random-Effects Meta-Analysis. *Educational and psychological measurement, 70*, 56-73. doi:10.1177/0013164409344534

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does “failure to replicate” really mean? *American Psychologist, 70*, 487-498. doi:10.1037/a0039400

Miller, J. J. (1978). The Inverse of the Freeman – Tukey Double Arcsine Transformation. *The American Statistician, 32*, 138-138. doi:10.1080/00031305.1978.10479283

Moher, D., Hopewell, S., Schulz, K. F., Montori, V., Gøtzsche, P. C., Devereaux, P. J., . . . Altman, D. G. (2010). CONSORT 2010 Explanation and Elaboration: Updated guidelines for reporting parallel group randomised trials. *Journal of Clinical Epidemiology, 63*, e1-e37. doi:10.1016/j.jclinepi.2010.03.004

Moher, D., Schulz, K. F., & Altman, D. G. (2001). The CONSORT statement: revised recommendations for improving the quality of reports of parallel group randomized trials. *BMC Medical Research Methodology, 1*, 2. doi:10.1186/1471-2288-1-2

Morris, C. N. (1983). Parametric Empirical Bayes Inference: Theory and Applications. *Journal of the American Statistical Association, 78*, 47-55. doi:10.1080/01621459.1983.10477920

Moshontz, H., Campbell, L., Ebersole, C. R., Ijzerman, H., Urry, H. L., Forscher, P. S., . . . Chartier, C. R. (2018). The Psychological Science Accelerator: Advancing Psychology Through a Distributed Collaborative Network. *Advances In Methods and Practices in Psychological Science, 1*, 501-515. doi:10.1177/2515245918797607

Obrecht, N. A., Chapman, G. B., & Gelman, R. (2007). Intuitive tests: Lay use of statistical information. *Psychonomic Bulletin & Review, 14*, 1147-1152. doi:10.3758/BF03193104

Orme, J. G., & Combs-Orme, T. D. (1986). Statistical power and Type II errors in social work research. *Social Work Research and Abstracts, 22*, 3-10. doi:10.1093/swra/22.3.3

Poisot, T. (2010). Getting data from an image *Data visualization (in R).* Retrieved from <https://rdataviz.wordpress.com/2010/03/05/getting-data-from-an-image-introductory-post/>

Quintana, D. S. (2017). Statistical considerations for reporting and planning heart rate variability case‐control studies. *Psychophysiology, 54*, 344-349.

R Development Core Team. (2018). R: A language and environment for statistical computing (Version 3.5.0). Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <http://www.R-project.org>

Rossi, J. S. (1990). Statistical power of psychological research: what have we gained in 20 years? *J Consult Clin Psychol, 58*, 646-656.

Schönbrodt, F. D., & Wagenmakers, E.-J. (2017). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*. doi:10.3758/s13423-017-1230-y

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105*, 309-316. doi:10.1037/0033-2909.105.2.309

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology. *Psychological Science, 22*, 1359-1366. doi:10.1177/0956797611417632

Smith, D. R., Hardy, I. C., & Gammell, M. P. (2011). Power rangers: No improvement in the statistical power of analyses published in Animal Behaviour. *Animal Behaviour, 81*, 347-352.

Snijders, T. A., & Bosker, R. J. (1993). Standard errors and sample sizes for two-level research. *Journal of Educational Statistics, 18*, 237-259. doi:10.2307/1165134

Stanley, T. D., Carter, E. C., & Doucouliagos, H. (2018). What Meta-Analyses Reveal About the Replicability of Psychological Research. *Psychological Bulletin*.

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

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin, 76*, 105-110. doi:10.1037/h0031322

Van Meter, E., & Charnigo, R. (2014). Strengthening Interactions between Statisticians and Collaborators: Objectives and Sample Sizes. *Journal of biometrics & biostatistics, 5*, e127. doi:10.4172/2155-6180.1000e127

Viechtbauer, W. (2010). Conducting Meta-Analyses in R with the metafor Package. *Journal Of Statistical Software, 36*, 48. doi:10.18637/jss.v036.i03

Wan, X., Wang, W., Liu, J., & Tong, T. (2014). Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range. *BMC Medical Research Methodology, 14*, 135. doi:10.1186/1471-2288-14-135

Westfall, J. (2015). PANGEA: Power ANalysis for GEneral Anova designs. *Unpublished manuscript. Available at* [*http://jakewestfall.org/publications/pangea.pdf*](http://jakewestfall.org/publications/pangea.pdf).

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, M. A. L. M. (2016). Degrees of Freedom in Planning, Running, Analyzing, and Reporting Psychological Studies: A Checklist to Avoid p-Hacking. *Frontiers in Psychology, 7*, 1832. doi:10.3389/fpsyg.2016.01832

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

Woolley, T. W. (1983). A comprehensive power-analytic investigation of research in medical education. *Journal of Medical Education, 58*, 710-715.

### Supplementary materials 1. List of articles included in the primary analysis

|  |  |  |  |
| --- | --- | --- | --- |
| Author(s) (year) | Area studied | Years studied | Number of surveyed Articles |
| Sindelar et al. (1988) | Education | 1932-1974 | 35 |
| Cohen (1962) | General Psychology | 1960 | 70 |
| Levenson (1980) | Clinical Psychology/Psychiatry | 1946-1977 | 30 |
| Osborne (2008) | Education | 1969 | 55 |
| Levenson (1980) | Clinical Psychology/Psychiatry | 1961-1977 | 26 |
| Haase (1974) | Education | 1968-1971 | 60 |
| Penick & Brewer (1974) | Education | 1969-1970 | 66 |
| Brewer (1972) | Education | 1969-1971 | 85 |
| Jones & Brewer (1972) | Sport and exercise psychology | 1969-1971 | 106 |
| Katzer & Sodt (1973) | Communication | 1971-1972 | 31 |
| Rothpearl, Mohs & Davis (1981) | Clinical Psychology/Psychiatry | 1966-1980 | 35 |
| Chase, & Tucker (1975) | Clinical Psychology/Psychiatry | 1973 | 46 |
| Kroll & Chase (1975) | Clinical Psychology/Psychiatry | 1973-1974 | 62 |
| Chase & Chase (1976) | General Psychology | 1974 | 121 |
| Chase & Baran (1976) | Communication | 1974 | 48 |
| Christensen & Christensen (1977) | Management / IO Psychology | 1975 | 43 |
| Schoonen (1991) | Clinical Psychology/Psychiatry | 1964-1989 | 13 |
| Schoonen (1991) | Clinical Psychology/Psychiatry | 1964-1989 | 9 |
| Woolley & Dawson (1983) | Education | 1977-1980 | 73 |
| Sawyer (1981) | Management / IO Psychology | 1979 | 23 |
| Daly & Hexamer (1983) | Education | 1978-1980 | 57 |
| Orme & Combs-Orme (1986) | Clinical Psychology/Psychiatry | 1977-1984 | 79 |
| Orme & Tolman (1986) | Education | 1976-1985 | 64 |
| Woolley (1983) | Education | 1980-1982 | 100 |
| Rossi (1990) | General Psychology | 1982 | 221 |
| Acklin & McDowell (1995) | Clinical Psychology/Psychiatry | 1975-1991 | 158 |
| Mazen, Hemmasi & Lewis (1987) | Management / IO Psychology | 1982-1984 | 44 |
| Mazen et al. (1987) | Management / IO Psychology | 1984 | 84 |
| Sedlmeier & Gigerenzer (1989) | Clinical Psychology/Psychiatry | 1984 | 54 |
| Dilullo (1998) | Education | 1975-1994 | 41 |
| Whittington, Podd, & Kan (2000) | Clinical Psychology/Psychiatry | 1978-1997 | 46 |
| Aguinis et al. (2005) | Management / IO Psychology | 1977-1998 | 106 |
| Kazantzis (2000) | Clinical Psychology/Psychiatry | 1980-1998 | 27 |
| Brown & Hale (1992) | Clinical Psychology/Psychiatry | 1989 | 24 |
| Tomcho, Thomas & Foels (2009) | Education | 1974-2006 | 193 |
| Kosciulek & Szymanski (1993) | Clinical Psychology/Psychiatry | 1990-1991 | 32 |
| Mone, Mueller, & Mauland (1996) | Management / IO Psychology | 1992-1994 | 210 |
| Clark-Carter (1997) | General Psychology | 1993-1994 | 54 |
| Cashen, Geiger (2004) | Management / IO Psychology | 1990-1999 | 43 |
| Borkowski, Welsh & Zhang (2001) | Education | 1993-1997 | 96 |
| Maddock & Rossi (2001) | Clinical Psychology/Psychiatry | 1997 | 187 |
| Okumura & Sakamoto (2011) | Clinical Psychology/Psychiatry | 1990-2006 | 311 |
| Bezeau & Graves (2001) | Clinical Psychology/Psychiatry | 1998-1999 | 66 |
| Osborne (2008) | Education | 1998-1999 | 96 |
| Ward (2002) | General Psychology | 2000 | 157 |
| Woods et al. (2006) | Neuropsychology | 1997-2004 | 30 |
| Overland (2014) | Education | 2000-2010 | 125 |
| Schweizer & Furley (2016) | Sport and exercise psychology | 2009-2013 | 337 |
| Schweizer & Furley (2016) | Sport and exercise psychology | 2009-2013 | 117 |
| Schweizer & Furley (2016) | Sport and exercise psychology | 2009-2013 | 109 |
| Schweizer & Furley (2016) | Sport and exercise psychology | 2009-2013 | 84 |
| Schweizer & Furley (2016) | Sport and exercise psychology | 2009-2013 | 45 |
| Szucs & Ioannidis (2017) | General Psychology | 2011-2014 | 3801 |

*Table [subfield].*

The subfields covered in the included power surveys.

|  |  |
| --- | --- |
| Subfield | n |
| Multiple fields of psychology research | 6 |
| Clinical Psychology/Psychiatry | 17 |
| Education | 14 |
| Management / IO Psychology | 7 |
| Sport and exercise psychology | 6 |
| Communication | 2 |
| Neuropsychology | 1 |

### Supplemental material 2. Search parameters

Table [database search].Databases and search terms used for data collection for a systematic review of power surveys performed on psychological research. Search performed on the 11th September 2017.

|  |  |  |
| --- | --- | --- |
| Database | Search terms | Number of records |
| Psychinfo, Ovid Interface | ("\*power\*" or "Determination" or "estimat\*" or "sampl\*").m\_titl. and ("power analysis" or "Statistical Power" or "Sample Size Estimation" or "Sample Size Determination" or "Sample size selection").mp. | 916 |
| Web of Science Core Collection | SU = (Psychology OR Psychiatry OR "Mathematical Methods In Social Sciences") AND TI = (Power\* OR Sampl\*) AND TS = ("power analysis" or "Statistical Power" or "Sample Size") | 1,072 |
| Total number of articles |  | 1988 |
| De-duplicated library |  | 1489 |

### Supplemental material 3. Code book for data collection

Code books for studies:

Coding rules:

* If a paper reports median power estimates separated by year, enter each year’s values into the database separately by year
* For studies which report median power estimates broken down into other categories (e.g., by journal), take the highest level (e.g., the values for the entire sample) at which median power levels are reported. If medians are not reported, record data at the highest level (e.g., “by APA published journals” as opposed to “by journal”)
* If a paper calculates observed power (i.e., power to detect the observed effect size of each study), exclude
* When studies include multiple investigations of the same articles (e.g., studies examining the power of mixed effects study designs to investigate power for main and interaction effects) report the higher estimate.
* If a paper calculates power for meta-analytically derived average, exclude, but retain data
* If a paper calculates power for other values, note and include (but exclude from meta-analysis)
* If power values are stated using multiple effect sizes, record the stated Cohen’s d, but preferably note the source for the estimates (e.g., “Cohen, 1988”)
* Note if an article explicitly notes having used any effect size apart from hedges g (i.e., the effect size that is often called Cohen’s d in papers, but which actually uses Hedges’ estimator)

|  |  |
| --- | --- |
| Variable ID | Explanation |
| id | unique paper ID |
| Author | Author name |
| Title | Paper title |
| Journal | Journal of publication |
| Year | Year of publication |
| exclude | Whether the paper should be excluded (include reason in "Notes" variable) |
| SamplingStrategy | The sampling strategy used to select the articles included in a particular paper, copy and pasted directly from article |
| SampleSource | Journals covered in article's sample, or a brief description of the article's sample (e.g., "articles included in Example's (1999) meta-analysis of the impact of x on y") |
| YearsStudied | The range of years covered in an article (e.g., 2001-2009) |
| MedianYear | Median of years included in an article, round down (e.g., for "2011 - 2013", "2012") |
| TargetTest | Statistical Tests that were included in the power survey (e.g., "all t-tests", "all statistical tests") |
| SubfieldClassification | Subfield of research examined in the power survey (e.g., "psychology", "clinical neuroscience", "organisational psych" etc.) |
| PowerEstimationTechnique | Copy and pasted copy of the way that power was reported to have been calculated (e.g., "Power for t-tests and F-tests was estimated using g\*power, a correlation of .5 was assumed between repeated measures") |
| AmalgomationMethod | Amalgamation method used, are the reported power summery statistics from individual tests within articles, or averaged at the article level or was power calculated for the "main test", etc. (e.g., "mean power of articles", "power of main statistical test", etc.) |
| DistinguishedStatisticalTestsAppropriatly | Did the method distinguish between different types of statistical procedures appropriately (if not wrong, then still "yes"?) - i.e., was power calculated for the particular types of statistical tests that were included in the power survey |
| NumberOfArticles | Number of articles included in power survey |
| NumberOfTests | Number of tests included in power survey |
| EffectSizeUsed | Effect size used (e.g., Cohen's d, Hedge's g, r, link to equation), or source for set of effect size benchmarks used (e.g., Cohen 1988) |
| SmallEffectBenchmark | Small effect benchmark used, or source for small benchmarks (e.g., ".2" or "Cohen 1988") |
| MediumEffectBenchmark | Medium effect benchmark used, or source for medium benchmarks (e.g., ".5" or "Cohen 1988") |
| LargeEffectBenchmark | Large effect benchmark used, or source for large benchmarks (e.g., ".8" or "Cohen 1988") |
| **PowerAtSmallEffectMedian** | Median power at small effect benchmark |
| FirstQuartilePowerAtSmall | First quantile of power at small effect benchmark |
| ThirdQuartilePowerAtSmall | Third quantile of power at small effect benchmark |
| **PowerAtMediumEffectMedian** | Median power at medium effect benchmark |
| FirstQuartilePowerAtMedium | First quantile of power at medium effect benchmark |
| ThirdQuartilePowerAtMedium | Third quantile of power at medium effect benchmark |
| **PowerAtLargeEffectMedian** | Median power at large effect benchmark |
| FirstQuartilePowerAtLarge | First quantile of power at large effect benchmark |
| ThirdQuartilePowerAtLarge | Third quantile of power at large effect benchmark |
| PowerAtSmallEffectMean | Mean power at small effect benchmark |
| PowerAtMediumEffectMean | Mean power at medium effect benchmark |
| PowerAtLargeEffectMean | Mean power at large effect benchmark |
| SDPowerAtSmall | Standard deviation at small effect benchmark |
| SDPowerAtMedium | Standard deviation at medium effect benchmark |
| SDPowerAtLarge | Standard deviation at large effect benchmark |
| SampleMedian | Median sample size |
| FirstQuartileSampleSize | First quartile of sample sizes recorded in power survey |
| ThirdQuartileSampleSize | Third quartile of sample sizes recorded in power survey |
| SampleMean | Mean of sample sizes recorded in power survey |
| SampleSizeSD | Standard deviation of sample sizes recorded in power survey |
| Notes | Any notes? Record reason for exclusion here |
| SDSmallAlgEstFromCDT | Standard deviation at small effect benchmark as estimated from frequency table |
| SDMedAlgEstFromCDT | Standard deviation at medium effect benchmark as estimated from frequency table |
| SDLargeAlgEstFromCDT | Standard deviation at large effect benchmark as estimated from frequency table |
| Solutions | Copy and pasted copy of the authors suggested solutions |
| SampleMin | Minimum sample size included in power survey |
| SampleMax | Maximum sample size included in power survey |
| **PowerSmallMin** | Minimum power value at small benchmark |
| PowerSmallMax | Maximum power value at small benchmark |
| PowerMedMin | Minimum power value at medium benchmark |
| PowerMedMax | Maximum power value at medium benchmark |
| PowerLargeMin | Minimum power value at large benchmark |
| PowerLargeMax | Maximum power value at large benchmark |
| NotInEnglish | Binary for reasons to have excluded articles - English text not available |
| FullTextUnavaliable | Binary for reasons to have excluded articles - full text not available |
| NoPowerOrSampleSizesReported | Binary for reasons to have excluded articles - does not calculate power at benchmark levels / report sample sizes for a body of research |
| DuplicateData | Binary for reasons to have excluded articles - duplicate data of another in this sample |
| NoPowerButSampleSizesReported | Binary for reasons to have excluded articles - does not calculate power at benchmark levels but does report sample size |
| OutsideScope | Binary for reasons to have excluded articles - area covered in power survey outside of the scope of the current research |

### Supplementary material 4. Sensitivity analysis

**Sensitivity analyses**

To investigate whether the results are sensitive to data imputation choices, multiple imputation methods were used (see Table S4 – 1 for model coefficient output). Firstly, analyses were run excluding any studies for which any data had to be estimated or imputed (i.e., only including those studies which reported a mean power estimate and standard deviation for a given effect size benchmark). We also imputed variances using multiple imputation, and using the median, maximum and minimum variances instead of mean imputation. All of these changes lead to differences in the coefficient for year of less than .005, differences in the intercept of less than .04, and no difference in statistical significance at the .05 alpha level.

To check whether the results are sensitive to analysis choices, we also ran these models without an effect for year (i.e., as a traditional random effects meta-analysis, see Table S4 - 2), and without random effects for study or area of research (as was preregistered analysis, see Table S4 - 3). The estimated change in power per year is within .003 of the main results presented in the main text of this article when the random effects are not included in the model, and the intercept parameters for either alternative parameterization model do not change by more than .04.

Because power is bounded between 0.05 and 1 for all included studies, studies which found estimated mean powers that are close to either bound will have lower variances. This means that the typical inverse variance weighting approach to meta-analysis (Hedges, 1992) will systematically overweight these studies. In order to account for this issue, meta-analyses were also run weighting by the number of articles included in papers following Hunter and Schmidt (2004). Although this approach avoids the issue of range restriction overweighting studies which showed mean powers near either bound, it is a less efficient estimator (Marín-Martínez & Sánchez-Meca, 2009). Again, the results showed that there was relatively little difference in outcome, with the estimated effect of year changing by less than .002, and the estimated average power of psychology research changing by -.001, -.002, and -.04 at the small, medium and large effect size benchmarks respectively, see row 3 of Table S4.1 for model output.

Table S4 - 1

*Model output for various imputation methods, model including year as a moderator, random effects for area of research and study and estimate level.*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Analysis type | Mean Power | | | Estimated change in average power per year | | |
| Small | Medium | Large | Small | Medium | Large |
| Mean imputationa | 0.233 | 0.620 | 0.797 | 0.000 | 0.001 | -0.001 |
| Multiple imputation of variancesb | .234 | .620 | 0.799 | 0.000 | 0.001 | -0.000 |
| Mean variance imputation, weighting by *n* articles | 0.232 | 0.624 | 0.84 | -0.001 | -0.002 | -0.003 |
| No values estimated or imputed | 0.23 | 0.582 | 0.819 | -0.002 | -0.003 | -0.002 |
| Maximum variance imputation | 0.233 | 0.621 | 0.796 | 0.000 | -0.002 | -0.001 |
| Minimum variance imputation | 0.234 | 0.621 | 0.798 | 0.000 | 0.001 | -0.001 |
| Median variance imputation | 0.233 | 0.62 | 0.797 | 0.000 | 0.001 | -0.001 |

Note: aModel is the model reported in the main text of this article. bModel was run on 20 imputed datasets with variances imputed using predictive mean matching, all articles where variances or standard deviations were not directly reported in the paper had their variances imputed.

Table S4 - 2

*Model output for a random effects meta-analysis with random effects for effect, article and area of research, but no fixed effect for year, with various missing data imputation methods.*

|  |  |  |  |
| --- | --- | --- | --- |
| Analysis type | Mean Power | | |
| Small | Medium | Large |
| Mean variance imputation | 0.233 | 0.621 | 0.794 |
| Mean variance imputation, weighting by *n* articles | 0.210 | 0.593 | 0.798 |
| No values estimated or imputed | 0.238 | 0.593 | 0.778 |
| Maximum variance imputation | 0.233 | 0.621 | 0.793 |
| Minimum variance imputation | 0.234 | 0.622 | 0.795 |
| Median variance imputation | 0.233 | 0.621 | 0.794 |

Table S4 - 3

*Model output for various imputation methods, model including year as a moderator, and random effects for effect, but not subfield of research or for each power survey, with various missing data imputation methods.*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Analysis type | Mean Power | | | Estimated change in average power per year | | |
| Small | Medium | Large | Small | Medium | Large |
| Mean imputation | 0.227 | 0.62 | 0.833 | -0.001 | -0.002 | -0.001 |
| Mean variance imputation, weighting by *n* articles | 0.23 | 0.624 | 0.838 | -0.001 | -0.002 | -0.003 |
| No values estimated or imputed | 0.217 | 0.582 | 0.818 | -0.002 | -0.003 | -0.002 |
| Maximum variance imputation | 0.226 | 0.618 | 0.832 | -0.001 | -0.002 | -0.001 |
| Minimum variance imputation | 0.228 | 0.621 | 0.834 | -0.001 | -0.002 | -0.001 |
| Median variance imputation | 0.228 | 0.621 | 0.833 | -0.001 | -0.002 | -0.001 |

### Supplementary material 5: Estimating the proportion of articles which report a power analysis in the psychology literature

Sample characteristics

Fifteen of the examined articles assessed how often power analyses were reported, meeting the inclusion criteria for the secondary analysis. Two additional articles were found through searches of these articles’ references lists, leading to 17 articles being included in the secondary analysis. Eight out of twenty-one of the included estimates examined research from clinical studies (e.g., clinical randomized controlled trials of psychological therapies), four examined educational research, three examined management / IO psychology, three neurocognitive/neuroimaging research, two examined general psychology and one examined communication research.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Paper | Years Studied | Subfield | Number of articles examined | Percent Reporting a PA |
| Woods et al., (2006) | 1997-2004 | Clinical | 30 | 0% |
| Thombs & Rice (2016) | 2013-2015 | Clinical | 89 | 3% |
| Ward (2002) | 2000 | Psychology | 103 | 7% |
| Dilullo (1998) | 1975-1994 | Education | 41 | 0% |
| ClarkCarter (1997) | 1993-1994 | Psychology | 54 | 2% |
| Larson & Carbine (2017) | 2010-2015 | Neuropsychology | 100 | 0% |
| Guo et al., (2014) | 2010-2011 | Neuropsychology | 100 | 1% |
| Sedlmeier & Gigerenzer (1989) | 1984 | Clinical | 64 | 0% |
| Short, Ketchen, & Palmer (2002) | 1980-1989 | Industrial-organisational | 149 | 3% |
| Short, Ketchen, & Palmer (2002) | 1990-1999 | Industrial-organisational | 288 | 6% |
| Katzer & Sodt (1973) | 1971-1972 | Communication | 31 | 6% |
| Kosciulek & Szymanski (1993) | 1990-1991 | Clinical | 32 | 3% |
| Osborne (2008) | 1969 | Education | 55 | 0% |
| Osborne (2008) | 1998-1999 | Education | 96 | 2% |
| Woolley & Dawson (1983) | 1977-1980 | Education | 193 | 0% |
| Cashen & Geiger (2004) | 1990-1999 | Industrial-organisational | 43 | 7% |
| Bezeau & Graves (2001) | 1998-1999 | Clinical | 66 | 3% |
| Faulkner, Fidler & Cumming (2008) | 1999-2003 | Clinical | 193 | 18% |
| Crosby et al. (2006) | 1980 | Clinical | 152 | 0% |
| Crosby et al. (2006) | 1990 | Clinical | 152 | 0% |
| Crosby et al. (2006) | 2000 | Clinical | 152 | 2% |

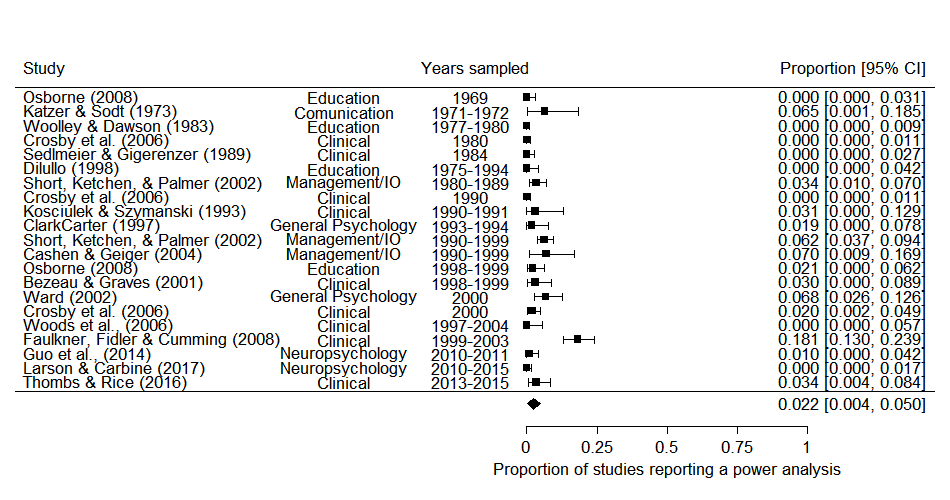
### 5.2.9 Secondary analysis:

A mixed effects meta-regression was conducted to examine the proportion of studies which report a power analysis and to estimate the change in power analysis reporting rates over time.

Proportions were transformed using the Tukey-Freeman Arcsine Transformation ( in the equation above), as this can act to normalize the sampling distributions of proportions (Miller, 1978). The mean year of the range of studies included in each paper was entered as a predictor in the meta-regression (, after being mean-centered making the intercept interpretable as the estimated proportion of studies for which power analyses are performed in the mean year included in this survey (1993). Articles which reported estimates for different year ranges separately in the same paper (e.g., 1980-1982 and 1990-1992) were entered into these meta-analyses separately, and random effects were included for individual estimates (, survey (), and area of research ( to account for non-independence between estimates and as the underlying population parameter being estimated in each study cannot be assumed to be the same. Restricted maximum likelihood estimation was used.

#### Results

The estimated mean proportion of researchers reporting a power analysis was 2.2%, 95% CIs [0.4%, 5.0%]. There is a very small estimated yearly increase in estimated power analysis reporting rates over time of 0.004 95% CI [0.001, 0.007] per year in Freeman-Tukey double arcsine transformed units. See Figure *[Secondary meta-regression with exclusions]* for a meta-regression scatterplot of the datapoints over time and *Figure [Secondary meta-analysis with exclusions]* for a forest plot of this analysis. The is significant heterogeneity in the proportion of studies reporting a power analysis, QE(19) = 85.86, *p* < .001, unsurprisingly given the heterogenous population included in this body of research.



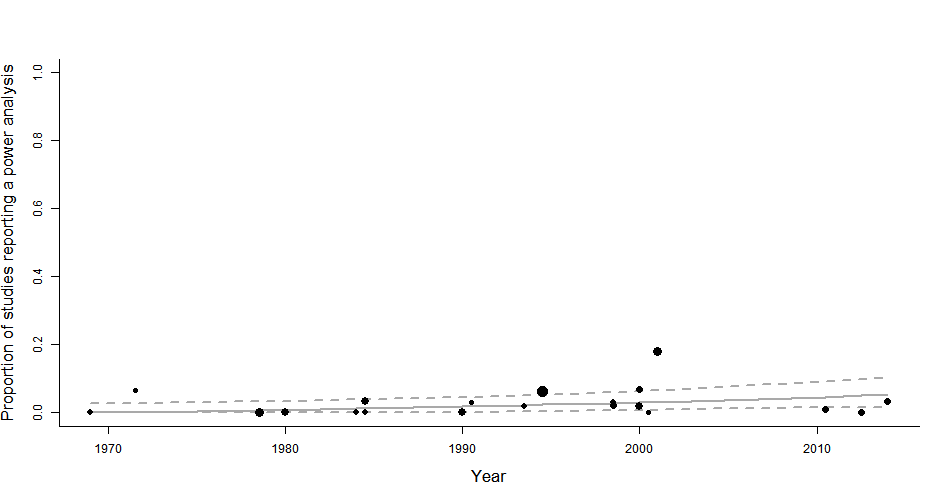
*Figure [Secondary meta-analysis with exclusions]*. Forest plot displaying the findings from seventeen examinations of the proportion of articles which reported a power analysis in psychology research.

Table [Meta-regression].

*Meta-regression of double arcsine transformed proportions of studies reporting a power analysis by (mean standardized) median year covered in each power analytic study. Note values in this table are Freeman-Tukey double arcsine transformed.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Coefficient | *b* | *b*  95% CI  [LL, UL] | *p* | Random effects and variance |
| Intercept | .168 | [0.0964 0.2393] | < .001 |  |
| Year | 0.003 | [0.0006 0.0067] | .02 |  |
|  |  |  |  | Effect σ2 = 0.000, n = 21 |
|  |  |  |  | Article σ2 = 0.009, n = 17 |
|  |  |  |  | Subfield σ2 = 0.003, n = 6 |
|  |  |  |  | QE(19) = 107.72, *p* < .001 |

*Note.* *LL* and *UL* indicate the lower and upper limits of a confidence interval, respectively.

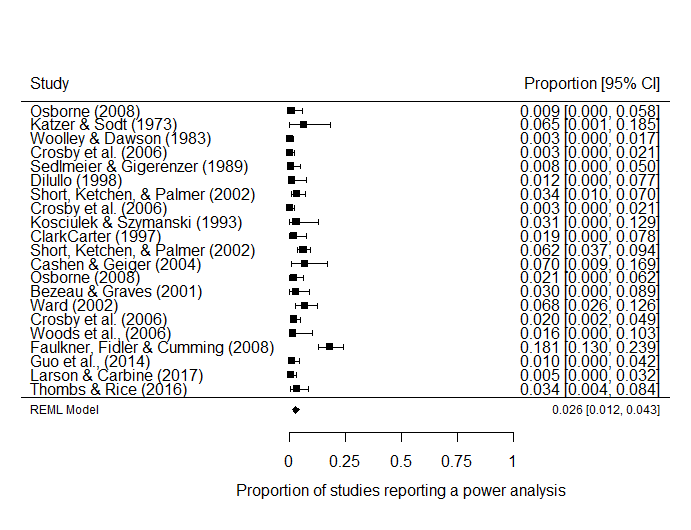


*Figure [Secondary meta-regression with exclusions]*. Meta-regression scatterplot of the effect of time (median year included in each study’s sample) on the proportion of sampled psychology articles reporting a power analysis. Dotted lines are 95% confidence intervals, and the solid line is the estimated proportion of studies reporting a power, circle sizes reflect the relative weighting of each article.

There are two main reasons to question the generalizability of the secondary analysis to psychology research more broadly. Firstly, many of the included literature surveys in this secondary analysis are from clinical psychology research. Secondly, there are few recent studies, and those that were published in the last 10 years only examine clinical and neuropsychology research. However, the results are so low that even if this analysis underestimated the proportion of articles reporting a power analysis by an order of magnitude, power analyses would still be quite rare. See below for sensitivity analyses of these results, and for a model which does not include random effects for study or area of research (as the preregistration did not specify that these effects would be included).

**Sensitivity analyses**

See figure [Secondary meta-analysis] for a forest plot and model output from a restricted maximum likelyhood model without any random predictors using the findings from seventeen examinations of the proportion of articles which reported a power analysis in psychology research. See Table [Meta-regression 2] for the model summary and *Figure [Secondary meta-regression with exclusions]* for a Meta-regression scatterplot of the effect of time (mean standardized, average year included in each study’s sample) on the proportion of sampled psychology articles reporting a power analysis, not accounting for random effects for article or area of research.



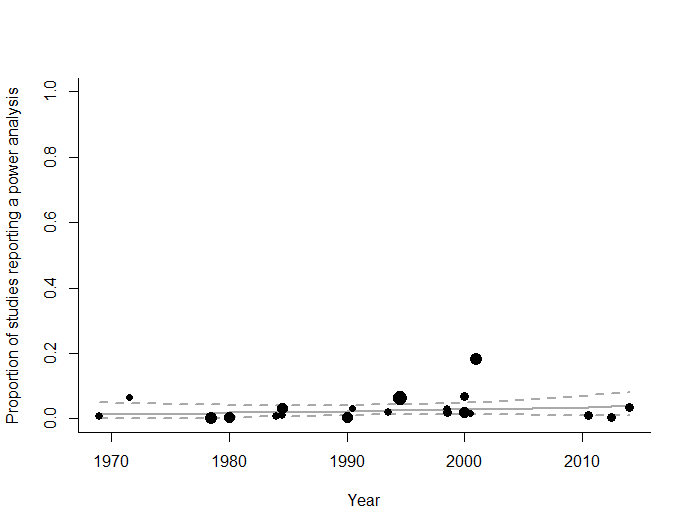
*Figure [Secondary meta-analysis]*. Forest plot displaying the findings from seventeen examinations of the proportion of articles which reported a power analysis in psychology research.

Table [Meta-regression].

*Meta-regression of double arcsine transformed proportions of studies reporting a power analysis by (mean standardized) median year covered in each power analytic study*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Predictor | *b* | *b*  95% CI  [LL, UL] | *p* | Fit |
| (Intercept) | -.1795 | [0.137, 0.222] | <.001 |  |
| Year | 0.016 | [-0.002, 0.005] | .384 |  |
|  |  |  |  | *R2* = .012 |
|  |  |  |  | 𝛕2 = .007 (SE = 0.003) |
|  |  |  |  | I2 = .738 |

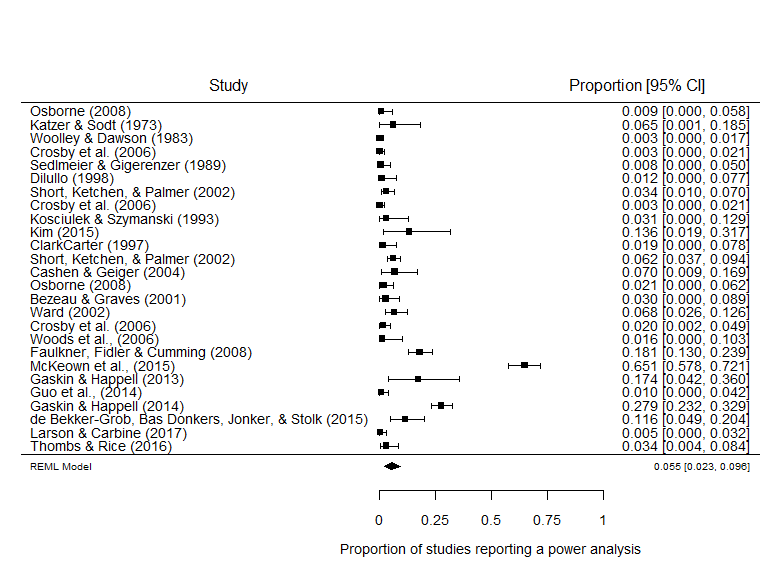
*Note.* *LL* and *UL* indicate the lower and upper limits of a confidence interval, respectively.



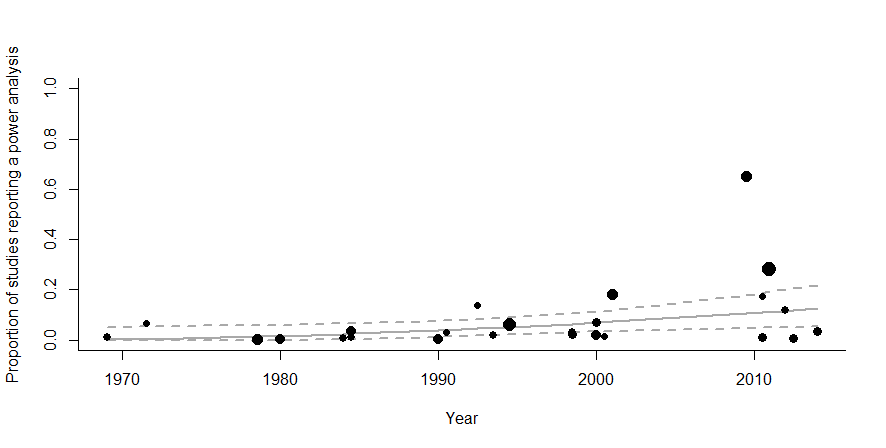
*Figure [Secondary meta-regression with exclusions]*. Meta-regression scatterplot of the effect of time (median year included in each study’s sample) on the proportion of sampled psychology articles reporting a power analysis. Dotted lines are 95% confidence intervals, and the solid line is the estimated proportion of studies reporting a PA, circle sizes reflect the relative weighting of articles.

**Including medical studies**

The secondary analysis excluded five studies which did not meet the pre-registered inclusion criteria, but for which data was collected. McKeown et al., (2015), de Bekker-Grob, Bas Donkers, Jonker, & Stolk (2015), Kim (2015), Gaskin & Happell (2014), and Gaskin & Happell (2013), and which examined, respectively, pharmacological and surgical interventions for pain, diverse discrete-choice experiments in healthcare, vision and blindness studies, and two studies which focused on nursing. While these studies do not meet the pre-registered inclusion criteria due to the fact that all three articles primarily include studies which examine psychiatric drugs or treatments for psychological patients it seems appropriate to consider the results of this analysis including these four additional studies.



*Figure [Secondary meta-analysis without regression and with medical studies]*. Forest plot of studies examining the proportion of articles which reported a power analysis in psychology research broadly defined. This analysis includes four articles not included in the previous study.



*Figure [Secondary meta-regression with medical studies]*. Meta-regression scatterplot of the effect of time (median year included in each study’s sample) on the proportion of psychology articles reporting a power analysis. Dotted lines are 95% confidence intervals, solid line displays the estimated proportion.

The meta-analytic estimate of the average proportion of studies which report a power analysis is slightly higher .055, 95% Cis [.023, .096]. As can be seen in figure *[Secondary meta-regression with medical studies],* a larger effect of time can be seen, b = 0.0065, se = 0.0028, z = 2.2827, *p* = .0224, 95% CI [0.0009, 0.0120]. However, this larger increase over time is clearly driven by the much larger proportion of studies which report a power analysis in medical research (Bland, 2009), with 65% of papers reporting a power analysis in McKeown et al., (2015), 28% in Gaskin & Happell (2014) and 17% in Gaskin & Happell (2013), and given that these papers happen to be recent, they cause an apparent increase in the proportion of studies reporting a power analysis over time.