Chapter 5. Systematic Review and Meta-analysis of Statistical power surveys of psychological research literatures

**5.1 Introduction**

Statistical power describes the probability of a frequentist statistical test rejecting the null hypothesis given a specific alternative hypothesis. 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. This means that 52% of the article published in the journal should have failed to reach statistical significance due to sampling variability alone if they were studying a ‘medium’ effect. Cohen used this fact to begin to argue for reform in the way that psychology research is performed and reported. Early power analysis advocates such as Cohen initially focused their criticism on the degree to which underpowered research would waste research funds and researcher time (Cohen, 1962, 1988). However in recent years underpowered research has been pointed to as one of the driving factors of the replication crisis in psychology (Maxwell, Lau, & Howard, 2015), as in the presence of publication and reporting biasesseveral negative consequences result if studies are routinely underpowered to reliably detect true effect sizes. Under publication bias, lower average power leads to effect size exaggeration and increased false positive error rates among published articles (Marjan Bakker, van Dijk, & Wicherts, 2012; DeCoster, Sparks, Sparks, Sparks, & Sparks, 2015; Ioannidis, 2008).

Since the publication of Cohen’s 1962 article over 60 power surveys have been performed systematically assessing the statistical power of bodies of psychological research. Various 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 various statistical power analysis computer programs (e.g., Faul, Erdfelder, Lang, & Buchner, 2007). An important question is whether these efforts have caused any change in the statistical power of psychology research. The current study examines the large body of literature published since Cohen’s initial power survey which examines statistical power in various areas of psychological research in order to (a) estimate the average power of psychology research, and (b) how statistical power has changed over the last half century.

Given that many of the included power surveys suggest that power analysis should be performed as part of research planning, a related and crucial question is whether power analysis performance rates have increased over time. Given that The American Psychological Association and CONSORT reporting guidelines have suggested that justification for the sample size included in research combined with the rapid increase in the availability and accessibility of power analysis tools, one might expect that the rate of power analysis reporting would have increased over time (APA Publications Communications Board Working Group on Journal Article Reporting Standards, 2008; Moher et al., 2010; Moher, Schulz, & Altman, 2001; Wilkinson, 1999). In order to estimate the proportion of studies which report having performed a power analysis, this study also uses multilevel meta-regression to examine question of how often published empirical research papers report having performed a power analysis during research planning. This allows for an assessment of (c) how common it is for a power analysis to be reported to have been performed in published psychology research, and (d) whether there has been an increase in the number of papers reporting having performed a power analysis over the last 30 years of psychological research.

**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, and before any analysis or summary statistics had been calculated. The pre-registration and pilot data is available at <https://osf.io/n6jfd/>, see table [pre-registration] for a list of deviations from the pre-registered protocol.

Table [preregistration].

*Deviations from preregistered protocol.*

|  |  |
| --- | --- |
| Preregistered | Performed, justification |
| Missing data to be excluded | Means and variances were imputed as large numbers of studies has some missing data. Analyses were also run without data imputation as was preregistered (see supplementary material 3). |
| Medians to be the subject of estimates | Meta-analysis estimated mean not medians. Due to the fact that means were reported slightly more often, equally often (in 45 compared to 47 articles), methods of estimating means from medians (with quartiles or other additional information) are more well developed, and as the standard error of means is smaller than that of medians (cerates paribus). |
| No estimation method was preregistered | Restricted maximum likelihood estimation used |
| No method of accounting for non-independence caused by different samples being reported in the same paper and co-occurrence of area of research | Multilevel meta-analysis used, including study and area of research to account for non-independence of results reported within the same paper and area of research. Model without any random effect were also performed and reported as sensitivity analyses. See appendix 4 for the output of these different models. |
| Reported sample sizes were also going to be used as an outcome measure | Not performed as few articles (7) reported mean sample sizes |
| Specific areas of research were preregistered for studies to be classified into. | “Sport and exercise psychology” and “communication research” were added as additional fields of research. |
| No random effects were specified in the secondary analysis | Study and area of research were included as random effects to account for non-independence between studies examining similar literatures and where articles reported results for multiple samples (e.g., separate year ranges). The model without any random effects was also run and reported as sensitivity analyses (see supplementary material 3). |

**5.2.3 Record identification**

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). 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/> and some further descriptive analysis of this database can be found at that link or in supplementary material [full search]. 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”.

**5.1.2 Inclusion criteria**

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 [effect sizes]). See table [effect sizes] for a list of the effect size benchmarks used.

Table [effect sizes]. 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.

All applicable articles in the psychology literature were included (broadly defined, including general social, cognitive, occupational, management, clinical, developmental and social psychology, as well as psychiatry, educational and neuroscience research). Articles which analysed the power of fewer than six articles were excluded in order to exclude studies which targeted a small underpowered body of research. Only articles with full texts available in English were included. Dissertations and other grey were included if they otherwise met the inclusion criteria. The secondary analysis includes all identified articles which examine a body of research in the same population of psychology research and report the (a) the number of articles examined and (b) the number of articles or the proportion of examined articles which reported a power analysis. The sample size in the current article was determined entirely by the number of applicable articles in the literature.

*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. After screening of abstracts, 92 records remained and were subjected to full text screening. During full text screening, 46 articles were excluded. See figure [PRISMA] for a PRISMA flow diagram of article identification, screening, eligibility analysis and inclusion. A total of 46 articles were included in the primary analysis.

Fifteen articles also reported the proportion of sampled articles which had reported a power analysis and were included in the secondary meta-analysis and meta-regression examining the proportion of articles which report a power analysis. Two additional articles were found through reference list searches of these articles’ references lists, and were included in the secondary analysis.

## Screening

Records excluded from this study  
(n = 1416)

Full-text articles excluded  
Reasons: 3 records were of other areas of research, 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)

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 [Prisma].* 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* = 1508)

Record abstracts screened (*n* = 1508)

**5.2.4 Data extraction**

The articles included in the primary analysis were examined in randomized order. When additional power surveys were identified during data extraction, 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. The following pieces of information were being extracted from these power surveys: (a) The sampling strategy used, (b) targeted type of statistical tests (e.g., are all tests assessed or just t-tests?), (c) the journals included, (d) the years studied, (e) subfield of research targeted, (f) power estimation technique, (g) whether the power survey distinguishes between different tests, (h) the number of articles included in analysis, (i) the number of tests included in analysis, (j) the unit of analysis of the power survey (i.e., article or statistical test), (k) the effect size benchmarks used, (l) the mean, median, and SD of power at these benchmarks, and (m) the mean, median, and SD of sample sizes.

Data extraction for the second analysis used the same randomization procedure as for the primary data extraction. The sampling strategy employed in examined articles, the years included in the paper’s sampling strategy, the area of research the paper sampled from, the total number of articles examined, and the proportion of sampled articles which reported a power analysis were extracted. See <https://osf.io/7ncke/> for data, and Supplementary Material 2 for the codebook as well as a full list of datapoints extracted from articles for both the primary and secondary analyses.

**5.2.5 Missing data handling**

The range and interquartile ranges of power at small and medium benchmarks was extracted from plots in one article (Smith, Hardy, & Gammell, 2011) using R’s ‘locator’ function (Poisot, 2010), see <https://osf.io/7f2q9/> for the code and output of this analysis. 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 value provided in the text, and the value extracted in this manner are treated as if they were reported directly in all further analyses.

For nine of the 16 articles in which no IQRs or standard deviations were given, but which displayed cumulative frequency tables, variances were estimated 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/> [UPDATE THIS file!]. One paper only provided a frequency table at the small benchmark (Cashen & Geiger, 2004), included in the count above. This method was also used to impute 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 imputation method, 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 imputed means and the reported means was just .022.

When mean power or sample power standard deviations were not reported at benchmark levels, Wan, Wang, Liu, and Tong (2014)’s method (equation C3) was used to estimate the mean and variances from the reported median and quartiles using the Varameta package [citation]. In order to validate this approach, this method was also used to estimate the means for all articles which reported medians, quartiles as well as means (18 articles reporting 52 estimated means), giving a mean absolute error of .04. This method appears to work better at the medium value (mean absolute difference = .026), than at the small benchmark (mean absolute difference = .040), or the large benchmark (0.059), performing worse towards either bound because of Wan et al’s method assuming normally distributed underlying data, an assumption necessarily broken as the underlying variable is bounded between 0.5 and 1, and exacerbated as estimates approach either end of that range. This approach was also validated against the 17 articles reporting 49 variances and for which these values could be calculated, providing a mean absolute error of .028.

Two power surveys had medians and quartiles which were all the same at the large effect size benchmark (all .99) which would be estimated as zero using this method. These values, along with three remaining articles which did not report variances or enough information for any of the above methods to be used had their variance estimated as the mean variance of all other studies. Sensitivity analyses were performed using the median, minimum and maximum of the other studies’ variances for all performed meta-analyses (see below and supplementary material 3 for further detail).

**5.2.5 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 (Viechtbauer, 2010). Wan et al’s All data and code used in these analyses are available at <https://osf.io/as7md/>.

**5.2.5.1 Primary analysis:**

At each benchmark level of power (small, medium, and large) a multilevel random effects meta-regression was performed. Article and area of psychology research were included as random effects to account for non-independence of sub-studies within articles (e.g., when an article reported multiple power estimates for different year ranges), and when studies covered the same areas of research. The (mean-centralized) year each study examined power for as a fixed predictor were also performed at each benchmark effect size. 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. All analyses used restricted maximum likelihood estimation.

**5.2.5.2 Sensitivity analyses**

To investigate whether the results are sensitive to data imputation and estimation methods, analyses were also run 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), weighting by number of included articles instead of inverse variances, and without random effects for year or field of research. None of these changes altered the intercept parameter by more than .04, altered the effect of year by more than .002, or provided results which would lead to substantially different conclusions being drawn. See supplementary materials 4 for coefficient values produced under these different scenarios. 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 benchmarks 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) 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). This analysis showed little difference in parameter estimates, 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.

**5.2.5.3 Bias assessment**

In order to assess for publication bias, we used an analogue to Egger’s Test including the number of articles which were surveyed in each study as a moderator (Egger, Smith, Schneider, & Minder, 1997). Number of included articles was used instead of sampling variances as the sampling variances are expected to be associated with outcome scores, as estimated mean power levels towards either bound (1 or .05) are expected to have reduced sampling variances.

**5.2.5.4 Secondary analysis:**

A multilevel random 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. The mean year of the range of studies included in each paper was entered as a predictor in the meta-regression, after being mean-centralized for interpretability. Proportions were transformed using the Tukey-Freeman Arcsine Transform for both the meta-analysis and meta-regression, as this can act to normalize the sampling distributions of proportions (Miller, 1978). 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 study was included as a random effect in order to account for non-independence of individual estimates. Restricted maximum likelihood estimation was used.

**5.3 Results**

5.3.1.1 Sample characteristics

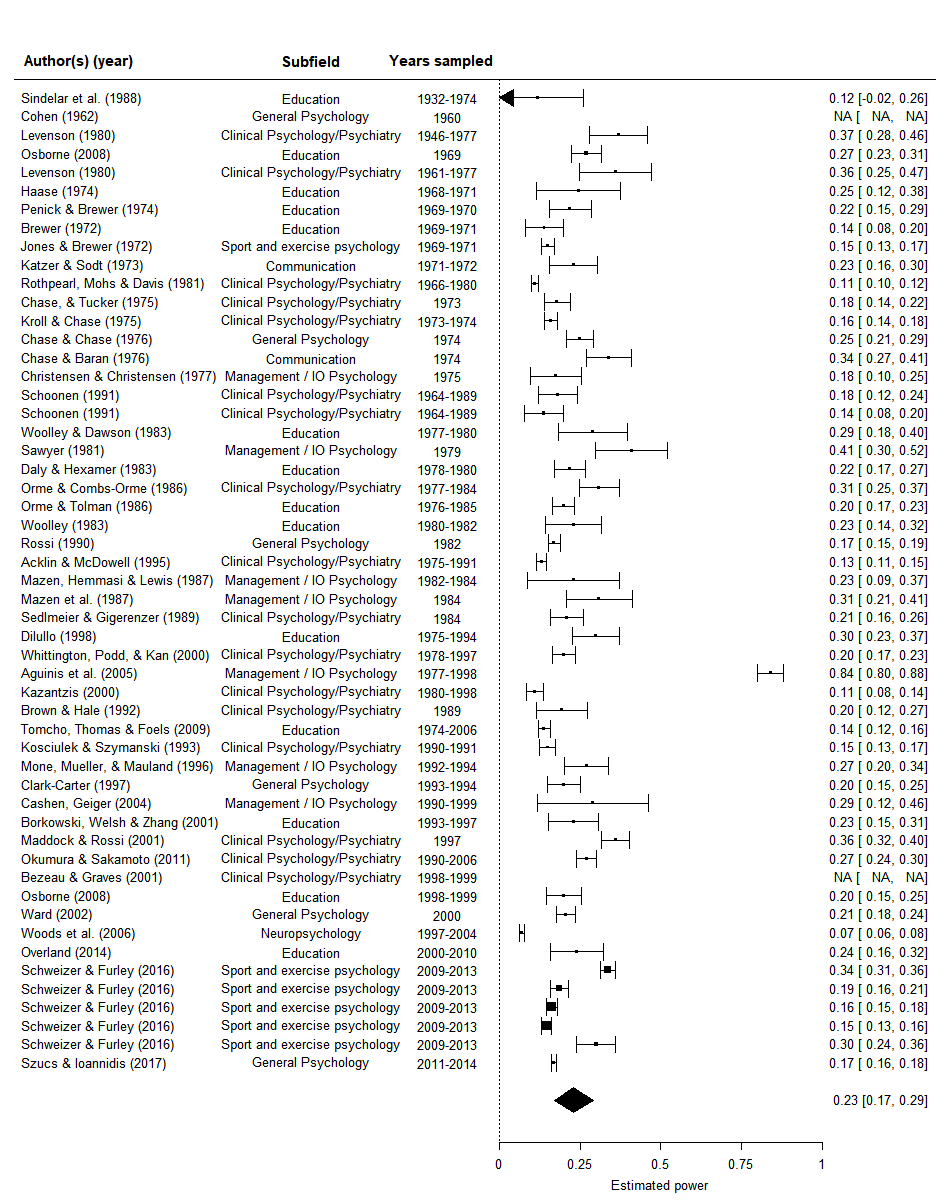
The

**5.3.1.2 results**

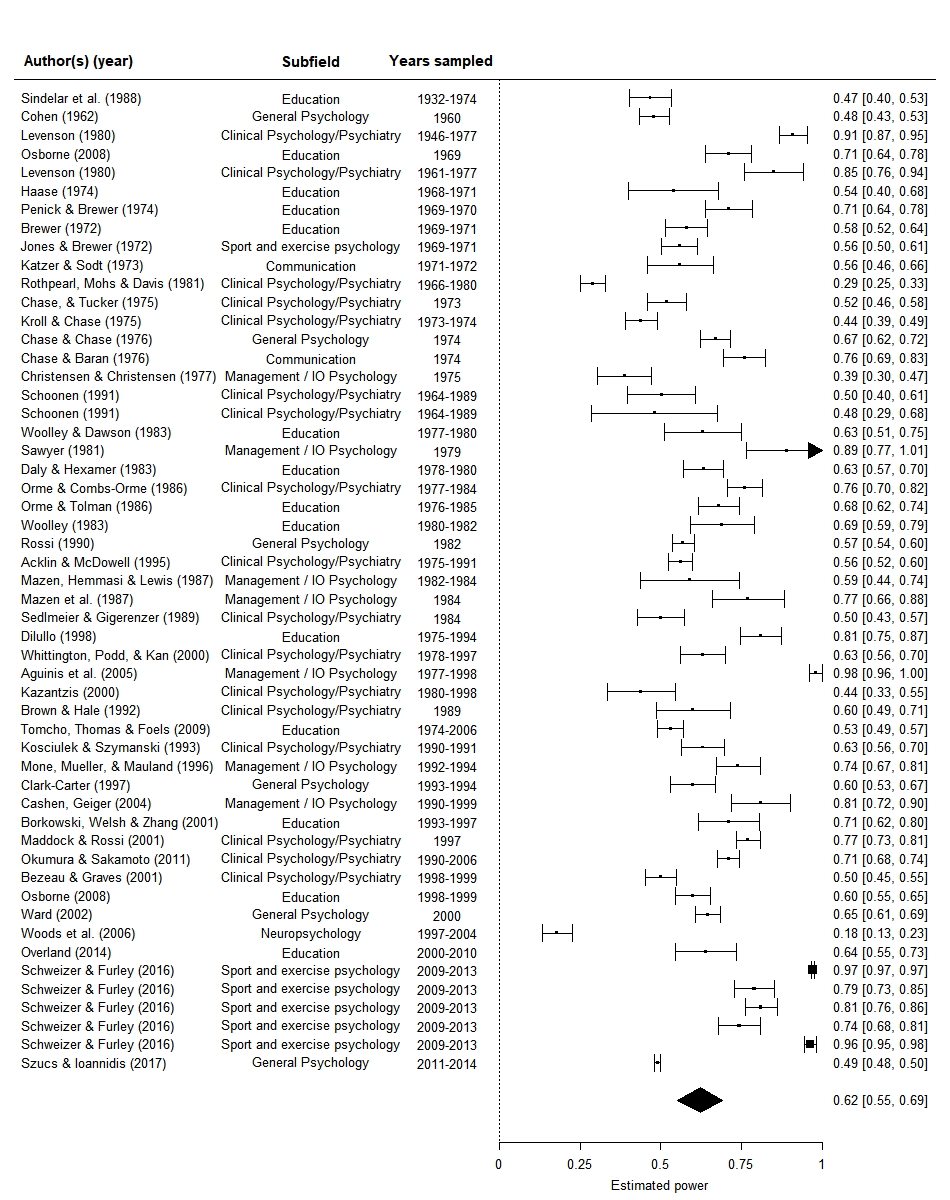
The multilevel meta-regression results suggest that the average power of psychology at the mean year included in this study was .23 (95% CIs [.17, .29]) for ‘small’ effects, .62 (95% CIs [.55, .69]) to detect ‘medium’ effects and .80 (95% CIs [.68, .92]) to detect ‘large’ effects following Cohen’s effect size benchmarks[[1]](#footnote-1). The estimated effect of time is negligible at all three benchmarks, at .001 (95% CIs [-0.003, 0.0006]), .002 (95% CIs [-0.004, 0.0007]) and .001 (95% CIs [-0.002, 0.0005]) at the small, medium and large benchmarks respectively. Random effects for article and subfield explain relatively small amounts of variance in all three models, and there is a significant amount of unexplained variance at all three benchmarks (see Table [Meta-regression primary small]-[Meta-regression primary large] for full model output, variance estimates and QE tests for excess heterogeneity). The degree of unexplained heterogeneity is unsurprising given the heterogenous populations included in this analysis, where researchers were estimating values as different as power for regression models in IO psychology compared to just t-tests in clinical psychology.

**5.3.1.1 Bias assessment**

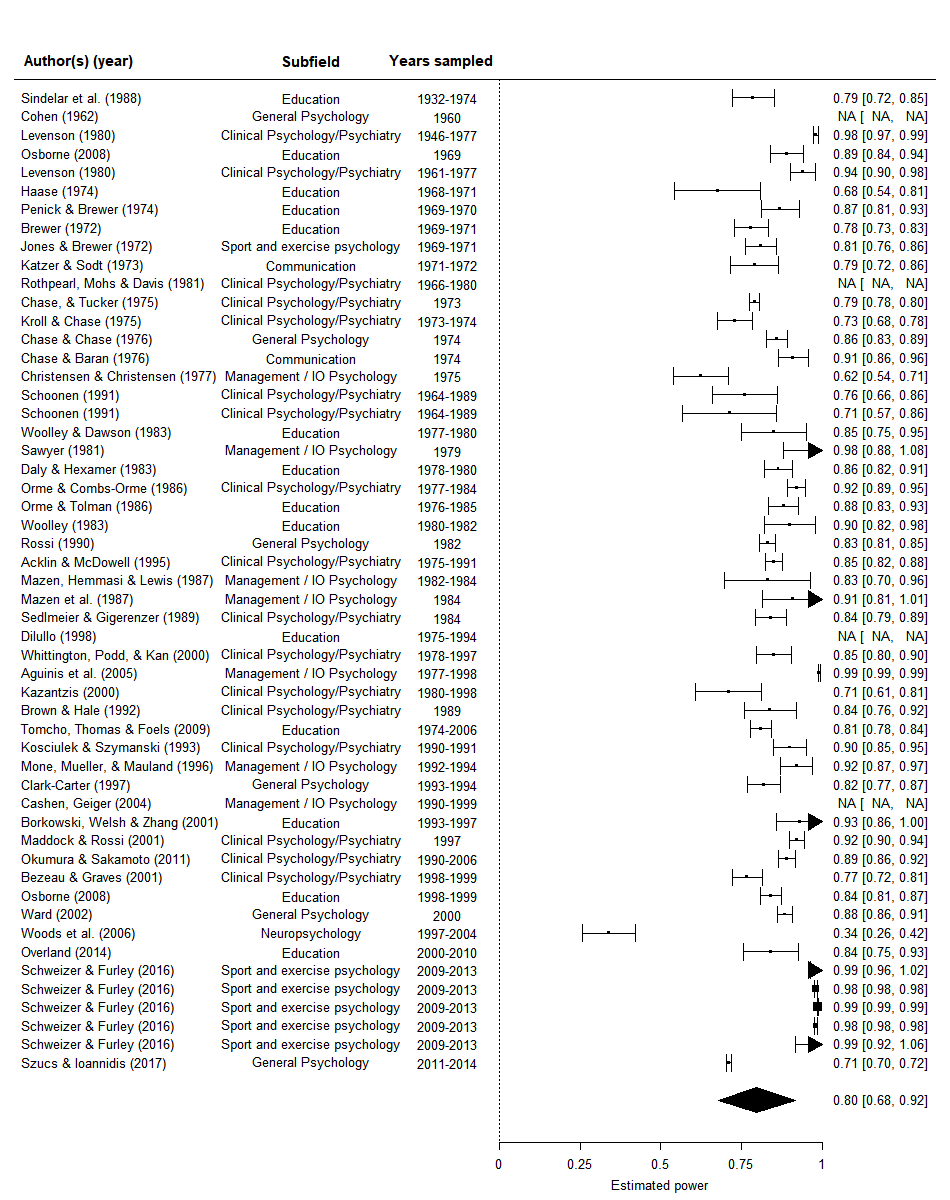
In order to assess for publication bias, we used an analogue to Egger’s Test; including the number of articles which were surveyed in each study as a moderator. This test showed that sample size was a significant positive predictor of average statistical power at the small and medium benchmark levels, but was not at the large effect size. Parameter estimates for the small medium and large effects respectively were (small b = 0.0005, p = <.001, medium b = .0001, p < .001, and large b = -0.000, *p* = .59), showing that larger studies tend to provide higher power estimates at the small and medium effect sizes. The fact that this is not seen at the large effect size is likely due to ceiling effects.



Plot [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.



Plot [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.



Plot [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.

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 | 0.231 | [0.172, 0.290] | < .001 |  |
| Year | -0.001 | [-0.003, 0.001] | .16 |  |
|  |  |  |  | Article σ2 = 0.0122, 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 | 0.622 | [0.553, 0.692] | < .001 |  |
| Year | -0.002 | [-0.004, 0.001] | .18 |  |
|  |  |  |  | Article σ2 = 0.0251, n = 46 |
|  |  |  |  | Subfield σ2 = 0.003, n = 7 |
|  |  |  |  | QE(51) = 10826.44, *p* < .001 |

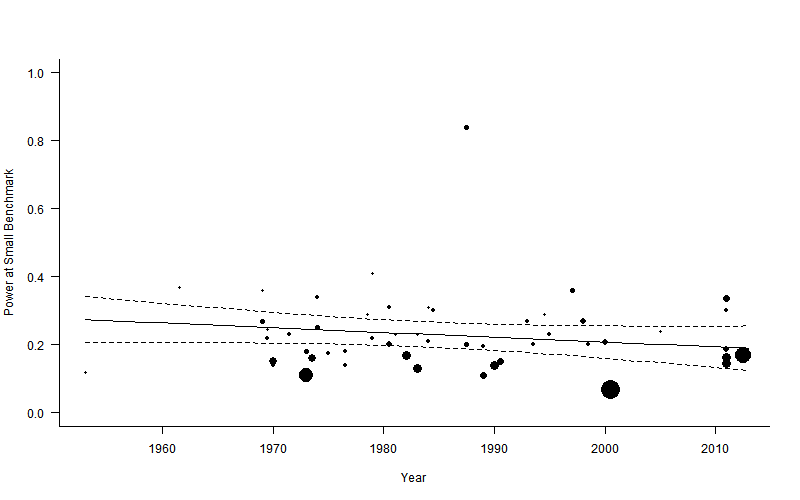
*Note.* *LL* and *UL* indicate the lower and upper limits of a confidence interval, respectively.

Table [Meta-regression primary large].

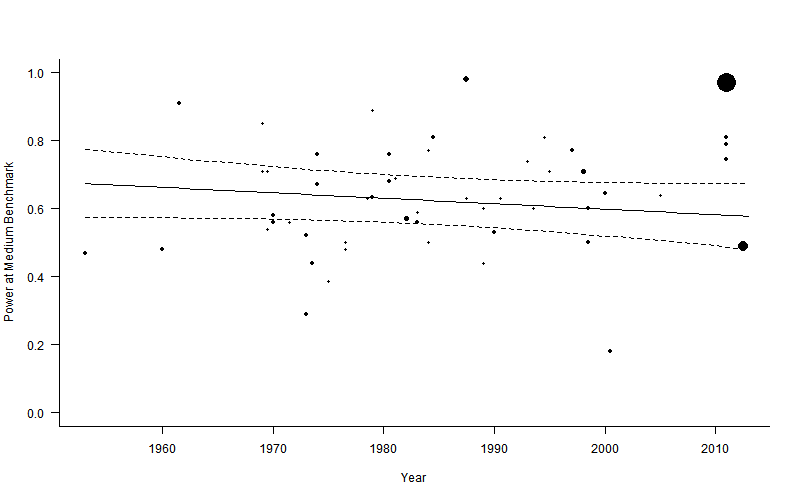
*Meta-regression of the power of psychology studies at a large effect size, including the year studied in each power survey as a moderator.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 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] | .21 |  |
|  |  |  |  | Article σ2 = 0.007, n = 42 |
|  |  |  |  | Subfield σ2 = 0.023, n = 7 |
|  |  |  |  | QE(48) = 6703.05, *p* < .001 |

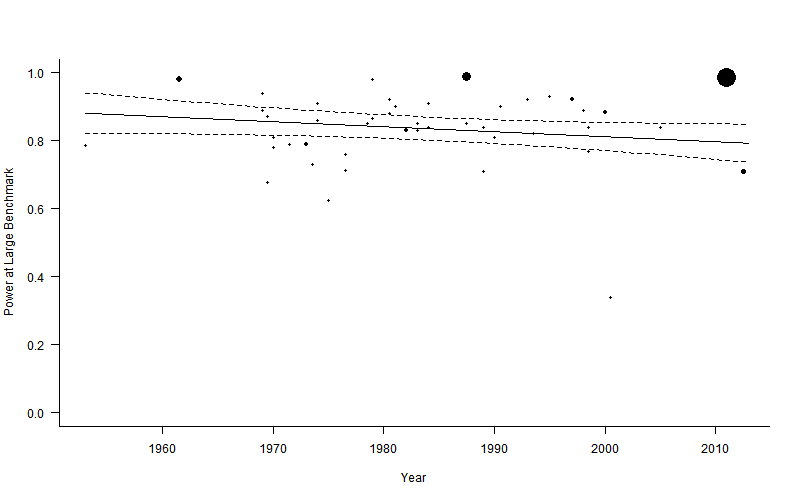
*Note.* *LL* and *UL* indicate the lower and upper limits of a confidence interval, respectively.



**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, circle 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, circle 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, circle sizes reflect the relative weighting of articles.

**5.3.2 Secondary analysis**

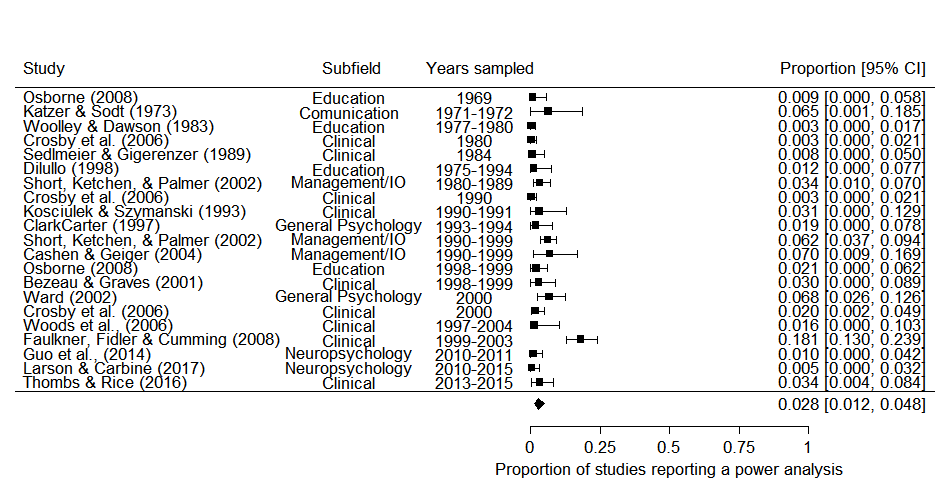
5.3.2.1 Sample characteristics

A plurality of data points (8 out of 21 data points included) were examinations of clinical research (e.g., examinations of 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% |

5.3.2.2 Secondary meta-regression results

The multilevel meta-regression of the proportion of studies which report a power analysis suggests that a very low proportion of psychology research reports a power analysis, with an the estimated proportion of researchers reporting a power analysis at the mean year included in this period being 2.8%, 95% CIs [1.2%, 4.8%]. There is a negligible change in increase in estimated power analysis reporting rates over time, with an estimated change in Freeman-Tukey double arcsine transformed units of 0.002 95% CI [-0.0007, 0.005] per year. See *Figure [Secondary meta-analysis with exclusions]* for a forest plot of the included studies, and Figure *[Secondary meta-regression with exclusions]* for a meta-regression scatterplot of the datapoints over time. The is significant unexplained heterogeneity in the proportion of studies reporting a power analysis remained, QE(19) = 85.86, *p* < .001**.** Random effects for article (σ2 = 0.01, n = 17) and subfield of research (Subfield σ2 = 0.00, n = 6) accounted for little to none of the overall variance in the model.



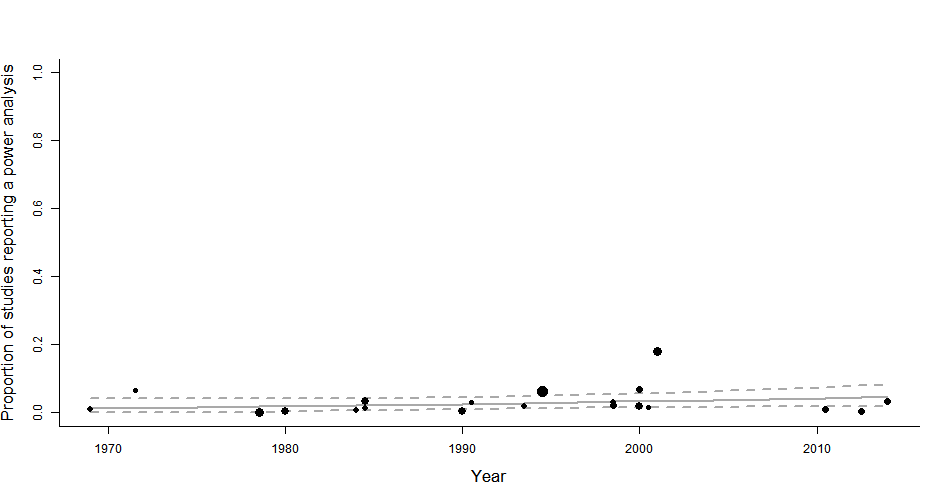
*Figure [Secondary meta-analysis with exclusions]*. Forest plot displaying the findings from seventeen examinations of the proportion of articles which reported a power analysis in psychology research.

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 | .186 | [0.136, 0.236] | < .001 |  |
| Year | 0.002 | [-0.0007, 0.005] | .13 |  |
|  |  |  |  | Article σ2 = 0.0077, n = 17 |
|  |  |  |  | Subfield σ2 = 0.000, n = 6 |
|  |  |  |  | QE(19) = 85.86, *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 PA, circle sizes reflect the relative weighting of each article.

**5.3 Discussion**

Assuming that effect sizes under study in psychology research have not increased, this analysis suggests that there has been little to no change in the statistical power of psychology research over the previous half century. Statistical power analysis reporting appears to be constantly low over time. 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., Amazon Turk studies) and larger undergraduate cohorts that should make larger scale research more tractable at least in many areas of psychological research.

HETROGENEITY IN FIELDS

Given that the average effect size seen in the psychology literature has been estimated to be around or below Cohen’s ‘medium’ effect size in psychology (e.g., Bosco, Aguinis, Singh, Field, & Pierce, 2015; Gignac & Szodorai, 2016; Quintana, 2017), this suggests that the average psychological research paper should fail to find significant results in almost 40% of occasions, assuming that the effect under study is in fact present. Despite this fact, over 90% of psychology papers report statistically significant findings (Fanelli, 2010). This means either a large proportion of performed research goes unreported (i.e., at least a third, again assuming that all studies are performed on true alternative hypotheses) or 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).

In order to prevent the performance of underpowered research, researchers should consider the likely power of their planned analyses during the planning of research. Given the evidence that our intuitions are very poor about the likely power and precision of research (Marjan Bakker, Hartgerink, Wicherts, & van der Maas, 2016; Obrecht, Chapman, & Gelman, 2007), formal power analysis, analyses to ensure a good probability of obtaining sufficiently narrow confidence intervals, or sufficiently convincing evidence via Bayes factors will be necessary for researchers to have accurate intuitions about the likely outcomes of their analyses under plausible alternative hypothesis. A variety of research planning packages and programs are freely available and should enable researchers to plan for relatively simple 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), although if more complex analyses are planned collaboration with a statistical consultant may be necessary (Van Meter & Charnigo, 2014).

Editors and reviewers 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 (APA Publications Communications Board Working Group on Journal Article Reporting Standards, 2008; Moher et al., 2010; Moher et al., 2001; Wilkinson, 1999). Although Power analyses are not easily interpretable after statistical analysis has been performed (Wagenmakers et al., 2015), requiring the truthful justification of sample sizes as a routine part of research forces researchers to think about these issues, and actively acknowledge that other constraints lead to sample size planning decisions when that is the case.

This advice, that researchers should consider the statistical power of their analyses during research planning and that reviewers should require their publication, are the suggested remedy in almost all of the papers included in this review. Given that these suggestions have clearly failed to have any impact on the practices of working scientists, I am not optimistic that anything will change in the near future. For research consumers this means that we must accept that a large portion of the research literature almost certainly provides exaggerated effect size estimates, does not have a constrained false positive error rate, and should be understood as inherently exploratory. This means that we should avoid focusing on the statistical significance of test results when interpreting results, instead focusing on the plausibility of the hypotheses and the strength of the evidence. Bayesian statistical approaches support these ends (Wagenmakers et al., 2015), although establishing minimum benchmarks for standards of evidence seems likely to lead to similarly problematic reporting habits in a different statistical approach.

Despite the bleak outlook for the field overall, individual researchers who want to do good research that will be publishable (i.e., the great majority of us) have a strong incentive to plan our research with a mind to the statistical power, precision or probability of finding strong evidence (Maxwell, Kelley, & Rausch, 2007; Schönbrodt & Wagenmakers, 2017). In cases where it is not possible to control the sample size that is available for analysis (e.g., like the current study), preregistering a data-analysis strategy and carefully assessing whether our conclusions rely on our data-dependent decisions becomes even more important. There are strong incentives for us to, consciously or otherwise, *p*-hack or HARK our way to more easily publishable findings (Fanelli, 2009; Fraser, Parker, Nakagawa, Barnett, & Fidler, 2018; John, Loewenstein, & Prelec, 2012; Kerr, 1998), and preregistration combined with careful sensitivity analyses provide tools to mitigate the possibility that we are using statistical tools to confirm our a priori hypotheses as opposed to enabling rigorous tests of our theories.

**Limitations:**

It would be unreasonable to assume that the included articles provide a random assessment of articles from the psychological research literature. It is possible that power surveys are only performed when a researcher has a suspicion that a particular area of research is underpowered, creating the impression that the overall literature is underpowered where as only the examined sections are. However, many of the included samples either target convivence samples (e.g., Szucs & Ioannidis, 2017), are explicitly chosen to be broadly representative of a subfield (e.g., Orme & Combs-Orme, 1986), or choose a population of high-impact journals in a subfield (e.g., Cashen & Geiger, 2004; Rossi, 1990), which should mitigate this probability.

Secondly, this study does not directly examine the statistical power of research, rather the statistical power of studies to detect Cohen’s benchmarks. Statistical power may have in fact increased if the average effect sizes that people are studying also increased, and this study would have no way of assessing this issue.Along similar lines, these studies almost uniformly target tests for which power can be easily estimated, ignoring more sophisticated analyses (e.g., complex SEM, factor analysis, or even multilevel models). This may mean that the included research underestimates the average power of psychological research, if larger studies tend to use these more sophisticated techniques.

The Egger’s test analogue results suggest that there is an association between estimated power and sample size. This could plausibly be accounted for by an association between subfield norms and the number of studies (i.e., fields were norms are to have larger sample sizes are also reflected in their meta-research), could be otherwise artefactual, or could suggest that publication bias is operating to suppress smaller studies which find higher power estimates. If the latter is the case, the estimates of mean power here could underestimate the true power of psychology research.

Additionally, a large amount of the heterogeneity seen in the power estimates examined may be caused by deviations from Cohen’s approach to power survey (e.g., not averaging the power of each article, only estimating power for the articles “primary analysis”, or only targeting a particular type of analysis). However, the overall trend in power estimates from these surveys is clear and many of these deviations seem likely to cause only small differences in observed power estimates (e.g., Cashen and Geiger (2004) showed that not averaging power estimates within articles lead to a difference in estimated power of less than .003).

There are a number of possible explanations for the estimated rate of power analysis reporting and lack of change over time that cannot be ruled out on the basis of the secondary analysis. First, it should be noted a plurality of these the studies included in this secondary analysis are from clinical psychology research, which could inflate the number of articles reporting a power analysis due to institutional review boards more regularly requiring power analysis if research deals clinical populations (Chan, Hróbjartsson, Jørgensen, Gøtzsche, & Altman, 2008; Moher et al., 2010). More problematically as none of the included articles were primarily being performed to estimate the proportion of articles which reported a power analysis, there is arguably an increased risk that only those articles in which the proportion of articles reporting a power analysis were worryingly low reported this value. Nonetheless, the results are so consistently low that this analysis provides strong evidence that power analyses are rarely reported across the psychology literature, even as the point estimate of the proportion should be interpreted with caution.

**Conclusion**

Statistical power analysis appears to be rarely reported, power to detect small to medium effects has been lower than suggested benchmarks, and neither of these facts appears to have changed despite over 50 years of repeated criticism on this topic. Research consumers should be aware that the proportion of studies which report statistically significant findings in psychology is implausibly high given the estimated power of studies in psychology, and should interpret published psychological literature with this fact in mind. Individual researchers should be aware of and make use of the tools that are available to help ensure that their research is likely to enable meaningful inferences.

**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

Chan, A.-W., Hróbjartsson, A., Jørgensen, K. J., Gøtzsche, P. C., & Altman, D. G. (2008). Discrepancies in sample size calculations and data analyses reported in randomised trials: comparison of publications with protocols. *BMJ, 337*. doi:10.1136/bmj.a2299

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. (2009). How Many Scientists Fabricate and Falsify Research? A Systematic Review and Meta-Analysis of Survey Data. *PLOS ONE, 4*, e5738. doi:10.1371/journal.pone.0005738

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

Fraser, H., Parker, T., Nakagawa, S., Barnett, A., & Fidler, F. (2018). Questionable research practices in ecology and evolution. *PLOS ONE, 13*, e0200303. doi:10.1371/journal.pone.0200303

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>

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

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science, 23*, 524-532. doi:10.1177/0956797611430953

Kerr, N. L. (1998). HARKing: Hypothesizing After the Results are Known. *Personality and Social Psychology Review, 2*, 196-217. doi:10.1207/s15327957pspr0203\_4

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., Kelley, K., & Rausch, J. R. (2007). Sample Size Planning for Statistical Power and Accuracy in Parameter Estimation. *Annual Review of Psychology, 59*, 537-563. doi:10.1146/annurev.psych.59.103006.093735

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

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

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

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

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

Wagenmakers, E.-J., Verhagen, J., Ly, A., Bakker, M., Lee, M. D., Matzke, D., . . . Morey, R. D. (2015). A power fallacy. *Behavior Research Methods, 47*, 913-917. doi:10.3758/s13428-014-0517-4

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.

**Supplemental material 1.**

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 2.**

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 |
| Jounral | 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 3.**

**Sensitivity analyses**

**Primary analysis**

To investigate whether the results are sensitive to data imputation choices, analyses were also run excluding including 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). All coefficient values were within .04 of the main results presented below, and no substantive differences in interpretation would result from this choice. Additionally, when no variances were reported or estimable, the main results reported below imputed the variance as the mean variance of all other studies. Sensitivity analyses were run imputing variances using the median, maximum and minimum variances instead of the mean showed that these lead to differences in any coefficient value of less than .001.

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, sensitivity analysis 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 less than .04; by -.001, -.002, and -.04 at the small, medium and large effect size benchmarks respectively.

See tables [Sensitivity analysis]- for full coefficient values from each model output, alongside output plots for models without random effects, or without random effects for area of research.

Table S4 - 1.

*Model output for various imputation and weighting methods, model including year as a moderator, random effects for study.*

|  |  |  |  |
| --- | --- | --- | --- |
| Analysis type | Mean Power | | |
| Small | Medium | Large |
| Mean variance imputation | 0.231 | 0.623 | 0.835 |
| Mean variance imputation, weighting by *n* articles | 0.21 | 0.593 | 0.798 |
| No values estimated or imputed | 0.226 | 0.593 | 0.825 |
| Maximum variance imputation | 0.23 | 0.621 | 0.834 |
| Minimum variance imputation | 0.231 | 0.624 | 0.836 |
| Median variance imputation | 0.231 | 0.623 | 0.836 |

Table S4 - 2.

*Model output for various imputation methods, model including year as a moderator, random effects for study.*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 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 |

Table S4 - 3.

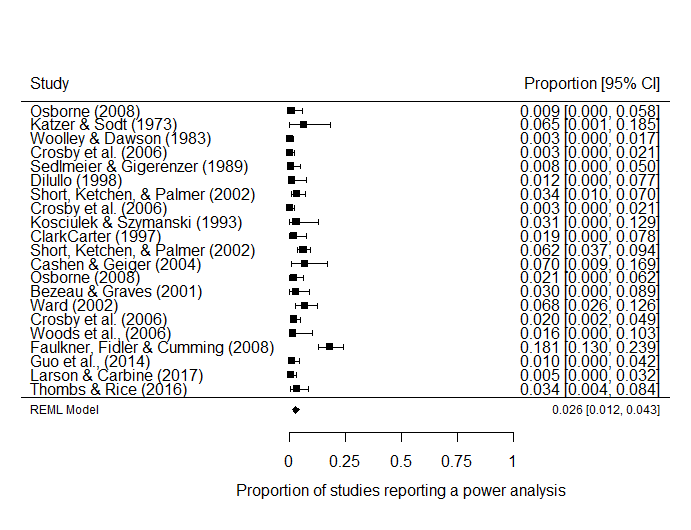
*Model output for various imputation methods, model including year as a moderator, random effects for area of research and study.*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Analysis type | Mean Power | | | Estimated change in average power per year | | |
| Small | Medium | Large | Small | Medium | Large |
| Mean imputationa | 0.231 | 0.622 | 0.797 | -0.001 | -0.002 | -0.001 |
| 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 | 0.231 | 0.621 | 0.796 | -0.001 | -0.002 | -0.001 |
| Minimum variance imputation | 0.232 | 0.623 | 0.798 | -0.001 | -0.002 | -0.001 |
| Median variance imputation | 0.231 | 0.623 | 0.797 | -0.001 | -0.002 | -0.001 |

Note: aModel is the model reported in the main text of this article.

**Secondary analysis**

See figure [Secondary meta-analysis] for a forest plot and model output from a REML 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.



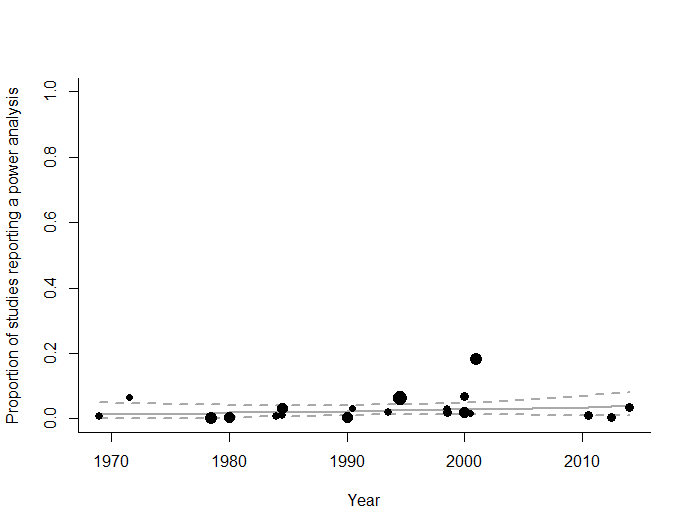
*Figure [Secondary meta-analysis]*. Forest plot displaying the findings from seventeen examinations of the proportion of articles which reported a power analysis in psychology research.

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 |

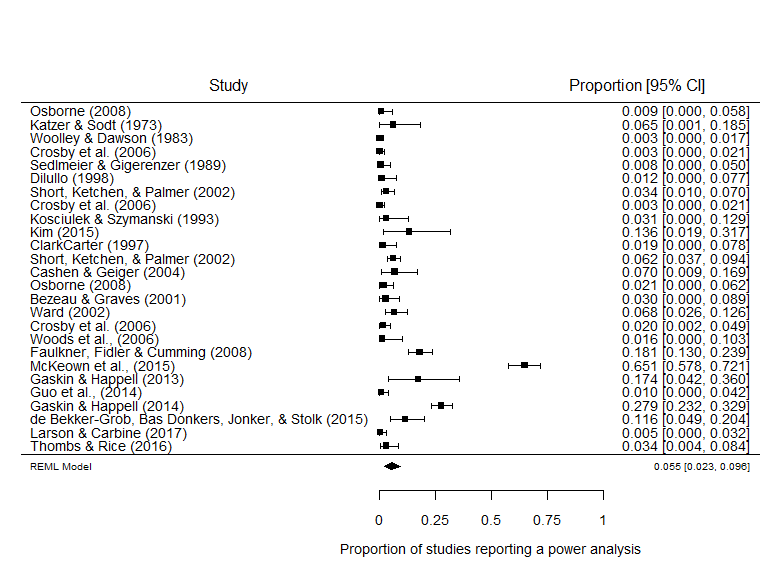
*Note.* *LL* and *UL* indicate the lower and upper limits of a confidence interval, respectively.



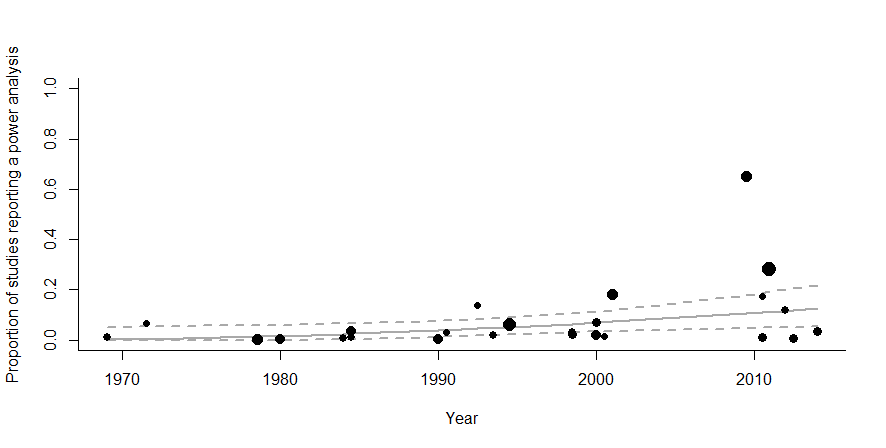
*Figure [Secondary meta-regression with exclusions]*. Meta-regression scatterplot of the effect of time (median year included in each study’s sample) on the proportion of sampled psychology articles reporting a power analysis. Dotted lines are 95% confidence intervals, and the solid line is the estimated 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 include some 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 average proportion of studies which report a power analysis is slightly higher .055, 95% Cis [.023, .096], although the population to which this average applies is unclear. As can be seen in figure *[Secondary meta-regression with medical studies],* a small but statistically significant effect of time can be seen [beta = 0.0065, se = 0.0028, z = 2.2827, *p* = .0224, 95% CI [0.0009, 0.0120]. This would lend support for a pre-registered hypothesis: “The number of articles reporting a power analysis will have increased over time, but is still low (i.e., below 10%).” However, this small 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 erroneously included papers happen to be recent, they cause an apparent increase in the proportion of studies reporting a power analysis over time.

**Supplementary Material 4 – Brief literature Overview**

On the 11th of September 2017 a literature search was performed using the PsycInfo and Medline databases in part to collect data for a meta-analysis of the power of psychology research (<https://osf.io/h8u9w/>), but also to allow for a brief overview of the amount and type of articles discussing statistical power in the psychology literature. This supplementary material briefly outlines the procedures used to classify the returned studies, and shows the total number of power surveys, papers giving power sample size determination advice or statistical procedures to estimate statistical power, articles which are conceptual discussions of statistical power, and the number of articles which have assessed the statistical power of statistical procedures under various conditions. Outlining even at this broad level the type and breadth of research that is available discussing statistical power, even just within domain specific fields of psychology, is important in illustrating just how much effort and research has gone into understanding, developing tools to estimate, and discussing statistical power within the field of psychology.

**Method**

**Literature search**

Psych info was searched using the medline interface for the search string *("\*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.* This search returned 916 documents. Web of Science was searched using the search string *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")*, returning 1,072 documents A total of 1988 documents were returned, leaving 1497 after 462 duplicates were removed.

**Article classification**

All 1487 remaining articles’ abstracts were screened and categorized. Full texts were accessed and assessed for applicability when there was ambiguity. All articles which primarily presented methods to plan the sample size of studies were selected.

This search excludes sample size planning articles for novel research paradigms (e.g., power analyses for large research projects). Conference proceedings were excluded. Software announcements were excluded, although articles about software included if they gave any examples of how the program could be used to estimate statistical power. Reanalyses of single previous studies or papers were excluded even if the focus of the reanalysis was on the paper’s likely statistical power.

Articles were classified into 6 groups, (1) power surveys, articles which estimate the statistical power of a group of studies, (2) sample size determination advice or techniques, (3) conceptual discussions of statistical power, (4) analyses of the statistical power of statistical procedures, (5) Analysis of the use of power analysis in a body of literature (i.e., studies which examine how power analysis has been reported and used in a body of literature), and (6) advice or techniques for effect size estimation for power analysis (e.g., articles detailing how to account for publication bias in estimating effect sizes from the published literature).

**Results**

This literature search identified 74 articles examining the statistical power of bodies of literature in psychology, 285 articles examining the power of different statistical procedures and some 478 articles providing methods of determining power or advice for sample size selection, and over 116 published conceptual discussions of statistical power. This database is available from <https://osf.io/t6jf8/>. A meta-analysis of those articles which attempt to estimate the statistical power of a body of research (power surveys), and analyses of the proportion of studies which report a power analysis can be found at (<https://osf.io/h8u9w/>). Although this is by no means a comprehensive overview of psychology research, it does show that there is a large body of research that has gone into discussing, assessing, and attempting to address psychology’s power problem, and provides a database of articles that may be useful to other researchers as a starting point for their own research.

Table SMLO 1.

*Classification of the types of article retrieved. Note that papers can be in multiple groups.*

|  |  |
| --- | --- |
| Type of article | Number of records |
| Articles discussing statistical power | 882 |
| Power surveys | 65 |
| Sample size determination advice or techniques | 432 |
| Conceptual discussions of statistical power | 102 |
| Analyses of the statistical power of statistical procedures | 272 |
| Analysis of the use of power analysis in a body of literature | 14 |
| Advice or techniques for effect size estimation for power analysis | 14 |
| Excluded articles | 607 |

Box SMLO 1. Example papers

|  |
| --- |
| **Power surveys**  Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology, 65*(3), 145-153.  ClarkCarter, D. (1997). The account taken of statistical power in research published in the British Journal of Psychology. *British Journal of Psychology, 88*, 71-83.  **Sample size determination advice or techniques**  Kelley, K. (2007). Sample size planning for the coefficient of variation from the accuracy in parameter estimation approach. *Behavior Research Methods, 39*(4), 755-766. doi:10.3758/bf03192966  Okumura, T. (2008). Sample size determination for data analysis by hierarchical linear models: An application to a social psychological research using data of TIMSS 2003. *Japanese Journal of Behaviormetrics, 35*(2), 221-228.  **Conceptual discussions of statistical power**  Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafo, M. R. (2013). Confidence and precision increase with high statistical power. *Nature Reviews Neuroscience, 14*(8), 585-586..  Chase, L. J., & Tucker, R. K. (1976). Statistical power: Derivation, development, and data-analytic implications. *The Psychological Record, 26*(4), 473-486.  **Analyses of the statistical power of statistical procedures**  Engelhardt, W. (1979). Non-parametric testing of interaction: Statistical power analysis. *Psychologische Beitrage, 21*(3-4), 439-449.  Visscher, P. M., & Posthuma, D. (2010). Statistical Power to Detect Genetic Loci Affecting Environmental Sensitivity. *Behavior Genetics, 40*(5), 728-733. doi:10.1007/s10519-010-9362-0  **Analysis of the use of power analysis in a body of literature**  Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafo, M. R. (2013). Confidence and precision increase with high statistical power. *Nature Reviews Neuroscience, 14*(8), 585-586..  Guo, Q., Thabane, L., Hall, G., McKinnon, M., Goeree, R., & Pullenayegum, E. (2014). A systematic review of the reporting of sample size calculations and corresponding data components in observational functional magnetic resonance imaging studies. *NeuroImage, 86*, 172-181.  **Advice or techniques for effect size estimation for power analysis**  Perugini, M., Gallucci, M., & Costantini, G. (2014). Safeguard Power as a Protection Against Imprecise Power Estimates. *Perspectives on Psychological Science, 9*(3), 319-332. doi:10.1177/1745691614528519  McShane, B. B., & Bockenholt, U. (2014). You Cannot Step Into the Same River Twice: When Power Analyses Are Optimistic. *Perspectives on Psychological Science, 9*(6), 612-625. doi:10.1177/1745691614548513 |

1. These values are estimates of power at the mean year of the studies included in the estimate, 1985. [↑](#footnote-ref-1)