Chapter 5. The Statistical Power of Psychology Research: A Systematic Review and Meta-analysis

### Abstract

Over the last half century 46 power surveys have been published in psychology, together estimating the statistical power that over 8,000 individual studies have at Cohen’s (1988) small, medium and large effect size benchmarks. This paper uses mixed effects meta regression to estimate the average statistical power of psychology research at each of these benchmarks, and to show whether and how these values have changed over time. We found that the average statistical power of published psychology research was extremely low for “small” effects, .23 (95% CI [.18, .29]), somewhat low for “medium” effects, .62 (95% CI [.54, .69]), and only acceptably high for “large” effects, .84 (95% CI [.81, .87]), and that there appears to have been little to no change in these values from 1960 to 2014. A secondary analysis of surveys that assessed how often power analyses are reported in psychology research suggests that power analysis reporting rates have increased slightly over time but remain low. This result suggests that efforts to increase the average power of research have not been effective and that novel methods of avoiding the negative implications of low statistical power are required.

Keywords: Publication bias, effect size, QRPs, statistical power, metascience, metaresearch

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 study should fail to reach statistical significance more than half the time even when studying a true “medium” effect. Since the publication of Cohen’s 1962 article, over 40 power surveys have systematically assessed the statistical power of bodies of psychology research at Cohen’s effect size benchmarks. The current study brings those papers together to estimate the average statistical power of psychology research at these benchmarks and to show how this value has changed over time using a systematic review and meta-analysis.

If studies in a body of literature have a low average level of statistical power, several major negative outcomes follow. Firstly, studies with a low level of statistical power can waste participants’ and researchers’ time and resources if statistical significance is used as the defacto standard for showing the presence of a non-zero effect (Cohen, 1962, 1988). 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 rate among significant reported results (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 behind 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 the sample sizes seen in 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 desired level of statistical power (Anderson & Maxwell, 2017).

Almost 1000 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., 1992, 1988) to statistical power analysis computer programs (e.g., Faul, Erdfelder, Lang, & Buchner, 2007). However, it is unclear whether this body of literature, the relative ease of use of these tools, and changes in journal reporting guidelines (e.g., APA Publications Communications Board Working Group on Journal Article Reporting Standards, 2008) 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. If these messages were being heard, we would expect the average statistical power of the published literature to have increased over time.

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; Appelbaum et al., 2018; Schulz, Altman, & Moher, 2010; 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.

# Methods

## Research design

The research design, hypotheses, and a detailed analysis plan for the primary and secondary analyses 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 were estimated when missing (see the section ‘Missing data handling and imputation’ below for full details) | 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. |
| A variance stabilizing transformation was used | A variance stabilizing transformation from Brown (1982) to account for the fact that power is bounded between .05 and 1 |
| 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. |
| “Sport and exercise psychology” and “communication research” were included as fields of research | “Sport and exercise psychology” and “communication research” are distinct areas of research not listed as subfields in the preregistration |

## Record identification

See Figure 1 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 articles in psychological research (broadly defined, including educational, occupational, management, clinical, psychiatry, and neuroscience research). Power surveys were included if authors 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). Surveys that analysed the power of fewer than six articles were excluded in order to avoid including articles that were not broad surveys of the power of an area of psychology research, but which were instead 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)*

|  |  |  |  |
| --- | --- | --- | --- |
| Type of test | Small | Medium | Large |
| t test on means (d) | 0.20 | 0.50 | 0.50 |
| t test on correlations (r) | 0.10 | 0.30 | 0.30 |
| F test ANOVA (f) | 0.10 | 0.25 | 0.25 |
| F test for multiple correlation or regression (f2) | 0.02 | 0.15 | 0.15 |
| Chi-square test (w) | 0.10 | 0.30 | 0.30 |

*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.

## Abstract and full text screening

One thousand, four hundred and thirty two 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, 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.



Figure []. Prisma flow diagram of article identification, screening and selection.

## 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 data-points extracted from articles for both the primary and secondary analyses. See Supplementary Materials 1 for a list of all included studies, the subfield of research they examined, and the number of articles they examined.

### 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 11 of these cases, no means were reported 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.

For two power surveys (Haase, 1974; Woolley, 1983) mean power levels were not reported, but frequency tables showing the number of articles achieving different levels of power at each effect size benchmark were presented. For those articles we estimated the mean power as the weighted average of the mid-interval values.

being the frequency within a particular bin and being the mid-interval value (e.g., for the bin .1 - .19, the mid interval value would be .145). In order to validate this mean estimation method, the difference between the estimated means and the reported means was calculated for 22 frequency tables from 8 studies that reported mean power and frequency tables, giving a mean absolute error of just 0.02 (with a mean error of 0.01 and an error standard deviation of 0.02). An R script with the data extracted from the frequency tables and the working for these estimates can be found at <https://osf.io/tdj6b/>.

Missing means were estimated using reported medians and interquartile power estimates following Wan, Wang, Liu, and Tong (2014)’s method (equation C3), using functions from the R package varameta (Grey, 2019) for five cases at each of the small, medium and large effect size benchmarks. In order to validate this approach, the means for all articles which reported medians, quartiles, and means were calculated (17 articles reporting 48 estimated means), which led to a mean absolute error of 0.05 (mean error = 0.00, sd = 0.07).

## Analysis

All data-analysis was conducted using R 3.5.1 (R Core Team, 2018), and meta-analyses were performed using the metafor package (version 2.0.0; Viechtbauer, 2010). The R Markdown document including all code and the data required to reproduce this paper document are available from <https://osf.io/as7md/>.

At each effect size benchmark (small, medium, and large) a mixed effects meta-regression was performed.

This analysis predicts estimated power at each benchmark () with an overall intercept (), and a fixed effect for year (). Random effects were included for the estimate (), survey (), and area of psychology research (). Random effects were included to account for non-independence between studies sampled from the same fields of research (e.g., clinical psychology or IO psychology), for non-independence in cases where surveys reported multiple estimates (e.g., when a power survey reported multiple power estimates for different year ranges), and at the effect level, making this a random effects meta-analysis. 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. All analyses used restricted maximum likelihood estimation.

Because statistical power is bounded between .05 and 1, we used a variance stabilizing transformation analogous to the Fisher r-to-z transformation to convert the mean level of power at each benchmark effect size into an unbounded quantity () following Brown (1982)

where is the estimated mean statistical power at a given effect size benchmark. The standard error of each estimate was calculating using

where is again the estimated mean statistical power from each survey and is the number of studies examined in each power survey. All estimates are back-transformed to raw estimated statistical power unless otherwise stated.

# Results

The mixed effects meta-regression suggests the mean power of psychology across this time period is .23 (95% CI [.18, .29]) for ‘small’ effects, .62 (95% CI [.54, .69]) for ‘medium’ effects, and .84 (95% CI [.81, .87]) for ‘large’ effects following Cohen’s effect size benchmarks. At each of the effect size benchmarks, the random effects at the article level have a greater estimated variance than the effect or subfield level, however these values are not precisely estimated enough to make definitive statements about even their rank order at the population level (e.g., at the medium benchmark the estimated standard deviations of the random effects are = 0.289, 95% CI [0.000, 0.289], = 0.551, 95% CI [0.000, 0.551], = 0.121, 95% CI [0.000, 0.121]). See tables 3 - 5 for meta-regression results including fixed and random effect estimates and Figures 2 - 4 for Forest plots at each effect size benchmark.

The estimated change in statistical power over time is negligible at all three benchmarks. The estimated change per year in transformed units is just -0.001 (95% CI [-0.015, 0.013]) for “small” effects, -0.001 (95% CI [-0.014, 0.012]) for “medium” effects, and -0.005 (95% CI [-0.020, 0.010]) for ‘large’ effects. Although this change is non-linear when back transformed to statistical power, these values represent an extremely small estimated change across the range of years covered in this study. For example, we can look at the predicted change per decade in the mean level of statistical power from the overall intercept (a value which represents an estimate of the mean level of statistical power in psychology at the mean year included in the power surveys, 1985). This gives us an estimated change in average statistical power per decade of just -.002 (95% CI [-.027, .022]) at the “small” benchmark, -.002 (95% CI [-.033, .030]) at the “medium” benchmark, and -.006 (95% CI [-.025, .014]) at the ‘large’ benchmark. See Figures 5 - 7 for scatter plots of the power estimates at each benchmark plotted against time.

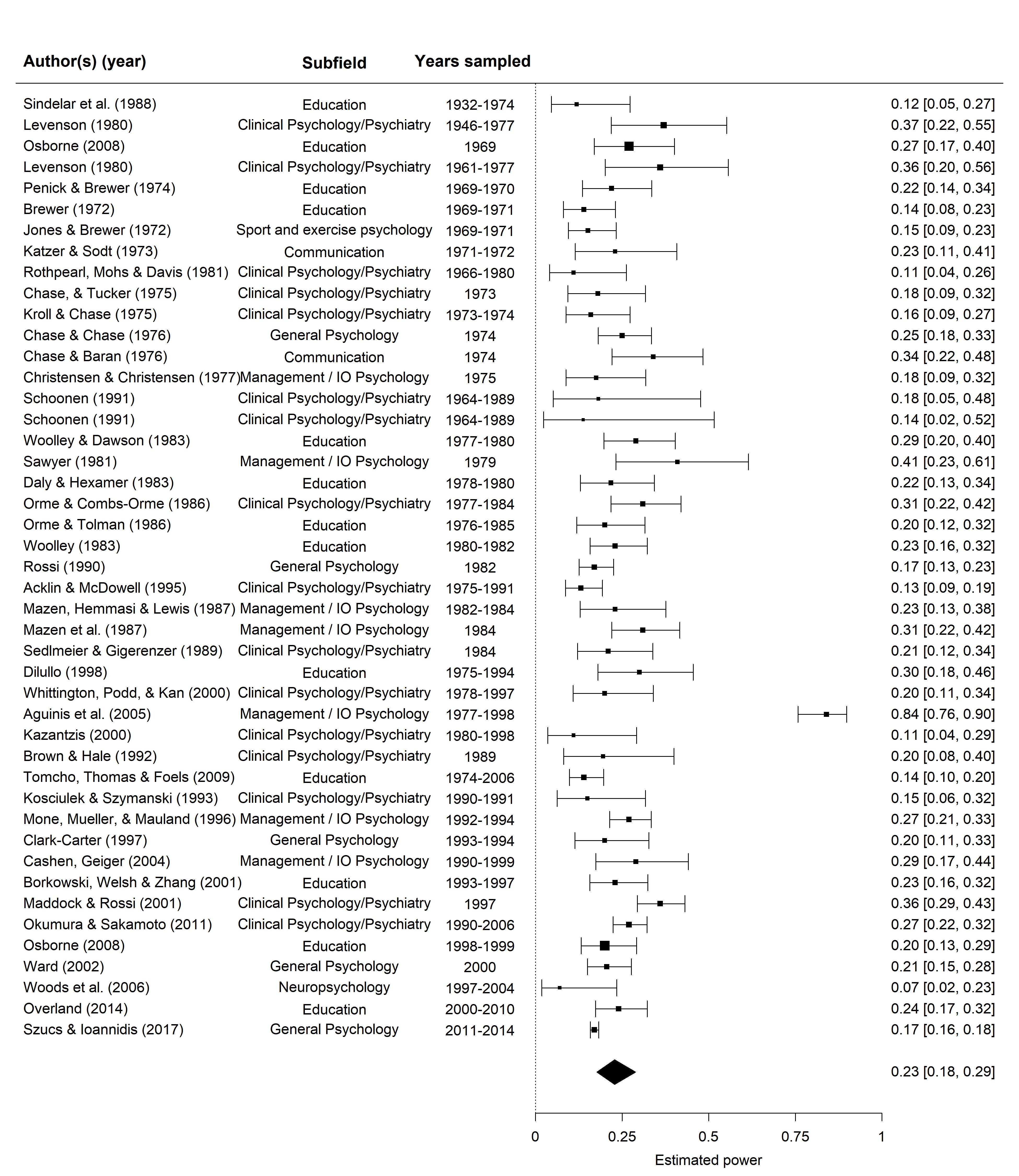


Figure 2. Forest plot of studies of the power of psychology research literatures at Cohen’s (1988) “small” effect size benchmark. The polygon shows the model intercept.

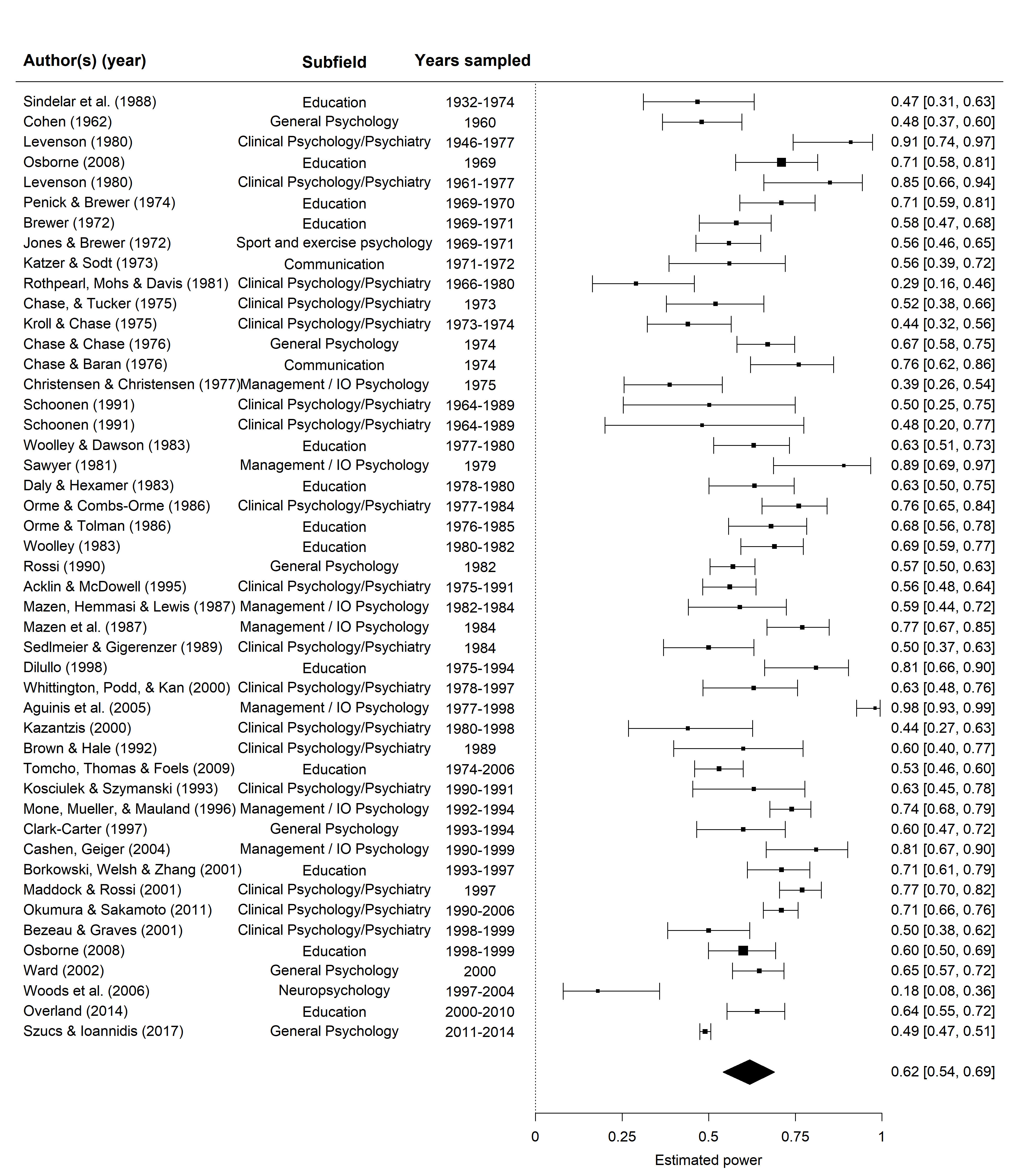


Figure 3. Forest plot of studies of the power of psychology research literatures at Cohen’s (1988) “medium” effect size benchmark. The polygon shows the model intercept.

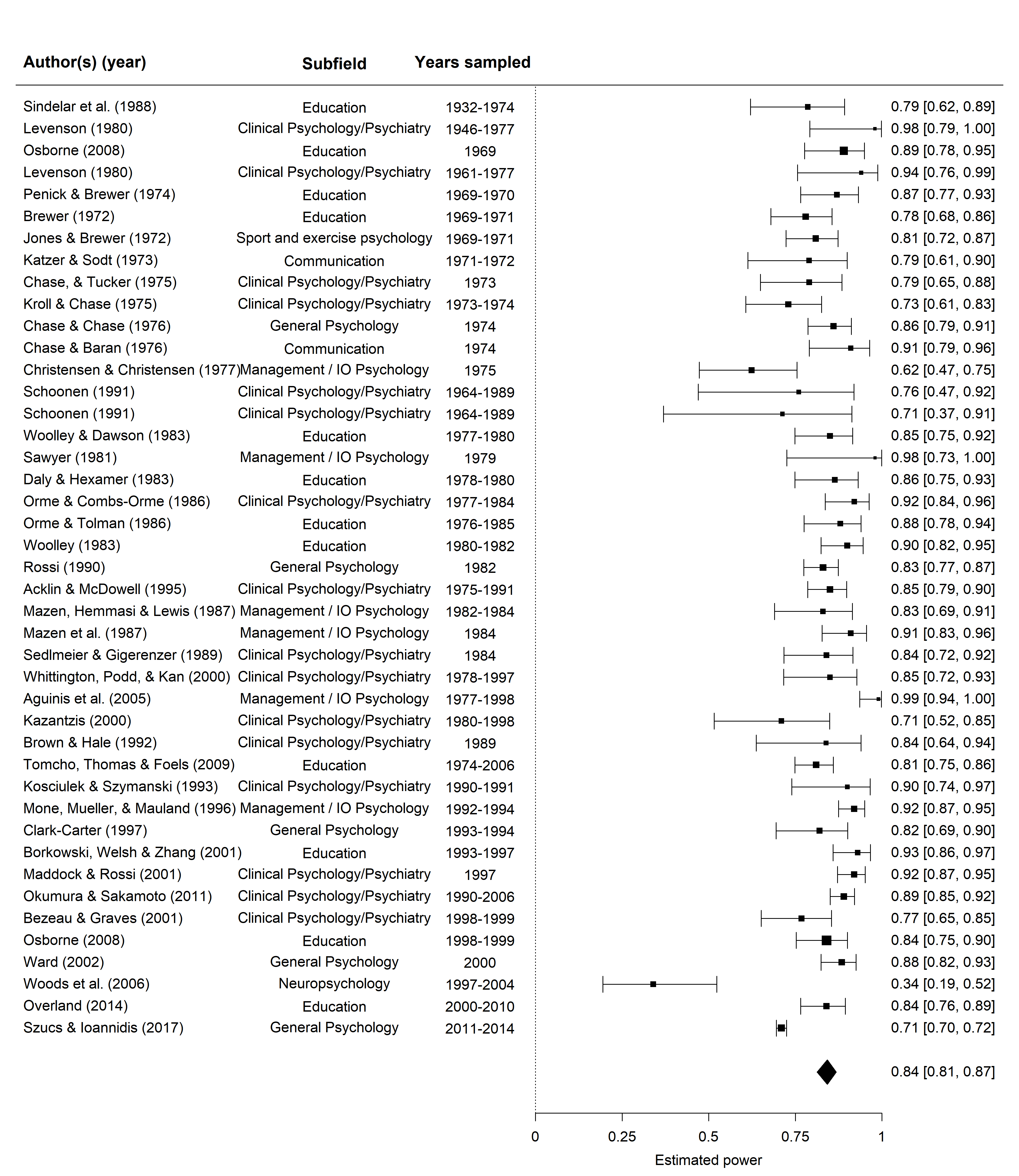


Figure 4. Forest plot of studies of the power of psychology research literatures at Cohen’s (1988) “large” effect size benchmark. The polygon shows the model intercept.

Table 3. *Results of a meta-regression of the power of psychology studies at Cohen’s (1988) “small” effect size benchmark, including the year studied in each power survey as a moderator.*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variable | Estimate | 95% CI LB | 95% CI UB | SE | p | Random effects |
| Intercept | -1.218 | -1.533 | -0.903 | 0.161 | < .001 |  |
| Year | -0.001 | -0.015 | 0.013 | 0.007 | .889 |  |
|  |  |  |  |  |  | Subfield variance = 0.085, n = 7 |
|  |  |  |  |  |  | Article variance = 0.263, n = 42 |
|  |  |  |  |  |  | Effect variance < .001, n = 45 |
|  |  |  |  |  |  | QE(44) = 232.7, p < .001 |

Table 4. *Results of a meta-regression of the power of psychology studies at Cohen’s (1988) “medium” effect size benchmark, including the year studied in each power survey as a moderator.*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variable | Estimate | 95% CI LB | 95% CI UB | SE | p | Random effects |
| Intercept | 0.484 | 0.167 | 0.800 | 0.161 | .003 |  |
| Year | -0.001 | -0.014 | 0.012 | 0.007 | .887 |  |
|  |  |  |  |  |  | Subfield variance = 0.084, n = 7 |
|  |  |  |  |  |  | Article variance = 0.303, n = 44 |
|  |  |  |  |  |  | Effect variance = 0.015, n = 47 |
|  |  |  |  |  |  | QE(46) = 279.65, p < .001 |

Table 5. *Results of a meta-regression of the power of psychology studies at Cohen’s (1988) “Large” effect size benchmark, including the year studied in each power survey as a moderator.*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variable | Estimate | 95% CI LB | 95% CI UB | SE | p | Random effects |
| Intercept | 1.676 | 1.469 | 1.883 | 0.106 | < .001 |  |
| Year | -0.005 | -0.020 | 0.010 | 0.008 | .530 |  |
|  |  |  |  |  |  | Subfield variance < .001, n = 7 |
|  |  |  |  |  |  | Article variance = 0.311, n = 40 |
|  |  |  |  |  |  | Effect variance < .001, n = 43 |
|  |  |  |  |  |  | QE(42) = 185.32, p < .001 |

### Bias assessment

The fact that the included surveys do not use traditional significance testing to assess their primary outcome means that it is not possible for a lack of statistically significant findings to act as a barrier to publication. However, it is still possible that smaller articles which find more “alarming” results are more likely to be published. If this were the case, this would lead to the current analysis producing pessimistic estimates of the average power of psychology. 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 surveyed in each study as a predictor in the meta-analyses. This analysis used the number of articles included in each study as opposed to the standard error of the transformed mean power estimates as the standard error is a function of the mean power estimates. This test suggests that there was little to no association between the number of articles included in each survey and statistical power at any of the benchmarks with coefficient estimates for sample size at the “small”, “medium”, and “large” benchmarks of -0.0001, 95% CI [-0.0004, 0.0002], *p* = .465, -0.0001 95% CI [-0.0004, 0.0002], *p* = .465, and -0.0002, 95% CI [-0.0005, 0.0001], *p* = .142, giving no obvious indication of publication bias.

### Sensitivity and robustness analysis

A number of sensitivity analyses were performed to assess whether data analysis and estimation choices influenced our inferences. We performed all analyses exuding the studies for which mean power values had to be estimated from medians and quartiles. We ran leave one out cross validation for all models (i.e., running each model dropping each power estimate included in the above analysis). We ran each model without random effects for article, without random effects for area of research, and without random effects for article or area of research. Finally, we performed these analyses treating the power estimates as simple means, i.e., not transforming the power estimates, and calculating their variances as their sample standard deviations divided by the square root of the number of articles included in each survey. None of these analyses would lead to any substantive difference in interpretation of results. The estimated mean power from these analyses does not change by more than .029 at the small benchmark, .035 at the medium benchmark, although it does decrease by .058 at the large benchmark when estimating mean power as a simple mean. Finally, the estimated change in statistical power per year did not change by more than .001 for any of these analyses (differences calculated using the back-transformed change per year from the model intercept for comparability between the simple means analysis which is in raw units of statistical power and the main, transformed analysis), and all models had comparable levels of precision to those presented. See Supplementary Materials 4 for output from all of these models.

# Discussion

This analysis suggests that the statistical power of psychology research was, on average, extremely low for “small” effects, .23 (95% CI [.18, .29]), somewhat low for “medium” effects, .62 (95% CI [.54, .69]), and only acceptably high for “large” effects, .84 (95% CI [.81, .87]) from the emergence of widespread statistical testing in psychology to 2014. It appears that there was little to no change in the average statistical power of psychology research to detect Cohen’s effect size benchmarks over this period. Looking at Supplementary Materials 5, the reporting of statistical power analysis appears to have become slightly more common over time, but to still only be reported in a small minority of studies. These results are extremely surprising 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., Cohen, 1962; Bezeau & Graves, 2001; Rossi, 1990), the increasing availability of user friendly power analysis tools (e.g., Cohen, 1992; Faul et al., 2007), as well as technological innovations (e.g., Amazon Turk studies) and larger undergraduate cohorts that could make larger sample research more tractable 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 below Cohen’s “medium” effect size (e.g., Bosco, Aguinis, Singh, Field, & Pierce, 2015; Gignac & Szodorai, 2016; Quintana, 2017), this suggests that an average psychological study should fail to find significant main results on 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). There are several possible explanations for this finding. Firstly, it is possible that a large proportion of performed research is simply going unreported (Greenwald, 1975; Sterling, 1959). Secondly, when read alongside evidence from surveys of researchers in psychology that find self reported rates of QRPs to be quite high (Agnoli, Wicherts, Veldkamp, Albiero, & Cubelli, 2017; John, Loewenstein, & Prelec, 2012; cf. Fiedler & Schwarz, 2016), it may suggests that some research is being presented as having found statistically significant main findings achieved in some part through p-hacking, HARKing or through the exploitation of researcher degrees of freedom (Bakker, van Assen, Crompvoets, Ong, & Soderberg, 2017; LeBel, McCarthy, Earp, Elson, & Vanpaemel, 2018; Sijtsma, Veldkamp, & Wicherts, 2016).

Given the evidence regarding how poor our intuitions are about the likely power and precision of research (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., G\*Power for the most common analyses such as ANOVA, regression or chi-square analysis; Faul et al., 2007; PINT 2.2 for two level hierarchical modeling; Snijders and Bosker, 1993; for sample size planning for Bayes factors see Schönbrodt and Wagenmakers, 2017; 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 following most of the existing reporting guidelines (e.g., APA Publications Communications Board Working Group on Journal Article Reporting Standards, 2008; Schulz et al., 2010; Wilkinson, 1999) and requesting a statement of justification for the included sample size included in a study. 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.

Fortunately, the recent development and rapid uptake of new research, publication and reporting initiatives give some reason for optimism, and provide mechanisms by which researchers can help to avoid the negative consequences of underpowered research. Preregistration of confirmatory analysis plans can help allow researchers to distinguish between confirmatory and exploratory analyses, helping to ensure that non-significant main findings are accurately reported when found (Simmons, Nelson, & Simonsohn, 2011). The use of preprint servers (e.g., https://psyarxiv.com) and data repositories (e.g., https://osf.io) allows researchers to disseminate findings outside of the traditional publication system, subverting publication bias. When “underpowered” or imprecise research occurs, ensuring that these results are available to future meta-analysts regardless of the outcome helps to avoid effect size inflation by ensuring that the results are available for future meta-analysts and can become part of a cumulative scientific literature. Finally, large-scale multi-lab collaborative efforts like the Psychological Science Accelerator (Moshontz et al., 2018) and the Many Labs projects (Klein et al., 2018) facilitate very large sample research, allowing for high powered research even when effect sizes may be small.

### Limitations

There are several limitations that should be noted in interpreting these results. Firstly, the individual articles included 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 published 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 estimates). Secondly, the power estimates included in this meta-analysis were calculated assuming that the studies that were surveyed used an alpha of .05. Because alpha corrections for multiple comparisons lead to lower statistical power, the current results may overestimate the average power of psychology research.

Finally, although this study suggests that the average level of statistical power of the psychology literature to detect Cohen’s benchmark effect sizes has not changed noticeably over time, this does not mean that the average statistical power of psychology research to detect the effect sizes under study has remained constant. If the effect sizes under study in psychological research have changed over time, the mean statistical power of this body will also have also changed over time. In order to address this question, the following chapter of this dissertation examines whether effect sizes have changed over time using a sample of database of over 130,000 effect size estimates from over 9,000 articles published in 5 APA journals from 1985 to 2013 (Nuijten, Hartgerink, van Assen, Epskamp, & Wicherts, 2015).

### Conclusion

Statistical power to detect small to medium effects appears to be substantially lower than recommended standards and power analysis to be rarely reported in psychology research. The average statistical power of the published psychology literature to detect these effect sizes does not appear to have increased substantially from the publication of the first power surveys to 2014, despite continued criticism of 'underpowered' research, and advocacy for the use of formal sample size planning techniques. 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.

As researchers, we all have limited budgets, limited numbers of potential participants, and limited amounts of time. We should ensure that we are making informed decisions about how we expend these resources. Formal sample size planning, performed for whatever goal we have for an experiment (be that reaching statistical significance given the presence of a particular effect, estimating a parameter with adequate precision, being able to reliably distinguish the correct model between a set of alternatives, or etc.), should play a role in helping us decide how we expend these limited resources. When we do perform research that may be "underpowered" or which is likely to lead to imprecise estimates, it's especially important that we make an effort to ensure that the results of our analyses will be available to other researchers regardless of their outcome.

## Supplementary Materials

### Supplementary Materials 1: Power surveys included in the primary analysis

Table SM1. Power surveys included in the primary analysis.

|  |  |  |
| --- | --- | --- |
| Articles | Subfield | Included articles |
| Sindelar et al. (1988). The power of hypothesis testing in special education efficacy research, The Journal of Special Education | Education | 35 |
| Cohen (1962). The statistical power of abnormal-social psychological research: A review, The Journal of Abnormal and Social Psychology | General Psychology | 70 |
| Osborne (2008). Sweating the small stuff in educational psychology: how effect size and power reporting failed to change from 1969 to 1999, and what that means for the future of changing practices, Educational Psychology | Education | 55 |
| Levenson (1980) . Statistical Power Analysis: Implications for Researchers, Planners, and Practitioners in Gerontology, The Gerontologist | Clinical Psychology/Psychiatry | 26 |
| Haase (1974). Power analysis of research in counselor education, Counselor Education and Supervision | Education | 60 |
| Penick & Brewer (1974). The power of statistical tests in science teaching research, Journal of Research in Science Teaching | Education | 66 |
| Brewer (1972). On the Power of Statistical Tests in the American Educational Research Journal, American Educational Research Journal | Education | 85 |
| Jones & Brewer (1972). An Analysis of the Power of Statistical Tests Reported in the Research Quarterly, Research Quarterly | Sport and exercise psychology | 106 |
| Katzer & Sodt (1973). An Analysis of the Use of Statistical Testing in Communication Research, Journal of Communication | Communication | 31 |
| Rothpearl, Mohs & Davis (1981). Statistical power in biological psychiatry, Psychiatry Research | Clinical Psychology/Psychiatry | 35 |
| Chase, & Tucker (1975). A power-analytic examination of contemporary communication research, Speech Monographs | Clinical Psychology/Psychiatry | 46 |
| Kroll & Chase (1975). Communication disorders: A power analytic assessment of recent research, Journal of Communication Disorders | Clinical Psychology/Psychiatry | 62 |
| Chase & Chase (1976). A statistical power analysis of applied psychological research, Journal of Applied Psychology | General Psychology | 121 |
| Chase & Baran (1976). An Assessment of Quantitative Research in Mass Communication., Journalism Quarterly | Communication | 48 |
| Christensen & Christensen (1977). Statistical Power Analysis of Health, Physical Education, and Recreation Research., Research Quarterly. American Alliance for Health, Physical Education and Recreation | Management / IO Psychology | 43 |
| Schoonen (1991). The internal validity of efficacy studies: Design and statistical power in studies of language therapy for aphasics, Brain and Language | Clinical Psychology/Psychiatry | 13 |
| Woolley & Dawson (1983). A Follow-up Power Analysis of the Statistical Tests Used in the Journal of Research in Science Teaching., Journal of Research in Science Teaching | Education | 73 |
| Sawyer (1981). Statistical power and effect size in marketing research, Journal of Marketing Research | Management / IO Psychology | 23 |
| Daly & Hexamer (1983). Statistical Power in Research in English Education., Research in the Teaching of English | Education | 57 |
| Orme & Combs-Orme (1986). Statistical power and Type II errors in social work research, Social Work Research & Abstracts | Clinical Psychology/Psychiatry | 79 |
| Orme & Tolman (1986). The Statistical Power of a Decade of Social Work Education Research., Social Service Review | Education | 64 |
| Woolley (1983). A comprehensive power-analytic investigation of research in medical education, Journal of Medical Education | Education | 100 |
| Rossi (1990). Statistical power of psychological research: What have we gained in 20 years?, Journal of Consulting and Clinical Psychology | General Psychology | 221 |
| Acklin & McDowell (1995). Statistical power in Rorschach research, Exner, John E Jr [Ed] (1995) Issues and methods in Rorschach research (pp 181-193) xiii, 324 pp Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc; US | Clinical Psychology/Psychiatry | 158 |
| Mazen, Hemmasi & Lewis (1987). Assessment of statistical power in contemporary strategy research, Strategic Management Journal | Management / IO Psychology | 44 |
| Mazen et al. (1987). Statistical power in contemporary management research, Academy of Management Journal | Management / IO Psychology | 84 |
| Sedlmeier & Gigerenzer (1989). Do studies of statistical power have an effect on the power of studies?, Psychological Bulletin | Clinical Psychology/Psychiatry | 54 |
| Dilullo (1998). A post hoc power analysis of inferential research examining the relationship between mathematics anxiety and mathematics performance, Dissertation Abstracts International Section A: Humanities and Social Sciences | Education | 41 |
| Whittington, Podd, & Kan (2000). Recognition memory impairment in Parkinson’s disease: Power and meta-analyses, Neuropsychology | Clinical Psychology/Psychiatry | 46 |
| Aguinis et al. (2005). Effect size and power in assessing moderating effects of categorical variables using multiple regression: A 30-year review, Journal of Applied Psychology | Management / IO Psychology | 106 |
| Kazantzis (2000). Power to detect homework effects in psychotherapy outcome research, Journal of Consulting and Clinical Psychology | Clinical Psychology/Psychiatry | 27 |
| Brown & Hale (1992). The power of statistical studies in consultation-liaison psychiatry, Psychosomatics | Clinical Psychology/Psychiatry | 24 |
| Tomcho, Thomas & Foels (2009). The power of teaching activities: Statistical and methodological recommendations, Teaching of Psychology | Education | 193 |
| Kosciulek & Szymanski (1993). Statistical power analysis of rehabilitation counseling research, Rehabilitation Counseling Bulletin | Clinical Psychology/Psychiatry | 32 |
| Mone, Mueller, & Mauland (1996). The perceptions and usage of statistical power in applied psychology and management research, Personnel Psychology | Management / IO Psychology | 210 |
| Clark-Carter (1997). The account taken of statistical power in research published in the British Journal of Psychology, British Journal of Psychology | General Psychology | 54 |
| Cashen, Geiger (2004). Statistical power and the testing of null hypotheses: A review of contemporary management research and recommendations for future studies, Organizational Research Methods | Management / IO Psychology | 43 |
| Borkowski, Welsh & Zhang (2001). An Analysis of Statistical Power in Behavioral Accounting Research., Behavioral Research in Accounting | Education | 96 |
| Maddock & Rossi (2001). Statistical power of articles published in three health psychology-related journals, Health Psychology | Clinical Psychology/Psychiatry | 187 |
| Okumura & Sakamoto (2011). Statistical power and effect sizes of depression research in Japan, Psychiatry and Clinical Neurosciences | Clinical Psychology/Psychiatry | 311 |
| Bezeau & Graves (2001). Statistical power and effect sizes of clinical neuropsychology research, Journal of Clinical and Experimental Neuropsychology | Clinical Psychology/Psychiatry | 66 |
| Ward (2002). Highly significant findings in psychology: A power and effect size survey, Dissertation Abstracts International: Section B: The Sciences and Engineering | General Psychology | 157 |
| Woods et al. (2006). Statistical power of studies examining the cognitive effects of subthalamic nucleus deep brain stimulation in Parkinson’s disease, Clinical Neuropsychologist | Neuropsychology | 30 |
| Overland (2014). Statistical Power in the Journal of Research in Music Education (2000-2010): A Retrospective Power Analysis, Bulletin of the Council for Research in Music Education | Education | 125 |
| Schweizer & Furley (2016). Reproducible research in sport and exercise psychology: The role of sample sizes, Psychology of Sport and Exercise | Sport and exercise psychology | 337 |
| Szucs & Ioannidis (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature., PLOS Biology | General Psychology | 3801 |

### Supplementary Materials 2: Search parameters

Table SM2. 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”) | 1072 |
| Total number of records |  | 1988 |
| De-duplicated library |  | 1489 |

## Supplementary Materials 3: Code book for data collection

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)

Table SM3. Codebook for data collection sheet.

|  |  |
| --- | --- |
| Column name | 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.) |
| DistinguishedStatisticalTestsAppropriately | 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 |
| 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 |
| Notes | Any notes? Record reason for exclusion here |
| 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 - article examines studies not in scope of literature review |

## Supplementary Materials 4: Sensitivity analyses and robustness checks

We performed a set of robustness and sensitivity analyses. First, we performed leave one out cross validation to assess whether any individual power estimate was overly influential. Secondly, we excluded all power estimates for which we had to estimate mean power using Wan and colleagues 2014 estimator. Thirdly, we performed all analyses treating each mean power estimate as an unbounded mean (i.e., not using a variance stabilizing transformation, instead treating it as a simple mean and estimating its variance as the standard deviation divided by the square root of the sample size included in this analysis). Finally, we tried alternative model specifications, removing random effects for the article, the article and the sub-field, and removing the fixed effect for year. None of these robustness or sensitivity analyses led to substantial differences in the results. See below for details on each of these analyses, and Tables SM4 to SM6 for model output from all of the sensitivity analyses.

### Leave one out cross validation

Leave one out cross validation was used to assess whether any individual power estimate has a large impact on the model coefficient values. The estimated mean level of power did not change by more than a power of 0.035 and the coefficient for year did not change by more than 0.001 at any of the effect size benchmarks when any individual estimate was dropped from these analyses.

### Excluding studies for which mean power had to be estimated

We re-performed each analysis excluding the five studies for which means had to be estimated from medians and quartiles using Wan et al. (2014)’s method. The results were not meaningfully different, with similar precision in the coefficient estimates (see table SM7 for estimates and confidence intervals), and the estimates of the mean statistical power and change per year changing by 0.002 or less.

### Using simple means

We also performed these analyses treating the power estimates as simple means, not transforming the power estimates, and calculating their variances as their sample standard deviations divided by the square root of the number of articles included in each survey. This analysis involved estimating the variances of a number of mean power estimates as standard deviations were not reported for 20 power estimates from 16 articles.

Several methods were used to estimate these missing variances. For 9 values at the small benchmark and for 8 at the medium and large benchmarks we estimated variances from frequency tables (i.e., tables showing how many articles had an estimated power at a given benchmark between .05 and .1, between .1 and .2, etc), using

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

We estimated variances using Wan et al. (2014)‘s C3 method (see equation 16 in Wan et al., 2014) for 5 mean power estimates at the small benchmark, 6 articles at the medium benchmark, and 4 at the large benchmark. For one of these articles the range and interquartile ranges of power at the small and medium benchmarks were extracted from plots using R’s ‘locator’ function. 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. This left a single article at the small effect size benchmark for which we could not estimate the variance of its power estimates (it did not report a power frequency table or quartiles), which was left out of this analysis.

No substantial differences in interpretation would result from the choice to treat the estimates as simple means, with the estimated mean power of psychology changing by 0.01 or less at the small and medium benchmarks, although estimated power is estimated to be notably lower at the large effect size benchmark with its estimate decreasing by 0.06, and the estimated change in statistical power by year being negligible and similarly precisely estimated.

### Alternative model parameterisations

Finally, we re-estimated the model under parameterisations, removing random effects for the article, the article and the sub-field, and removing the fixed effect for year. 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.

Table SM6. *Mean estimated power to detect a small effect and estimated change per year (in both transformed units and back-transformed to units of statistical power) for the primary analysis model presented in the article body (“Main model”), a model without a coefficient for year, without random effects for field or subfield, without random effects for article, and a model estimated using untransformed power estimates (i.e., treating effects as simple means).*

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Mean Power | | | Change per year in transformed units | | | Change per year in units of power | | |
| Model | Estimate | 95% CI LB | 95% CI UB | Estimate | 95% CI LB | 95% CI UB | Estimate | 95% CI LB | 95% CI UB |
| Main model | 0.228 | 0.178 | 0.288 | -0.001 | -0.015 | 0.013 | 0.000 | -0.003 | 0.002 |
| No coefficient for year | 0.202 | 0.181 | 0.225 |  |  |  |  |  |  |
| No random effects for subfield or article | 0.202 | 0.181 | 0.225 | 0.002 | -0.009 | 0.012 | 0.000 | -0.001 | 0.002 |
| No random effects for article | 0.202 | 0.181 | 0.225 | 0.002 | -0.009 | 0.012 | 0.000 | -0.001 | 0.002 |
| Untransformed power estimates | 0.233 | 0.174 | 0.293 |  |  |  | -0.002 | -0.004 | 0.000 |

*Note*. Estimates of the change per year in back-transformed units are calculated from the model intercept, and represent the estimated change per year at the mean year included in this study, 1985.

Table SM6. *Mean estimated power to detect a medium effect and estimated change per year (in both transformed units and back-transformed to units of statistical power) for the primary analysis model presented in the article body (“Main model”), a model without a coefficient for year, without random effects for field or subfield, without random effects for article, and a model estimated using untransformed power estimates (i.e., treating effects as simple means).*

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Mean Power | | Change per year in transformed units | | | | | Change per year in units of power | | |
| Model | Estimate | | 95% CI LB | 95% CI UB | Estimate | 95% CI LB | 95% CI UB | Estimate | 95% CI LB | 95% CI UB |
| Main model | 0.619 | | 0.542 | 0.690 | -0.001 | -0.014 | 0.012 | 0.000 | -0.003 | 0.003 |
| No coefficient for year | 0.608 | | 0.566 | 0.649 |  |  |  |  |  |  |
| No random effects for subfield or article | 0.610 | | 0.567 | 0.650 | -0.003 | -0.016 | 0.011 | -0.001 | -0.004 | 0.003 |
| No random effects for article | | 0.610 | 0.567 | 0.650 | -0.003 | -0.016 | 0.011 | -0.001 | -0.004 | 0.003 |
| Untransformed power estimates | | 0.606 | 0.529 | 0.684 |  |  |  | -0.002 | -0.005 | 0.001 |

Note. Estimates of the change per year in back-transformed units are calculated from the model intercept, and represent the estimated change per year at the mean year included in this study, 1985.

Table SM7. *Mean estimated power to detect a large effect and estimated change per year (in both transformed units and back-transformed to units of statistical power) for the primary analysis model presented in the article body (“Main model”), a model without a coefficient for year, without random effects for field or subfield, without random effects for article, and a model estimated using untransformed power estimates (i.e., treating effects as simple means).*

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Mean Power | | | Change per year in transformed units | | | Change per year in units of power | | |
| Model | Estimate | 95% CI LB | 95% CI UB | Estimate | 95% CI LB | 95% CI UB | Estimate | 95% CI LB | 95% CI UB |
| Main model | 0.842 | 0.813 | 0.868 | -0.005 | -0.020 | 0.010 | -0.001 | -0.003 | 0.001 |
| No coefficient for year | 0.842 | 0.813 | 0.868 |  |  |  |  |  |  |
| No random effects for subfield or article | 0.843 | 0.814 | 0.868 | -0.004 | -0.020 | 0.012 | 0.000 | -0.003 | 0.002 |
| No random effects for article | 0.823 | 0.753 | 0.876 | 0.000 | -0.016 | 0.016 | 0.000 | -0.002 | 0.002 |
| Untransformed power estimates | 0.785 | 0.663 | 0.907 |  |  |  | -0.001 | -0.003 | 0.001 |

*Note*. Estimates of the change per year in back-transformed units are calculated from the model intercept, and represent the estimated change per year at the mean year included in this study, 1985.

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

The results of the primary analysis suggest that the average power of published psychology literature is low for small or medium effects and only acceptable for large effects, and that these values have been remarkably stable over time. Given that many of the included power surveys suggest that the routine performance and reporting of power analysis is a key mechanism by which the average statistical power of psychology research could be improved, a important question is whether power analysis has become more common over time. A number of the power surveys included in the primary analysis also reported the proportion of articles they examined that reported a power analysis. This means that we can begin to examine this body of literature to see whether there has been any obvious increase in the number of articles that report a power analysis. In order to begin to address this question, we performed a meta-regression of the proportion of studies that report a power analysis, aggregating the results of all of the articles which reported how often power analyses were reported in their samples captured in the current literature search.

### Secondary Analysis Methods

Data extraction for the second analysis used the same randomization procedure as was used in the primary data extraction. During eligibility screening for the primary data extraction, we also examined whether articles reported the proportion of their surveyed articles that reported a power analysis. We additionally searched the reference lists of each applicable article for additional applicable articles. A total of 17 surveys were found, 15 of which were identified during eligibility screening for the primary analysis, and two that were identified through reference list searches of applicable articles. We extracted the years surveyed in each survey, the area of research examined (classified as in the primary analysis), the total number of articles examined in each survey, and the proportion of sampled articles which reported a power analysis. Data are available from <https://osf.io/h8u9w/>.

#### Sample characteristics

The 17 included articles reported the proportion of studies that reported a power analysis from a total of 21 distinct samples. See Table SM8 for the number of articles included in each sample, the population sampled from, and the years surveyed in each sample. Eight out 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.

Table SM8. *Original articles included in the secondary analysis.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Paper | Years Studied | Subfield | n | % Reporting a PA |
| Osborne (2008) | 1969 | Education | 55 | 0% |
| Katzer & Sodt (1973) | 1971-1972 | Communication | 31 | 6% |
| Crosby et al. (2006) | 1980 | Clinical | 152 | 0% |
| Sedlmeier & Gigerenzer (1989) | 1984 | Clinical | 64 | 0% |
| Crosby et al. (2006) | 1990 | Clinical | 152 | 0% |
| Kosciulek & Szymanski (1993) | 1990-1991 | Clinical | 32 | 3% |
| Kim (2015) | 1992-1993 | Medicine | 22 | 12% |
| ClarkCarter (1997) | 1993-1994 | General Psychology | 54 | 2% |
| Short, Ketchen, & Palmer (2002) | 1990-1999 | Management/IO | 288 | 6% |
| Cashen & Geiger (2004) | 1990-1999 | Management/IO | 43 | 7% |
| Osborne (2008) | 1998-1999 | Education | 96 | 2% |
| Bezeau & Graves (2001) | 1998-1999 | Clinical | 66 | 3% |
| Ward (2002) | 2000 | General Psychology | 103 | 7% |
| Crosby et al. (2006) | 2000 | Clinical | 152 | 2% |
| Woods et al., (2006) | 1997-2004 | Clinical | 30 | 0% |
| McKeown et al., (2015) | 2006-2013 | Medicine | 172 | 65% |
| Gaskin & Happell (2013) | 2010-2011 | Nursing | 23 | 17% |
| Guo et al., (2014) | 2010-2011 | Neuropsychology | 100 | 1% |
| Gaskin & Happell (2014) | 2011 | Nursing | 333 | 28% |
| de Bekker-Grob, Bas Donkers, Jonker, & Stolk (2015) | 2012 | Clinical | 69 | 12% |
| Larson & Carbine (2017) | 2010-2015 | Neuropsychology | 100 | 0% |
| Thombs & Rice (2016) | 2013-2015 | Clinical | 89 | 3% |

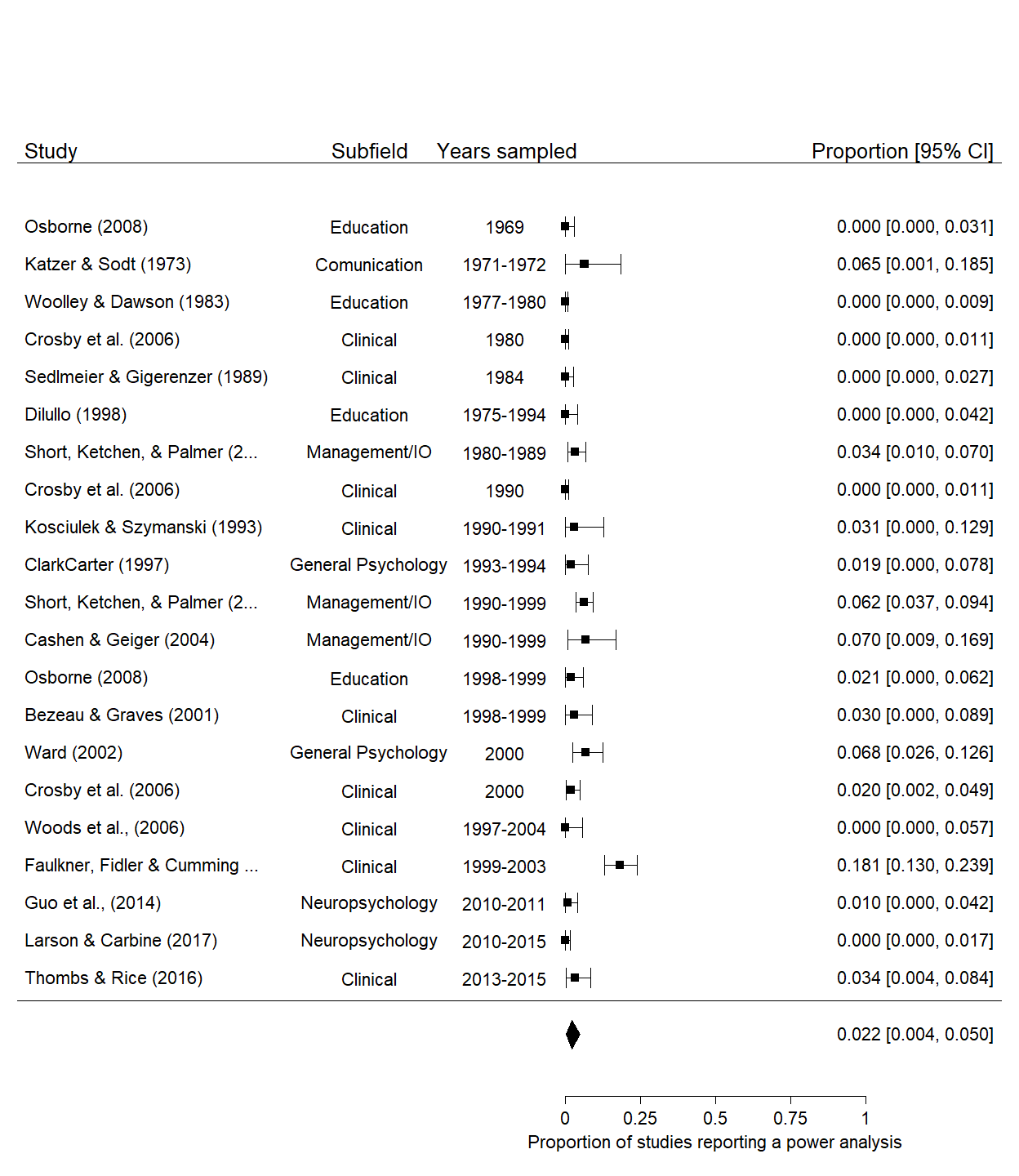
### Analysis

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.

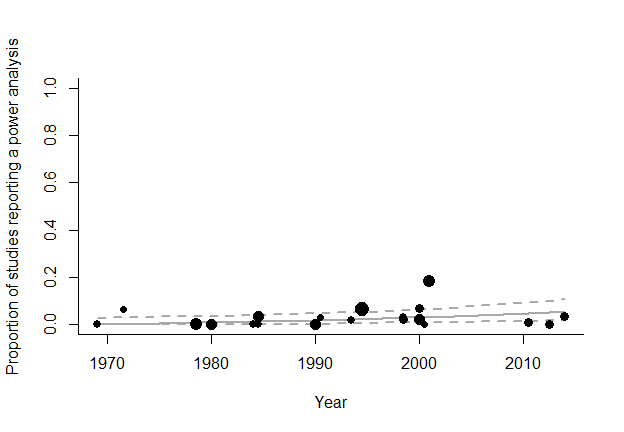
Proportions ( in the equation above) were transformed using the Tukey-Freeman Arcsine Transformation, as this can act to normalize the sampling distributions of proportions (Miller, 1978). This analysis included a parameter for year (), which was mean-centered. This means the intercept is interpretable as the estimated Tukey-Freeman Arcsine transformed proportion of studies for which power analyses are performed in the mean year included in this survey (1993). Random effects were included for individual estimates (), the survey that each estimate was collected as a part of (), and area of research (, e.g., clinical psychology). This analysis therefore accounts for non-independence between estimates from the same subfield of research, and from the same survey (i.e., when a survey reported estimates from multiple year ranges), and the random effect for each estimate makes this a random effects meta-analysis (i.e., relaxing the assumption that each proportion is an estimate of the same population parameter). If a survey examined articles from a a range of years, the mean year covered in the analysis was entered as the predictor (e.g., if a study examined 1980-1982, 1981 was entered as its value for year). Surveys that reported estimates for multiple year ranges (e.g., 1980-1982 and 1990-1992) were entered separately (i.e., a case was included for the 1980-1982 and for the 1990-1992 range). Restricted maximum likelihood estimation was used.

#### Results

The estimated percentage of papers reporting a power analysis is 2.2%, 95% CI [0.4%, 5%]. 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 A1 for a forest plot of this analysis and Figure A2 for a meta-regression scatterplot of the datapoints over time.



*Figure A1.* Forest plot of the proportion of articles reporting a power analysis. The polygon shows the model intercept.



*Figure A2.* Scatter plot of the estimated proportion of psychology articles reporting a power analysis. 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.

Table SM9. *Results of a meta-regression of the proportion of studies reporting a power analysis, including the year studied in each power survey as a moderator.*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variable | Estimate | 95% CI LB | 95% CI UB | SE | p | Random effects |
| Intercept | 0.168 | 0.096 | 0.238 | 0.036 | < .001 |  |
| Year | 0.004 | 0.001 | 0.007 | 0.002 | .018 |  |
|  |  |  |  |  |  | Subfield variance = 0.003, n = 6 |
|  |  |  |  |  |  | Article variance = 0.009, n = 17 |
|  |  |  |  |  |  | Effect variance < .001, n = 21 |
|  |  |  |  |  |  | QE(20) = 107.72, p < .001 |

In interpreting these results, 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, few recent studies were identified, and those which were identified 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 a considerable margin, 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).

#### Secondary analysis sensitivity analyses

No random effects were preregistered in the case of the secondary analysis, as are included above, so we also re-performed this analysis with out any random effects at the article or subfield level. Using this model, the estimated percentage of papers reporting a power analysis is slightly higher at 1.9%, 95% CI [0.6%, 3.6%]. There is again a very small estimated yearly increase in estimated power analysis reporting rates over time of 0.002, 95% CI [-0.002, 0.006] per year in Freeman-Tukey double arcsine transformed units. See Figure A1 for a forest plot of this analysis and Figure A2 for a meta-regression scatterplot of the datapoints over time. Although rerunning the analysis without random effects leads to a non-significant estimated change per year, as the null hypothesis of no change over time is implausible (given that techniques for power analysis for almost all statistical techniques we now use were not available in 1960, meaning power analyses could not be reported at that time, and as they are now reported at least occasionally), this shouldn’t be treated with much weight beyond reinforcing the fact that the change in reporting practices appears to be small in this sample.

Table SM10. *Results of a meta-regression of the proportion of studies reporting a power analysis, including the year studied in each power survey as a moderator and no random effects for study or area of research.*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Parameter | Estimate | 95% CI LB | 95% CI UB | SE | p |
| Intercept | 0.161 | 0.112 | 0.209 | 0.025 | < .001 |
| Year | 0.002 | -0.002 | 0.006 | 0.002 | .269 |