# Calibrating effect size estimates for bodies of research

Despite psychological research’s reliance on frequentist null hypothesis testing, in the great majority of cases the main question of interest is the nature (i.e., size and direct) of the effect or relationship between variables. Most readers will have the experience of seeing papers or talks where relationships between variables or the magnitude of effects are summarised with *p* values which do not characterise the size or direction of effect, telling us instead the probability of obtaining data as or more extreme under a specific null hypothesis. Given the evidence that researchers are (consciously or unconsciously) optimising their analysis methods to minimise *p* values and selectively reporting their results based on significance (Bakker, van Assen, Crompvoets, Ong, & Soderberg, 2017; John, Loewenstein, & Prelec, 2012), *p* values are often not interpretable in a strict Neyman-Pearson framework (i.e., as constraining type one, false positive, error rates). In order to allow for maximally interpretable results, the results of experiments need to be reported fully, and the direction and magnitude of effects need to be expressed.

When an effect or relationship exists, often the most meaningful summary will be on expressed in the raw units (e.g., a mean difference or SD), or a direct visualisation of the data. However, when the raw units are difficult to interpret, standardised effect sizes such as Cohen’s *d*, , or Pearson’s *r*, play an important role in succinctly expressing information about the effect or relationship. Although in many cases the expression of the effect of an intervention or the size of an effect might be more easily and understandably expressed in raw units, standardised effect sizes are now encouraged as they facilitate meta-analysis, comparisons between studies and across research contexts (Appelbaum et al., 2018). Standardised effect sizes allow for the results of an experiment to be expressed clearly and succinctly, and are helpful in performing formal sample size planning such as power analysis. Developing an understanding of effect sizes is becoming more important as they become more commonly reported and as psychology moves away from focusing only on statistical significance as the only indication of the presence or absence and importance of a given effect (Gigerenzer & Marewski, 2014).

In order to be able to understand and make use of standardised effect sizes in the context of scientific research, researchers need to have some understanding of the mathematical details of how they are estimated, as well as intuitive sense of what effects can be expected in a given area of research or how large observed effects are in context. There are many texts which provide an outline of the mathematical details (e.g., for a brief summary see [chapter effect sizes] or Lakens (2013)), but relatively few studies have attempted to address the latter issue of what effect sizes are routinely reported and what could reasonably be classified as a small or a large effect. Part of the reason for this is that the meaning and importance of a given standardised effect size is highly context dependent. If someone is studying a treatment for a common disease, an effect of Cohen’s *d* of .1 may represent an effect that could save thousands of lives. However, if someone is studying, for example, social media addiction, it is unlikely that a treatment that has an effect of .1 Cohen’s *d* would be pursued further. For this reason, attempting to provide universally applicable firm benchmarks on what a “small”, “medium” or “large” is foolhardy if not impossible.

Nonetheless the consumers and producers of research that is often reported and conveyed in standardised effect sizes need to be able to understand what effects can reasonably be expected in their area of research to effectively plan their research (e.g., using a power analysis), and to understand the relative import of observed effects in context. This paper focuses on three of the most common standardised effect size measures (Cohen’s *d* and and Pearson’s r), and uses three approaches to help provide researchers and research students an intuitive basis on which to understand these effect sizes; firstly by presenting non-technical definitions of each, secondly by providing examples from real-world, non-scientific scenarios, and finally by bringing together previous efforts which have been made to survey the effect sizes seen in various bodies of research to provide an idea of the distribution of effect sizes in fields of research.

## Cohen’s Benchmarks

**“The definitions are arbitrary, such qualitative concepts as "large" are sometimes understood as absolute, sometimes as relative; and thus they run a risk of being misunderstood.”**

**Cohen (1988, p. 12)**

Many of the most commonly used standardised effect size benchmarks were first proposed by Cohen (1962, 1970, 1988) in the context of power analysis (Huberty, 2002). In so far as current standards for classifying the importance and relative magnitude of observed effects, it seems that people have largely relied upon the standardised effect size benchmarks given by Cohen (1962, 1970, 1988), despite the practice being argued against as anything less than a last resort since their proposal (e.g., Thompson, 2007), including by Cohen himself (Cohen, 1988).

Table [effect sizes]. Effect size benchmarks following Cohen (1977, 1988, 1992)

|  |  |  |  |
| --- | --- | --- | --- |
| Effect size | Small | Medium | Large |
| d | .2 | .5 | .8 |
| r | .1 | .3 | .5 |
| w (φ) | .1 | .3 | .5 |
| OR b | 1.49 | 3.45 | 9 |
| *f* | .1 | .25 | .4 |
| *f 2* | .02 | .15 | .35 |
| a | .0099 | .0588 | .1379 |
| R2 | .02 | .13 | .26 |

Notes: a Transformed from Cohen’s benchmarks for *f*. Converted from Cohen’s benchmarks for *w* Cohen (1962) used slightly different estimates for small and large benchmarks (e.g., for *t* tests for mean differences small was a *d* of .25 and large a *d* of 1) although the medium benchmarks has remained the same.

**Using these effect sizes in power analysis**

Although selection of standardised effect sizes for use in power analysis using benchmark values, derived from commonly cited benchmarks or even those derived from a literature survey such as those which are presented here are the least preferred way of planning sample sizes, knowledge of what effect sizes can be reasonably expected in different areas of research are essential to developing reasonable effect size estimates. Without either performing a formal meta-analysis to derive effect sizes from previous studies, or using effect sizes directly seen in previous research (both approaches which can have their own issues (Kenny & Judd, in press; McShane & Böckenholt, 2016) and rely on their being a suitably comparable previous set of studies), researchers must select a minimum effect size of interest or assess whether it is likely that a given effect size is a plausible outcome from their experiment, both operations which require an intuitive understanding of what effect sizes mean and what can be reasonably expected in their area of research.

In part in order to prevent researchers from relying on these benchmarks, several efforts over the past half century have attempted to extract empirical benchmarks from bodies of psychological literature by systematically surveying papers and extracting the effect sizes that are reported. This paper presents a literature review of previous effect size surveys performed in psychology and educational research in order to bring together this body of literature and allow for us to begin to understand what types of effects are commonly reported across areas of psychology research, allowing researchers to develop expectations and intuitive understandings about the magnitude of other effects seen in the literature.[chapter effect size measures] presents the estimators that are associated with each of the effect sizes this paper refers to, as well as providing methods for estimating them from the published literature using more commonly reported statistics.

**Common language definitions of effect sizes**

Cohen’s *d* is the most commonly reported effect size in the psychological literature (Cumming et al., 2007) and in the case of independent groups describes the mean difference between groups standardised by their pooled standard deviations. In other words, Cohen’s *d* describes the size of the difference between two groups divided by how much variability is observed among individuals in the groups. The most commonly used estimator for Cohen’s *d* is upwardly biased in small sample sizes (i.e., it tends to overestimate the true population effect size), and Hedges’ *g* corrects for this fact, although the amount of bias is negligible in per group sample sizes above approximately 30 (Hedges, 1981). Glass’s delta is a similar effect size but only uses the standard deviation of the control group as opposed to assuming equal variances across groups, and is now rarely used (Smith & Glass, 1977).

Examples:

Male female height difference Cohen’s *d* of approximately 1.8 according to data from Garcia and Quintana-Domeque (2007) ( M – F / Sp = 174.5865 - 162.8726 / 6.424281 ).

For repeated measures designs there are multiple estimators for Cohen’s *d* (Lakens, 2013). For the purposes of easy interpretability, it will often be advisable to use Cohen’s *d* which does not incorporate the correlation between groups, or to use another estimator which is more easily comparable across situations (see for example Bonett, 2008). However, for the purposes of power analysis, Cohen’s dz, the Mean of the difference scores standardised by the standard deviation of the difference scores, is useful in that the specification of a dz value along with a sample size and alpha level is all that is required to provide a power estimate for a two group repeated measures design. However, the following benchmarks focus on effect sizes for mean differences (i.e., Cohen’s *d* for independent groups).

**Understanding Cohen’s *d***

Reiser and Faraggi (1999) provide a convenient way of transforming Cohen’s *d* for independent groups; as the expected proportion overlap of two populations given that each are normally distributed, have equal variance, and equal sample sizes. See Figure *[Cohen’s d as population distributions]* for a visual depiction of the proportion overlap expected at each of Cohen’s effect size benchmarks.



*Figure [Cohen’s d as population distributions]*. Population distributions and percentage overlap with a mean difference of .2, .5, .8 and 1.2 Cohen’s d (calculated assuming that populations are normally distributed, have equal variance, and equal sample sizes using equations from (Reiser & Faraggi, 1999)).

## Methods

### Review protocol

A review protocol was designed to return all articles which surveyed an area of research and reported the effect sizes reported in those articles. All searches were performed on the 11th August, 2018. The PsychInfo database was searched through the Ovid interface for “Effect size benchmarks.mp.” (“mp” searches for matches in the title, abstract, heading word, table of contents and key concepts), identifying 15 articles. Web of Knowledge was searched for “SU = Psychology AND TI = effect size benchmarks” (i.e., subject area psychology, and titles including ‘effect’ ‘size’ and ‘benchmarks’), identifying 5 articles. Additional searches for “average effect size” and “effect size benchmarks” in Google Scholar identified a further 6 articles. Hand searches of the references lists of all articles including during full text screening identified an additional 3 articles. I knew of two articles outlining effect size benchmarks from the grey literature, a pre-print (Lovakov & Agadullina, 2017) and book (Hattie, 2009) which are included. After deduplication and full text screening, 18 documents remained, 3 of which were removed after full text screening (see figure [prisma]) leaving 15 documents.



Figure [prisma]. Prisma diagram of the article search and screening procedure.

## Results

See tables [education - effect sizes not r or d] for the extracted values from each paper.

Table [education]. The mean effect size and standard deviation reported in educational studies

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Authors (year) | Sampled groups | Unit of Analysis | Number of Effects | Mean effect size (Cohen’s *d*) | SD of effect sizes |
| Hill et al., (2008) | Elementary school RCTs | Effect sizes | 389 | 0.33 | 0.48 |
| Hill et al., (2008) | Middle school RCTs | Effect sizes | 36 | 0.51 | 0.49 |
| Hill et al., (2008) | High school RCTs | Effect sizes | 43 | 0.27 | 0.33 |
| Hill et al., (2008) | Meta-analyses of elementary school intervention studiesa | Meta-analytic effect size estimates | 32 | 0.23 | 0.21 |
| Hill et al., (2008) | Meta-analyses of middle school intervention studiesa | Meta-analytic effect size estimates | 27 | 0.27 | 0.24 |
| Hill et al., (2008) | Meta-analyses of high school intervention studiesa | Meta-analytic effect size estimate | 28 | 0.24 | .15 |
| Hattie (2009) | Meta-analyses of educational interventions | Effect sizes | 146,626 | 0.4 | *NA* |

Note: aInterventions included in sourced from Bloom et al., 2007b or Lipsey et al., 2007. Hattie 2009 included 816 meta-analyses, including a total of 52,649 articles.

Hill et al., did not report the total number of meta-analyses or effects included in their study.

Table [effect sizes d psychology]. Results of effect size surveys reporting Cohen’s *d* and examining psychology research

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Authors (year) | Area of research | Location effects sampled from | n effects | n meta-analyses | n articles | Mean effect | SD effect size | 25th Percentile | Median effect | 75th percentile |
| Cooper, & Findley (1982) | Social psychology | Results reported in social psychology textbooks | 14 | NA | 14 | 1.19 | 0.62 |  |  |  |
| Lipsey & Wilson (1993) | Psychological interventions | Meta-analytic estimates of psychological interventions’ effects | 302 | 302 | NA | 0.5 | 0.29 |  | 0.47 |  |
| Szucs, & Ioannidis (2017)a | Cognitive neuroscience, psychology and psychiatry | Statistical tests reported in cognitive neuroscience, psychology, psychiatry articles published in high impact journals, 2011 - 2014 | 26841 | NA | 3801 | 0.938 |  |  | 0.654 |  |
| Szucs, & Ioannidis (2017) a | Cognitive neuroscience | Statistical tests reported in cognitive neuroscience articles published in high impact journals, 2011 - 2014 | 7888 | NA | 1192 |  |  | 0.34 |  | 1.22 |
| Szucs, & Ioannidis (2017) a | Psychology | Statistical tests reported in psychology articles published in high impact journals, 2011 - 2014 | 16887 | NA | 2261 |  |  | 0.29 |  | 0.96 |
| Szucs, & Ioannidis (2017) a | Psychiatry | Statistical tests reported in articles published in high impact journals, 2011 - 2014 | 2066 | NA | 348 |  |  | 0.23 |  | 0.91 |
| Qunitana (2017) | Hear rate variability studies | Effect sizes from meta-analyses of Heart Rate Variability Studies | 297 | 9 | 293 |  |  | 0.26 | 0.51 | 0.88 |
| Bergmann et al., (2018) | Language Acquisition Research | Effects reported in articles included in Meta-lab (http://metalab.stanford.edu.) | NA | 12 |  |  |  |  | 0.45 |  |
| Smith & Glass (1977)b | Clinical psychology | Effect sizes from studies of psychotherapy with a non-treatment control group published before 1977 | 833 | NA | 375 | 0.68 | 0.67 |  |  |  |
| Andrey & Agadullina (2018) | Social psychology | Effects included in in meta-analyses published in 29 journals in the "Psychology, Social" category of Social Sciences Citation Index | 3498 | 42 | 1922 |  |  | 0.15 | 0.38 | 0.69 |

Table [rs]. Results of effect size surveys of Pearson correlation coefficients.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Authors (year) | Area of research | Sampled groups | n effects | n meta-analyses | n articles | Mean effect | SD effects | 25th Percentile | Median effect | 75th percentile |
| Cooper, & Findley (1982)a | Social psychology | Main result of articles reported in social psychology textbooks reporting r | 23 | NA | 23 | 0.48 | 0.22 |  |  |  |
| Richard, Bond Jr, & Stokes-Zoota (2003) | Social psychology | “Conclusions” from literature search for Social psychology meta-analyses | 474 | 322 | NA | 0.21 | 0.15 |  | 0.18 |  |
| Hemphill (2003) | Clinical psychology | Meta-analytic effect size estimate from articles included in Meyer et al., 2001 or Lipsey & Wilson, 1993 | 380 | 380 | NA |  |  | 0.15 |  | 0.35 |
| Paterson et al., (2015) | Management and applied psychology | Effect sizes from meta-analyses published in the top 30 impact factor management journals before 2012 | 776 | 258 | NA | 0.227 | 0.135 |  | 0.2 |  |
| Bosco et al. (2015) | Management and applied psychology | Effects reported in the first correlation table of articles published in the Journal of Applied Psychology and Personnel Psychology from 1980 to 2010 | 147328 | 816 | 1660 | 0.32 | 0.22 | 0.07 | 0.16 | 0.16 |
| Gignac & Szodorai (2016) | Personality and Social psychology | Effects of studies included in meta-analyses of correlational studies published in Personality and Individual Differences, Psychological Bulletin, Journal of Research in Personality, Journal of Personality and Social Psychology, Journal of Personality, and Intelligence, from 1985-2015 | 708 | 199 | NA |  |  | 0.11 | 0.19 | 0.29 |

Note: a Cooper, & Findley (1982)results also reported in Table [effect sizes not r or d].

Table [effect sizes not r or d]. Results of effect size surveys of assorted effect size benchmarks.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Authors (year) | Area of research | Sampled groups | n Effects | N articles | Mean effect size | SD effect size | 25th Percentile | Median effect size | 75th percentile | Effect size unit |
| Haase, Waechter & Solomon, (1982) | Clinical psychology | Each univariate inferential statistic reported in the Journal of Counselling Psychology, 1970-1979 | 11,044 | 701 | 0.1589 |  | 0.0428 | 0.083 | 0.2682 |  |
| Cooper, & Findley (1982) | Social psychology | Main result of articles reported in social psychology textbooks reporting f (df = 1) | 113 | 113 | 0.45 | 0.3 |  |  |  | f (df = 1) |
| Cooper, & Findley (1982) | Social psychology | Main result of articles reported in social psychology textbooks reporting f (df > 1) | 72 | 72 | 0.6 | 0.54 |  |  |  | f (df > 1) |
| Cooper, & Findley (1982)a | Social psychology | Main result of articles reported in social psychology textbooks reporting r | 23 | 23 | 0.48 | 0.22 |  |  |  | r |
| Cooper, & Findley (1982) | Social psychology | Articles reported in social psychology textbooks reporting w (df = 1) | 15 | 15 | 0.26 | 0.16 |  |  |  | w (df = 1) |

Note: aCooper, & Findley (1982) correlational studies results also reported in Table [rs]

## Discussion

A total of 15 articles were identified provided empirical benchmarks for areas of psychology research. The earliest study to examine the distribution of effect sizes in a body of literature identified is by two of the progenitors and popularisers of meta-analysis, Smith and Glass (1977). This early meta-analysis included 375 studies of psychotherapy which compared a treatment group to a control group, using an unusual effect size measure roughly equivalent to Cohen’s d (Glass’ delta, an effect size measure that standardised the effect size with the standard deviation of the control group not a pooled estimate).

These studies do not provide a comprehensive assessment of all areas of psychology research, and often present values that if taken at face value as estimates of the average power of an area of research are likely to be severe overstates. For example, Cooper and Findley (1982) examine effect sizes reported in social psychology textbooks, articles which seem likely to show particularly large effects compared to other studies. In so far as the studies reported in textbooks are seen as illustrations of important effects worthy of coverage and due to the “Proteus phenomenon” (Button et al., 2013), that initial studies published on a topic may exaggerate effect sizes as compared to later studies.

Additionally, the database presented above illustrates the degree of heterogeneity that can be seen in different areas of published psychological research. With means effects in various areas of psychology as different as a *d* of .5 from meta-analyses of Psychological interventions (Lipsey & Wilson, 1993) to d = .94 from a text scrapping study examining recently (2011 – 2014) published *t*-tests reported in cognitive neuroscience, psychology, psychiatry articles in a sample of high impact journals.

Additionally, the fact that the medians tend to be lower than the means points to an important insight, that effect sizes reported in psychology are likely to be heavily positively skewed. I.e., there some very large effects reported and some which are much smaller. Importantly, none of these articles attempt to address the issue of publication bias increasing average effect sizes in the published literature. Given the difference between original study and replication attempt effect sizes that has been seen in all of the large scale replication studies it is likely that these are all severe overestimates and should be understood as upper bounds on the size of the effects under study (Anderson & Maxwell, 2017; Maxwell, Lau, & Howard, 2015; Open Science Collaboration, 2015).

Anderson, S. F., & Maxwell, S. E. (2017). Addressing the "Replication Crisis": Using Original Studies to Design Replication Studies with Appropriate Statistical Power. *Multivariate Behavioral Research, 52*, 305-324. doi:10.1080/00273171.2017.1289361

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report (Vol. 73, pp. 3-25). US: American Psychological Association.

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.*

Bonett, D. G. (2008). Confidence intervals for standardized linear contrasts of means. *Psychol Methods, 13*, 99-109. doi:10.1037/1082-989x.13.2.99

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafo, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*, 365-376. doi:10.1038/nrn3475

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. (1970). Approximate power and sample size determination for common one-sample and two-sample hypothesis tests. *Educational and Psychological Measurement, 30*, 811-831.

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

Cooper, H., & Findley, M. (1982). Expected Effect Sizes. *Personality and Social Psychology Bulletin, 8*, 168-173. doi:10.1177/014616728281026

Cumming, G., Fidler, F., Leonard, M., Kalinowski, P., Christiansen, A., Kleinig, A., . . . Wilson, S. (2007). Statistical reform in psychology: Is anything changing? *Psychological Science, 18*, 230-232. doi:10.1111/j.1467-9280.2007.01881.x

Garcia, J., & Quintana-Domeque, C. (2007). The evolution of adult height in Europe: A brief note. *Economics & Human Biology, 5*, 340-349. doi:<https://doi.org/10.1016/j.ehb.2007.02.002>

Gigerenzer, G., & Marewski, J. N. (2014). Surrogate Science: The Idol of a Universal Method for Scientific Inference. *Journal of Management, 41*, 421-440. doi:10.1177/0149206314547522

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. London, England: Routledge.

Hedges, L. V. (1981). Distribution Theory for Glass's Estimator of Effect size and Related Estimators. *Journal of Educational Statistics, 6*, 107-128. doi:10.3102/10769986006002107

Huberty, C. J. (2002). A history of effect size indices. *Educational and psychological measurement, 62*, 227-240. doi:10.1177/0013164402062002002

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

Kenny, D. A., & Judd, C. M. (in press). The Unappreciated Heterogeneity of Effect Sizes: Implications for Power, Precision, Planning of Research, and Replication.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in Psychology, 4*, 863. doi:10.3389/fpsyg.2013.00863

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist, 48*, 1181-1209. doi:10.1037/0003-066X.48.12.1181

Lovakov, A., & Agadullina, E. (2017). Empirically Derived Guidelines for Interpreting Effect Size in Social Psychology. doi:10.31234/osf.io/2epc4

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

McShane, B. B., & Böckenholt, U. (2016). Planning sample sizes when effect sizes are uncertain: The power-calibrated effect size approach. *Psychological Methods, 21*, 47-60. doi:10.1037/met0000036

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*.

Reiser, B., & Faraggi, D. (1999). Confidence Intervals for the Overlapping Coefficient: the Normal Equal Variance Case. *Journal of the Royal Statistical Society: Series D (The Statistician), 48*, 413-418. doi:10.1111/1467-9884.00199

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist, 32*, 752-760. doi:10.1037/0003-066X.32.9.752

Thompson, B. (2007). Effect sizes, confidence intervals, and confidence intervals for effect sizes. *Psychology in the Schools, 44*, 423-432. doi:10.1002/pits.20234