Understanding effect sizes in context

There are various uses for standardised effect sizes such as Cohen’s *d*, , or Pearson’s *r*, in psychology research. 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 (i.e., when raw units are directly interpretable or widely understood), standardised effect sizes are now encouraged as they facilitate meta-analysis, comparisons between studies and across research contexts (Appelbaum et al., 2018). Often standardised effect sizes they 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 a 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 understand and make use of standardised effect sizes in the context of scientific research, one needs to have an understanding of the technical or 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 .01 may signifiy a treatment that could save thousands of lives. However, if someone is studying social media addiction, it is unlikely that a treatment that has an effect of .01 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. Several efforts over the past half century have attempted to provide these benchmarks on a literature wide scale by systematically surveying bodies of literature, and extracting the effect sizes that were observed in the literature. However, these efforts have never been brought together to facilitate readers understanding of not just the effects seen in their narrow field of expertise but also the variety of effects sizes that are observed across fields. In order to fill this gap, this paper is a literature review of previous effect size surveys performed in psychology and educational research. Bringing together this body of literature allows for us to begin to understand what types of effects are commonly reported to have been found in different areas of psychology research, allowing researchers to develop expectations and intuitive understandings about the magnitude of other effects seen in the literature.

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 consistently since their proposal (e.g., Thompson, 2007) [MORE CITATIONS]. For example, when asked in a survey what effect size they expect to see in their research over half of academic psychology researchers who provided an effect size in Pearson r or Cohen’s *d* responded that they expected an effect size equal to one of Cohen’s benchmarks (see effect size chapter). This paper provides academic researchers a go-to source for examining empirical benchmarks as opposed to those provided by Cohen in order to rely on the closest possible comparison group, as well as providing and educational resource for those who want to understand the types of effects seen in their own and other people’s areas of research.

Many of the most commonly used standardised effect size measures and benchmarks were first proposed by Cohen (1962, 1970, 1988) in the context of 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 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-anlaysis 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.

Importantly, the reported effect sizes in the liteartuer may not be representative of the true effects (… publication bais etc).

Other ways of expressing effect sizes

Proportion overlap (U) –



*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)).

Examples of effect sizes

Table [Cohen] Cohen’s (1988) Benchmarks for different types of effect size

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

Notes: a Transformed from Cohen’s benchmarks for *f*. Note that the benchmarks for *f 2* are not simply *f* benchmarks squared. ….. THIS IS BECAUSE

Literature search method

A systematic 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]. Article search and screening procedure.

Results Discussion

A total of 19 articles were identified provind 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. Although the explicit aim of this study was not to,

It is also noteworthy that this is the only study to use 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 provide an overview of the degree of heterogeneity that can be seen in different areas of published psychological research.

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 (Anderson & Maxwell, 2017; Maxwell, Lau, & Howard, 2015; Open Science Collaboration, 2015).

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 |

Note:

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]

FIND SOME WAY OF VISULISING HERE ~ Maybe use the kernel density estimatiors, imputing data haphazardly

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.

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.

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.

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

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