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. Most obviously, they allow for the results of an experiment to be expressed clearly and succinctly transferring information. 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). Furthermore, they 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, for a brief summary see [chapter effect sizes] or Lakens (2013). The current paper focuses on the second aspect, on developing an intuitive understanding of effect sizes. Toward that end, this paper outlines various approaches to understanding standardised effect sizes, giving intuitive real-world examples and providing a systematic review of previous efforts to provide empirical effect size benchmarks, benchmarks that are based on the observed effect sizes in the literature.

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, either those derived from a specific literature, or those derived from suggestions from researchers (e.g., Cohen, 1988), 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 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)), 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.

… Explain how Cohen came up with these benchmarks.

Other ways of expressing effect sizes

Proportion overlap (U) –



*Figure [Cohen’s d as population distributions]*. Population distributions with a mean difference of .2, .5, .8 and 1.2 Cohen’s d, along with the percentage overlap between populations (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) | .1 | .3 | .5 |
| OR (converted from w) CHECK COHEN FOR THIS | 1.49 | 3.45 | 9 |
| *f* | .1 | .25 | .4 |
| *f2* | .02 | .15 | .35 |
| b | .0099 | .0588 | .1379 |
| R2 | .02 | .13 | .26 |

Notes: a Transformed from Cohen’s benchmarks for *f*

**Difference Between Two Means\***

|  |  |  |
| --- | --- | --- |
| Size of effect | *d* | % variance |
| small | .2 | 1 |
| medium | .5 | 6 |
| large | .8 | 16 |

Cohen’s *d* is not influenced by the ratio of *n1* to *n2*, but *rpb* and eta-squared are.

**Pearson Correlation Coefficient**

|  |  |  |
| --- | --- | --- |
| Size of effect | *ρ* | % variance |
| small | .1 | 1 |
| medium | .3 | 9 |
| large | .5 | 25 |

**Contingency Table Analysis**

|  |  |  |
| --- | --- | --- |
| Size of effect | *w* = φ | odds ratio\* |
| small | .1 | 1.49 |
| medium | .3 | 3.45 |
| large | .5 | 9 |

\*For a 2 x 2 table with both marginals distributed uniformly.

**ANOVA Effect**

|  |  |  |
| --- | --- | --- |
| Size of effect | *f* | % of variance |
| small | .1 | 1 |
| medium | .25 | 6 |
| large | .4 | 14 |

**Multiple *R2***

|  |  |  |
| --- | --- | --- |
| Size of effect | *f2* | % of variance |
| small | .02 | 2 |
| medium | .25 | 13 |
| large | .4 | 26 |

Systematic review 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 serachers 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”, identifying 5 articles. Additional searches for “average effect size” and “effect size benchmarks” in Google Scholar identified 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 articles remained.



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 of psychological interventions which reported results in Cohen’s d.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Authors (year) | Area of research | Sampled effects | n effects | n meta-analyses | n articles | Mean effect | SD effect size | 25th Percentile | Median effect | 75th percentile |
| Cooper, & Findley (1982) | Social psychology | Cited 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) | 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 |  |  |  | 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