# Effect sizes for planning and interpreting research, definitions and empirical benchmarks

## x.1 Standardized effect size typology

Over the last 30 years an increased focus has been placed on the reporting and interpretation of effect sizes as an important part of the development of a cumulative and interpretable research literature (e.g., Cumming, 2013; Hedges, 1981; Kruschke & Liddell, 2017; Wilkinson, 1999). Effect sizes can be expressed in standardised or unstandardized units. Unstandardized effect sizes (e.g., mean differences) are presented in the units the measured variables, and may be particularly useful when the units of analysis are directly interpretable (e.g., income, IQ scores, measures of height or weight). Standardised effect sizes (e.g., Cohen’s *d* for mean differences) have several distinct uses. Some measures may be useful for facilitating interpretation when the units of measurement are not themselves interpretable (e.g., a newly developed measure), as they express observed patterns in the data in a way interpretable without reference to the units of measurement. This chapter provides a definitions and explanation of the most common effect sizes used and reported in psychological research, and presents the results of review of studies which report empirical benchmarks from the published literature. This chapter aims to act as a resource to aid the development of an understanding of the technical details of these effect sizes alongside an intuitive understanding of the distribution of effects that are seen across bodies of published research. Secondarily, this chapter acts to define the estimators that are associated with the effect sizes used in this dissertation in the notation that is followed throughout this thesis.

In many if not most studies reported in the Behavioural Sciences literature the main question of interest is the nature (i.e., size and direction) of an effect or relationship between variables. When an effect or relationship exists, often the most meaningful way of summarising some data would be to express the effect in the raw units (e.g., a mean difference and SD), or a directly visualising 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 size, direction and strength of an effect or relationship. The reporting of standardised effect sizes is now encouraged as this practice facilitates meta-analysis, allows for studies’ effects to be compared and collapsed, as well as allowing for comparisons between studies and across research contexts (Appelbaum et al., 2018).

Standardised effect sizes are also helpful in performing formal sample size planning (such as power analysis). In order to perform formal sample size determination like power analysis, researchers must specify an alternative hypothesis in sufficient detail to determine the sampling distribution of the test statistic under a specific alternative hypothesis. For relatively simple designs (e.g., for a comparison of the mean scores of two independent groups or correlational analysis) the specification of a single standardised effect size characterises the sampling distribution under the alternative hypothesis adequately for power analysis (Cohen, 1988). For more complex designs (e.g., when covariates are to be included or when repeated measures designs are used) additional parameters may need to be specified. One of the major difficulties often cited by researchers in performing a power analysis is that they have trouble developing appropriate parameters for use in power analysis [cite interviews and survey].

### 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 size benchmarks 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.

**Aims of the current paper**

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 or relationship (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 effect sizes are estimated, as well as intuitive sense of what effects can be expected in a given area of research. There are many texts which provide an outline of the mathematical details (e.g Lakens (2013)), but there are few articles have attempted to address the issue of what effect sizes 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. In part in order to prevent researchers from relying on arbitrary benchmarks, a number of papers published over the past half century have presented empirical benchmarks extracted from bodies of psychological literature by systematically surveying papers and extracting the effect sizes that are reported. This chapter has three main goals. One is to bring 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 various fields of research. The second goal is to present and explain effect size definitions and estimators alongside these benchmarks. Thirdly, this chapter provides a reference or glossary for the nomenclature that will be used throughout this thesis when referring to effect sizes. This chapter groups standardized effect sizes into three main categories; effect sizes for group differences (e.g., Cohen’s *d* and Hedge’s *g*), variance explained effect sizes (e.g., r, R2, eta2, partial eta2, omega2), and probability effect sizes (e.g., odds ratios, Cohen’s w).

### Methods - Review protocol

In order to identify articles provided effect size benchmarks for a body of literature the PsychInfo and Web of knowledge databases were searched. Psychinfo 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, 15 independent articles were identified which provided empirical effect size benchmarks for fields of research. All searches were performed on the 11th August, 2018. See Figure [prisma] for a Prisma diagram of the article search and screening procedure.



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

## Effect sizes for Mean differences

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. Cohen’s *d*, was originally proposed as an measure of the size of effect in Cohen’s first power survey, and was explicitly developed to aide in sample size determination (Cohen, 1962). There are a number of estimators for the population parameter the difference between groups divided by the pooled standard deviation. The estimates produced by all of these estimators are commonly called “Cohen’s *d*”, and all use equation x.1.

(x.1)

(adapted from McGrath & Meyer, 2006, p. 386)

Whereis the mean of sample 1, and is the mean of sample 2, and is the pooled standard deviation. The pooled standard deviation is most often calculated for samples as:

(x.2)

(Cohen, 1977, p. 67)

Or equivalently as equation x.3[[1]](#footnote-1).

(x.3)

(adapted from Hedges, 1981, p. 110)

Where is the sample variance for each group, calculated as per equation x.4

(x.4)

j indicating the group. The pooled standard deviation should be calculated for populations (i.e., if all possible units of analysis have been collected) using n1+n2 in the denominator as opposed to n1+n2-2, without Bessel’s correction (Cohen, 1977, 1988; McGrath & Meyer, 2006).

Terminology around these effect sizes is remarkably inconsistent, and sometimes Cohen’s d is reserved to describe the estimator that doesn’t use Bessel’s correction, and the estimator outlined in equation x.1 to x.3 is called Hedges’ g (e.g., (Rosenthal, 1991)). However, as Cohen outlined both estimators (e.g., Cohen, 1977) before Hedges (1981), and as the population version is rarely applicable, it seems reasonable to use Cohen’s *d* to refer to the estimator outlined in equations x.1 to x.4. This estimator for Cohen’s d is consistent (that is, as the n increases its expectation increasingly accurately approximates the population parameter), but it is upwardly biased (it tends to overestimate the population parameter, especially when the included sample size is small). Hedges (1981) outlines a correction factor to produce an unbiased estimator:

(x.5)

(Originally from Hedges, 1981; this version adapted from Hedges & Olkin, 1985, p. 104).

Where for an independent groups design, d is calculated as per equation x.1 and is the gamma function. However, this correction factor is fairly complex (although trivial on modern computers), and Hedges also provides a computationally simple approximation which performs well for all practical scenarios (Hedges, 1981, p. 114).

Hedge’s approximate bias corrected *g*\* is calculated as:

(x.6)

Where for an independent groups design and *d* is Cohen’s d as calculated in equation x.1 using x.2 as the estimator for the variance.

(this version adapted from Borenstein, Hedges, Higgins, & Rothstein, 2011, p. 27; originally from Hedges, 1981)

In the literature ‘Hedge's g’ or ‘Cohen's d’ are used interchangeably to refer to , and (Lakens, 2013). , and are all virtually identical for most practical purposes when *n* > 30, and all are estimators of the same population parameter. For the purposes of power analysis, it is important to realise that is upwardly biased and increasingly so in smaller samples. However, simple sampling variability and selective reporting are likely to cause greater difficulties in determining the ‘true’ effect size than the bias of the estimator that has been used. Also worth noting is Glass’s delta, a similar effect size which only uses the standard deviation of the control group as opposed to assuming equal variances across groups (Smith & Glass, 1977), however this effect size is rarely used in contemporary research.

If effect sizes have not been reported, Cohen’s d can be calculated using the results of an independent samples t tests using the formula

x.7 (Lakens, 2013, equation 2)

Where and are the sample sizes for groups 1 and 2 respectively, and *t* is the result of an independent samples t test.

Alternatively, if only the total sample size is available x.8 can be used, although it is only correct if the groups are equal and will be an underestimate if the groups are unequal. However, even if the ratio of samples sizes in each group is as extreme as 70 to 30 the underestimation will be no more than 8% (Rosenthal, 1991).

x.8 (Rosenthal, 1991, p. 17)

Cohen’s *d* can be estimated from *r*, the Pearson product moment correlation coefficient

(Borenstein et al., 2011) equation 7.5

Where d’s variance is :

(Borenstein et al., 2011) equation 7.6

And Vr is the variance of *r.*

#### Understanding Cohen’s d in context

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)).

A number of projects have also developed empirical effect size benchmarks from various fields of research for Cohen’s *d*. See tables [education] for a summary of the average effect sizes seen in educational research, and Table [effect sizes d psychology] for the average effect sizes seen more broadly in psychological research. In order to put the average effect sizes seen in these fields in context, the height difference between people who identify as male (with an average height around 174 cm) and people who identify as female (with an average height around 164 cm) represents a Cohen’s *d* of approximately 1.8 , where the pooled standard deviation is 6.4 (calculation performed on data from Garcia and Quintana-Domeque (2007)).

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 sizes | 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 |

#### Standardised mean differences for the comparisons of two repeated measures:

The most common effect size measure for mean difference between repeated measures is also commonly called Cohen’s d. For repeated measures designs there are multiple estimators for Cohen’s *d* (Lakens, 2013). Following Cohen (1977, 1988) I will refer to the repeated measures version as Cohen’s . This effect size follows a similar general form to the independent samples Cohen’s d (x.1), except the numerator is the mean difference between measures.

x.9 Lakens (2013) equation 6

Where is the mean difference score, and is the standard deviation of the difference scores calculated as:

Where is the difference scores for case *i*, is the mean difference score, and is the standard deviation of the difference scores. Equivalently, can be calculated as

(Cohen 1988 p. 48)

Where and are the variances of groups one and two, and is equal to the Pearson correlation between subjects measures on measure one and measure two. Notably, this equation x.9 highlights an important fact Cohen’s , and repeated measures t tests, that the effect size is dependent upon the correlation between scores on repeated measures. The higher the correlation, the greater the . This can make a large difference to the , for example, taking the example the two groups have equal variance, and there is a one standard deviation difference between groups, can vary between .707 (when ) and infinity (when approaches 1). For this reason it has been argued that classical Cohen’s d (equation x.1) should be interpreted in lieu of for maximum interpretability and comparability across experimental designs (Morris & DeShon, 2002). However, for the purposes of power analysis, it is beneficial to use , as the correlation between repeated measures increases the standard error of the difference decreases, or equivalently, the size of the test statistic increases, as can be seen in equation [Rosen]:

equation [Rosen] from Lakens (2013) Equation 7

Where is the *t* statistic calculated as per a repeated measures *t* test, and *n* is the sample size.

is also biased, and an equivalent to Hedges’ correction can be applied to adjust for this bias similarly to the independents samples Cohen’s d (Gibbons, Hedeker, & Davis, 1993):

(x.Gibbons) (equation 7, p. 274 Gibbons, Hedeker & Davis, 1993)

Where is degrees of freedom (i.e., as per repeated measures *t*-tests) and is the gamma function.

## Effect sizes for association/variance explained:

Almost certainly the most commonly used measure of association and one that almost all psychological scientist will be family with is Pearson’s Product Moment Correlation Coefficient, Pearson *r*. One of the oldest standardised effect sizes commonly used today, r measures the degree of linear association between two variables and was pioneered by Galton and further developed by Karl Pearson (Pearson, 1903).

Where *x* are the values of x, y are the values of y, and n is the number of pairs of scores.

r2 equals the total variation in one variable that can be predicted through its linear association with another.

*r* can be estimated from *d* values as

(Borenstein et al., 2011) equation 7.7

In which a is a correction factor for unequal group sizes

(Borenstein et al., 2011) equation 7.8

In this case the variance of Vr can be calculated as

A number of papers have extracted effect sizes from various areas of psychological research, see table [correlations] for a summary of the studies which have reported empirical benchmarks alongside a description of their sampled populations.

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].

Multiple R2

Multiple R2 will be familiar to people from a regression context, and describes the proportion of variance explained, but with multiple predictor variables. When there are two predictor variables,

Where is the correlation between the dependent variable and the first predictor, and

This estimator is increasingly upwardly biased (i.e., it overestimates the amount of variance explained at the population level) as more predictors are introduced, so (Theil, 1958) proposed an alternative estimator which adjusts for the sample size and number of included predictor variables often called adjusted R2.

Where *n* is the sample size and *k* is the number of predictor variables (Miles, 2004).

An analogous effect size to R2 in the context of ANOVA is η2 (eta squared), the ratio of the sums of squares between groups to the total sums of squares which can be estimated as equation [eta].[[2]](#footnote-2) Eta squared is one of the oldest standardised effect sizes outlined here apart from natural effect sizes such as correlation, and has been around since at least 1939 (Goulden, 1939).

Equation [eta]

Where *SSeffect* is The sums of squares between groups, , where *n* is the sample size in each group, is the kth group’s mean and is the grand mean. *SStotal* is equal to the total sums of squares, , with being the ith item’s value on y.

Whereas in general R2 is used to discuss the variability accounted for in a model, η2 is generally used to describe the proportion of variance accounted for by a single factor.

This is different to another related and commonly reported effect size, or partial eta squared, which describes the proportion of variance that can be attributed to a particular factor after excluding variance explained by other factors in the model.

Equation 2, (Levine & Hullett, 2006).

Where is the sum of squared residuals, , with being the predicted value of i or equivalently the mean of the group under study. and are equal in one-way ANOVAs as all summed and squared errors are included in the error term, in multiway or repeated measures ANOVA partial eta squared will be larger as the additional factors are not included in the denominator.

Although it has been argued that researchers should favour one over the other (e.g., Levine & Hullett [2006] ague for favouring , whereas Richardson (2011) argues that will be the more meaningful statistic in most cases), both effect sizes are meaningful in different scenarios. If a researcher wants to describe the effect of a single factor such also has additional factors in their model will be more meaningful, whereas if a researcher wants to describe the total variance explained by all factors in their model they should report . Adding to the confusion between these two statistics is the fact that although was produced in multi-way ANOVAs in SPSS, it was mislabelled as in SPSS versions 7 to 10 meaning that it is likely that many of the eta square values in the literature produced using SPSS are in fact partial eta squared (Levine & Hullett, 2006; Richardson, 2011).

can be calculated from reported F statistics (Richardson, 2011).

However and are upwardly biased. Two other estimators have been proposed,ε2 (Epsilon squared) and ω2 (omega squared).

#### Epsilon squared ε2

Epsilon squared is equivalent to adjusted R**2**, adjusting the effect size downwards as the number of factors gets larger and as the sample size decreases. can be calculated as

Equation 4 (Carroll & Nordholm, 1975), or equivalently

Equation 7 (Carroll & Nordholm, 1975)

Where N equals the sample size and is equal to the number of levels of the factor minus one, and isthe mean squares error (or the within groups mean square). also estimates the proportion of variance explained after all other sources of variance included in the model have been partialled out and can be calculated from observed F statistics and associated degrees of freedom.

Appendix A (Albers & Lakens, 2018)

#### Omega squared ω2

A similar alternative estimator is ω2 (omega squared).

Equation 6 (Carroll & Nordholm, 1975)

Where is the sum of squares for the effect, is total sum of squares, and is the error mean squares.

, which again estimates the proportion of variance explained by a given factor after all other sources of variance have been partialled out can be calculated as equation [maxwell]

**(S. Maxwell, Camp, & Arvey, 1981) equation 26**

can also be calculated from reported F statistics

Appendix A, (Albers & Lakens, 2018)

It has been pointed out that and are “essentially the same in practice” as they only differ by (Carroll & Nordholm, 1975, p. 544), an amount that will be negligible for most practical purposes. For use in power analysis for statistical tests of the effect of a factor, the main effect size of interest is the ‘partial’ version of these statistics (i.e., , and ) as they are estimates of the proportion of variance explained of the otherwise unexplained variance, the same value that impacts the size of the F statistic and the significance of these types of statistical tests. As for which of these estimators is preferable, although both and upwardly biased and more variable than (Levine & Hullett, 2006). Albers and Lakens (2018) argue that it is preferable to use or for power analysis as simulations show that they lead to better power studies on average, although further research is necessary to determine the optimal effect size to use in other situations.   
Generalised variance explained measures

, and have been criticised in that they will differ between designs when some factors are measured in some designs but not measured in another (e.g., when a covariate is included in some studies, or when a factor that can account for some variance is accounted for such as gender in some analyses but is not included in others). In these cases, the partial variance explained effect sizes will not be comparable with the same value calculated in another study. The variance explained by the covariate or measured factor will be partialled out of the denominator when the covariate is included in the model, but this variance would be included in the error variance when the measured variable is not included in the model (Olejnik & Algina, 2003).

To make this more concreate take the example of a study on the impact of chocolate advertising on the variable “professed enjoyment of chocolate”. If gender could account for 5% of the variance in “professed enjoyment of chocolate”, the for the impact of chocolate would be higher if it is included in the ANOVA model than if it were not. However, the true impact of the advertisement would not have changed, and the effect sizes from these two studies would not be comparable. Similar issues occur in repeated measure or mixed designs. Correlations between individuals’ scores over levels of a factor (e.g., correlations between individuals’ scores over time) reduce the error sums of squares, decreasing the value of the denominator and inflating the for the factor of interest compared to a between subjects design (i.e., when repeated measures are not taken).

(Generalised eta squared) and were developed by Olejnik and Algina (2003) in order to avoid this issue. It is functionally identical to except in that it include any the measured, non-manipulated factors in the denominator.

Equation 5 (Olejnik & Algina, 2003)

sums the sums of squares of all measured factors (the unmanipulated factors that are included in the model, e.g., gender) and sums over all the sums of squares for subjects or covariates (i.e., it plays the role of and additionally includes any variance from covariates included in the model). acts as an indicator variable which takes the value of 0 if the effect if interest involves measured factors (e.g., age or sex or interactions between measured and manipulated factors) or 1 if this is not the case. prevents from being counted twice when the effect is measured not manipulated; first as and then as part of .

Olejnik and Algina also developed a generalised .

Equation 7[[3]](#footnote-3) (Olejnik & Algina, 2003)

*N* is the total number of data points in the analysis, degrees of freedom for the effect under study, is the degrees of freedom for all measured effects. is the error mean square for testing the effect, whereas is the error mean square for testing the effect labelled . is equal to the sum of all plus .

Although in general or may be preferable for comparing effects across studies, it is often impossible to extract this information from published papers, researchers rarely report all of the information necessary to calculate these effect sizes. In fact, many statistical programs do not produce all of the values necessary to calculate these metrics by default (Olejnik & Algina, 2003)

#### f

Although *f* and *f*2 are now are now relatively rarely used, these effect size metrics are worth understanding as they are the metrics in which Cohen defined his benchmark values for ANOVA and regression designs. *f* is equal to the ratio of the standard deviations of the means of groups compared to the standard deviation of all included data, and *f*2 can be interpreted as the variance of the means of each group divided by the variance of all included data.

*f*2can be calculated from eta squared as

Equation 8.2.19, Cohen (1988)

Other estimators can also be used (i.e., ω2 or ) as the basis for *f*2 calculations. Whichever estimator was used, e.g., here , can be calculated from *f*2 as

Equation 8.2.20, from Cohen (1988).

*f* can be calculated from the F statistic produced by an ANOVA as

Appendix A, (Albers & Lakens, 2018).

**Empirical benchmarks for association**

Few efforts to identify empirical benchmarks for measures of association outside of correlation coefficients were found in the current literature survey, possibly because until recently the estimation and publication of effect sizes for association was difficult (these values not being reported by typical statistical software), and relatively difficult to calculate by hand. See table [effect sizes not r or d] for all of the identified attempts to find empirical benchmarks for association in the psychology literature.

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]

Categorical effect sizes:

There are a number of effect sizes for categorical variables, the most common which is probably Odds ratios, and Cohen’s W which is useful in power analysis and more general in that it is not constrained to 2 by 2 contingency tables.

#### W

Cohen (1988, 1977) proposes the effect size measure **w** for chi square tests for tests of frequencies or proportions.

Following Cohen (1988) equation 7.2.1

Where Poi is the null hypothesised proportion in cell i, P1i is the alternative hypothesised proportion in cell i, and m is the total number of cells. This means that w is the sum of the deviation from the null hypotheses standardised by the size of the null hypothesized value. w is beneficial in that it scales to any number of cells, however for 2 by two contingency tables more easily interpretable values are often employed. See Table [effect sizes not r or d] for benchmarks extracted from a small number of articles which used this effect size and which were reported on in social psychology textbooks.   
Odds ratios

Given two by two contingency tables, a commonly employed effect size is the odds ratio, the ratio of the odds of an event occurring in one group (e.g., a treatment group) to the odds of it occurring in other group (e.g., a control group).

Table [Contingency table example]

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Outcome | |
|  |  | Positive | Negative |
| Treatment group | Active | a | b |
|  | Control | c | d |

(J. A. Morris & Gardner, 1988)

Although there is no transformation that converts odds ratio into r or *d* without knowledge of other parameters, odds ratios can be used to approximate product-moment correlations and cohen’s *d* (Bonett, 2007) either accepting additional assumptions about the underlying data or with further knowledge about the data. Odds ratios can be converted to *d* without knowledge of any other parameters or sample statistics under the assumption that the data is representative of a dichotomisation of a logistically distributed variable in each group (Borenstein et al., 2011):

Borenstein et al. (2011) equation 7.1

With ln being the natural logarithm, π being the mathematical constant pi, and OR being the odds ratio. When a researcher has access to the sample size in each group, the Ulrich-Writz approximation can be used (Ulrich & Wirtz, 2004),

Bonett (2007), page 3

With *n1* and *n1* being the sample size from the first and second groups respectively, and *n* being the total sample size.

This can then be used to approximate Cohen’s d,

Bonett (2007) , page 3

More accurate Pearson product moment correlations can be estimated from odds ratios with additional information about the marginal proportions, see Bonnett (2007) for further detail.

### Conclusion and a note of caution

A note of caution is advisable in interpreting the above reported empirical effect size benchmarks. The effect sizes explained above and the empirical benchmarks that were identified do not provide a comprehensive assessment of all effect sizes or 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 in part because of smaller sample sizes and publication bias. It is also noteworthy that the median benchmarks tend to be much lower than the reported mean benchmarks as effect sizes reported in psychology tend to be heavily positively skewed, an important consideration when thinking about what effect sizes should be expected from research. In the same vein, none of the included 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 the reported benchmarks are overestimates (Anderson & Maxwell, 2017; S. E. Maxwell, Lau, & Howard, 2015; Open Science Collaboration, 2015).

In so far as having a description of the distribution of effect sizes in areas of research are useful in guiding researchers intuitions, it is clear that there is a need for wider descriptions of the types of effect sizes that can be seen across areas of psychology research. There are large areas of psychology research that have not been surveyed. However, even this small body of research illustrates the degree of heterogeneity among effect sizes 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.

It’s also important to note the competing interests that are seen in this body of research. The larger the scope of any of these effect size surveys the less useful they are. Ideally, for the purposes of understanding what effect sizes a researcher should expect, they want to have empirical benchmarks for as closely matched a body of research as is possible. However, the more closely matched a body of research the less widely applicable these benchmarks are. There are also major methods issues that should be considered, as almost every paper reports multiple statistical tests, averaging over effects may lead to erroneous inferences, as may just focusing on the main result as reported by the authors, which could conceivably be upwardly biased if the authors choice of main result is influenced by the size of the effects they observed.

This chapter provides the definitions and methods of calculation for the most common standardised effect sizes used in power analysis, and provided empirical effect size benchmarks for these effect size measures where they are available. All of these standardised effect sizes are useful in certain scenarios, and there are numerous estimators and other effect size measures that are not covered above. None of the presented benchmarks should be used as the sole basis for a power analysis, but having an accurate understanding of the types of effect sizes seen in the literature seems essential for having a starting point at which to base an effect size estimate for a power analysis, be that in estimating a minimum effect size of interest or in assessing the results of a sensitivity analysis (see chapter [effect size estimation for PA] for more information on these approaches).

References

Albers, C., & Lakens, D. (2018). When power analyses based on pilot data are biased: Inaccurate effect size estimators and follow-up bias. *Journal of Experimental Social Psychology, 74*, 187-195. doi:<https://doi.org/10.1016/j.jesp.2017.09.004>

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.

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

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2011). *Introduction to Meta-Analysis*. West Sussex, United Kingdom: John Wiley & Sons.

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

Carroll, R. M., & Nordholm, L. A. (1975). Sampling Characteristics of Kelley's ε and Hays' ω. *Educational and psychological measurement, 35*, 541-554. doi:10.1177/001316447503500304

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. (1977). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

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. (2013). The New Statistics. *Psychological Science, 25*, 7-29. doi:10.1177/0956797613504966

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>

Gibbons, R. D., Hedeker, D. R., & Davis, J. M. (1993). Estimation of Effect Size From a Series of Experiments Involving Paired Comparisons. *Journal of Educational Statistics, 18*, 271-279. doi:10.3102/10769986018003271

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

Goulden, C. H. (1939). *Methods of statistical analysis*. New York, USA: John Wiley &. Sons, Inc.

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

Hedges, L. V., & Olkin, I. (1985). Statistical Methods for Meta Analysis. San Diego, CA: Academic Press.

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

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

Kruschke, J. K., & Liddell, T. M. (2017). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*. doi:10.3758/s13423-016-1221-4

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

Levine, T. R., & Hullett, C. R. (2006). Eta Squared, Partial Eta Squared, and Misreporting of Effect Size in Communication Research. *Human Communication Research, 28*, 612-625. doi:10.1111/j.1468-2958.2002.tb00828.x

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

Maher, J. M., Markey, J. C., & Ebert-May, D. (2013). The Other Half of the Story: Effect Size Analysis in Quantitative Research. *CBE Life Sciences Education, 12*, 345-351. doi:10.1187/cbe.13-04-0082

Maxwell, S., Camp, C., & Arvey, R. (1981). Measures of strength of association: A comparative examination. *The Journal of applied psychology, 66*, 525-534.

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

McGrath, R. E., & Meyer, G. J. (2006). When effect sizes disagree: the case of r and d. *Psychological Methods, 11*, 386-401. doi:10.1037/1082-989x.11.4.386

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

Miles, J. (2004). R-Squared, Adjusted R-Squared. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of Statistics in Behavioral Science*.

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychol Methods, 7*, 105-125.

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: measures of effect size for some common research designs. *Psychological Methods, 8*, 434-447. doi:10.1037/1082-989x.8.4.434

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

Pearson, K. (1903). On the influence of natural selection on the variability and correlation of organs. *Philosophical Transactions of the Royal Society of London., 200*, 1.

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

Richardson, J. T. E. (2011). Eta squared and partial eta squared as measures of effect size in educational research. *Educational Research Review, 6*, 135-147. doi:<https://doi.org/10.1016/j.edurev.2010.12.001>

Rosenthal, R. (1991). Meta-Analytic Procedures for Social Research. Thousand Oaks, California: SAGE Publications, Inc. Retrieved from <http://methods.sagepub.com/book/meta-analytic-procedures-for-social-research>. doi:10.4135/9781412984997

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

Theil, H. (1958). *Economic forecasts and policy*: Amsterdam : North-Holland Pub. Co., 1958.

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

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594-604. doi:10.1037/0003-066X.54.8.594

Supplementary material [Cohen’s d/Hedges’ g]

Although the equivalence between formulas x.3 and x.2 is relatively trivial, it seems worth highlighting this equivalence more explicitly as this appears to be a common source of confusion for students and researchers. For example, (Maher et al., 2013) reported that the difference between d and g is that Hedge’s g uses equation x.3 to calculate the pooled standard deviation instead of equation x.2, despite the fact that those formulas are mathematically identical.

equation x.2

However, this simplifies to equation x.2 Both “na – 1” and “nb – 1” in the numerator of the fraction cancel out, as can be seen more clearly when the s is replaced with the formula for calculating the standard deviation in [x.2 expanded].

[x.2 expanded]

Algebraic manipulation then simplifies this formula to equation [Simplified1].

[simplified1]

Multiplying the elements in the numerator out, we get equation [simplified2], which is identical to equation x.2 above.

[simplified2]

1. See supplementary material [Cohen’s d/Hedges’ g] for a demonstration of the equivalence between x.2 and x.3. This is explicitly provided in the supplementary material as this appears to be a common point of confusion among students and researchers (e.g., (Maher, Markey, & Ebert-May, 2013) misidentifies equation x.3 as the equation for Hedge’s *g* and contrasts that with equation x.2 of Cohen’s *d*, although the results of each equation is in fact the same.) [↑](#footnote-ref-1)
2. Following Albers and Lakens (2018) I write the estimators \eta^2, \epsilon^2, or \omega^2 without using the hat notation often used to distinguish between parameter and estimator. [↑](#footnote-ref-2)
3. Both the numerator and denominator are divided by N in Olejnik and Algina, 2003, which has been removed for clarity here. [↑](#footnote-ref-3)