# Chapter 5 – Effect sizes for planning and interpreting research, definitions and empirical benchmarks

## Abstract

This chapter presents a literature review of effect size surveys (studies which have systematically extracted effect sizes from the empirical research literature) in the behavioural sciences research. In order to be able to understand and make use of units-free effect sizes (e.g., Cohen’s *d*, correlation coefficients or ) for the purpose of power analyses, reading, or communicating research, researchers need to know what these effect sizes describe and have a sense of what effect size magnitudes are typically seen in a given area of research. In order to facilitate the interpretation of effect sizes in psychological research, this chapter brings together these previous efforts to provide empirical effect sizes benchmarks for the behavioural sciences research and presents their results alongside common language effect size descriptions, estimators for each effect size, and advice on the appropriate effect sizes and estimators to use when planning research.

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). The magnitude of effects can be expressed in unit-free or unit-dependent effect sizes. Unit-dependent effect sizes (e.g., mean differences) are presented in the units of the measured variable, and may be particularly useful when the units of analysis are directly interpretable (e.g., income, IQ scores, measures of height or weight). Units-free effect sizes (e.g., Cohen’s *d* for mean differences or R2) can be useful for facilitating understanding when the units of measurement are not themselves interpretable (e.g., a newly developed measure), are essential for meta-analysis, and can aide in formal sample size determination. This chapter describes the results of a systematic review used to identify previous efforts to survey the effect sizes reported in various bodies of behavioural sciences research. These results are presented alongside common language descriptions of each effect size type, estimators for the included effect sizes, and advice on the appropriate effect sizes to use when planning research.

In order to be able to understand and make use of standardised effect sizes for power analysis, researchers need to have some understanding of the details of how they are calculated, know which quantities effect sizes describe, and have a sense of the effect sizes are typically seen in a given area of research. There are many texts which provide an outline of the mathematical details (e.g., Lakens 2013, Cohen 1988), but relatively few which have attempted to give guidance on what could reasonably be classified as a small or a large effect. Part of the reason for the relative scarcity of advice on how to interpret effect sizes is that that the meaning and importance of a given standardised effect size is highly context dependent. If someone is studying a treatment for a common and deadly disease, an effect of Cohen’s *d* of .1 may represent an effect that could save hundreds of thousands of lives. If someone is studying, for example, self-reported agreeableness, it is unlikely that an intervention that has an effect of .1 Cohen’s *d* would be pursued further. For this reason, attempting to provide universally applicable benchmarks on what a “small”, “medium” or “large” is foolhardy if not impossible.

**“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)**

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 a | 1.49 | 3.45 | 9 |
| *F* | .1 | .25 | .4 |
| *f 2* | .02 | .15 | .35 |
| b | .0099 | .0588 | .1379 |
| R2 | .02 | .13 | .26 |

Notes: 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 have remained the same. a Converted from Cohen’s benchmarks for *w.* b Transformed from Cohen’s benchmarks for *f*.

Despite the issues inherent in suggesting that any particular effect size benchmarks should be widely employed, there is a clear need for effect size guidance aimed at practicing researchers and research students. The consumers and producers of research that is often reported in standardised effect sizes need to be able to understand what effects are typically seen in their area of research. Having an accurate sense of the magnitudes of the effects seen in an area of research is helpful not just in allowing researchers interpret the relative importance of observed effects, but also in allowing them to effectively plan future research.

In so far as current standards for classifying the importance and relative magnitude of observed effects people have largely relied upon the standardised effect size benchmarks given by Cohen (1962, 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). In part in order to prevent researchers from relying on Cohen’s benchmarks in interpreting reported effect sizes and in planning their research, a number of papers have been published providing empirical benchmarks, benchmarks developed by examining bodies of research and extracting reported effect sizes. This chapter collects these studies and presents them alongside common language explanations and estimators for each of the included each effect size metrics and a brief discussion of how widely applicable these benchmark values are.

### Methods Review protocol

In order to identify articles which have extracted effect size benchmarks from bodies of literature a snowballing sampling method was used. This sampling method used targeted literature searches to identify an initial sample, and then all articles which either cite or are cited by the original article were screened to achieve good coverage of this small research literature. 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 and citing articles (identified using Web of Science) of all articles including identified an additional three articles. A further search of these three articles’ citing and cited articles did not identify any further applicable article. Two articles outlining effect size benchmarks were also included from the grey literature, a pre-print (Lovakov & Agadullina, 2017) and a book (Hattie, 2009). After deduplication and full text screening, a total of 15 records were identified that provided empirical effect size benchmarks for fields of research. Database searches were performed on the 11th August, 2018.

### Analysis and data presentation

This chapter presents the results of this literature search grouped into three categories; effect sizes for mean differences (Cohen’s d, and Hedge’s *g*), categorical effect sizes (Cohen’s w), and variance explained effect sizes (r, R2, *η2*,ε2 and ω2). No aggregation is performed as a part of this paper for two reasons. Firstly, as aggregation of these values would lead to the loss of these studies main value; providing an indication of the distribution of effect sizes in specific sub-populations. Secondly, the sampling strategies in the examined articles are as varied as selecting effects from textbooks to effects reported in meta-analyses of clinical studies. This means that aggregating these efforts would produce estimates that are unlikely to describe any identifiable population. The sampling strategy used by each included effect size survey is identified alongside each reported result and a brief discussion of which populations the sample is likely to be representative of.

This chapter presents the empirical benchmarks alongside the estimators for each effect size, and a common language description of the estimated quantity. In the tables below, “NA” reflects a cell for which there is no applicable response (e.g., “number of meta-analyses included” when effects were not extracted from meta-analyses) and “-” indicates that a value was not reported in the effect size survey.

## Results

### Effect sizes for Mean differences

A number of projects have extracted empirical effect size benchmarks from various fields of research for Cohen’s *d* for independent groups. 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. None of the identified studies report effect size benchmarks for mean differences in repeated measures designs (also called Cohen’s *d*, supplementary materials [d] for a detailed description of this estimator). For a visual depiction of the proportion overlap at each of Cohen’s benchmarks, see Figure *[Cohen’s d as population distributions]* calculated assuming equal variances in each group and that the populations are normally distributed.



*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, calculated using methods from Reiser and Faraggi (1999).

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 deviation. In other words, Cohen’s *d* describes the size of the difference between two groups divided by a pooled measure of the variability among individuals in each group. 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.

Whereis the mean of sample 1, and is the mean of sample 2, and is the pooled standard deviation (adapted from McGrath & Meyer, 2006, p. 386). The pooled standard deviation is most often calculated for samples as

following Cohen (1977, p. 67). Equivalently, this can be calculated as

following Hedges (1981, p. 110) [[1]](#footnote-1). Here, is the sample variance for each group, calculated as

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

This estimator for Cohen’s d is consistent (that is, as the sample size 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

where for an independent groups design, d is calculated as per above and is the gamma function (originally outlined in Hedges, 1981; this version is adapted from Hedges & Olkin, 1985, p. 104). This correction factor is fairly complex (although computationally trivial on modern computers), and Hedges provides a simple approximation which performs well enough for most practical purposes (Hedges, 1981, p. 114). Hedge’s approximate bias corrected *g*\* is calculated as

where for an independent groups design and *d* is Cohen’s d as calculated as above (this version adapted from Borenstein, Hedges, Higgins, & Rothstein, 2011, p. 27; originally from Hedges, 1981).

Confusingly, in the literature ‘Hedge's g’ or ‘Cohen's d’ are often used interchangeably to refer to , and (Lakens, 2013). For most practical purposes , and are virtually identical when *n* ⪆ 30, and all are estimators of the same population parameter. Simple sampling variability and selective reporting are likely to cause greater difficulties in determining effect sizes for power analysis than the bias of the estimator that has been used, although Hedges’ *g* is the most preferred for the purposes of power analysis as it is unbiased. See supplementary materials [conversions] for methods of extracting these measures from reported test statistics.

This effect size has had the largest number of empirical effect size surveys. The median reported Cohen’s d values tend to be around Cohen’s “medium” estimate (.5), with median values varying from .38 (extracted from 3498 effects in meta-analyses published in 42 meta-analyses published in 29 journals in the "Psychology, Social" category of Social Sciences Citation Index ) to 0.654 (extracted from some 26841 effects extracted from *t* tests in cognitive neuroscience articles published in high impact journals from 2011 to 2014). The mean Cohen’s *d* values in the non-educational interventions tend to be much higher than Cohen’s estimated “medium” benchmarks, and the extracted effect sizes are highly skewed (Szucs & Ioannidis, 2017). In those studies that examined only educational interventions, the mean values appear to be somewhat smaller, ranging from a minimum of .23 (examining meta-analytic outcomes of elementary school intervention studies; Hill, Bloom, Black, & Lipsey, 2008) to .51 (examining the results of randomised controlled trials of educational interventions performed in middle schools; Hill et al., 2008), see Table [education].

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 Randomised controlled trials (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 studies a | Meta-analytic effect size estimates | 32 | 0.23 | 0.21 |
| Hill et al., (2008) | Meta-analyses of middle school intervention studies a | Meta-analytic effect size estimates | 27 | 0.27 | 0.24 |
| Hill et al., (2008) | Meta-analyses of high school intervention studies a | Meta-analytic effect size estimate | 28 | 0.24 | .15 |
| Hattie (2009) | Meta-analyses of educational interventions b | Effect sizes | 146,626 | 0.4 | *NA* |

Note: a Hill et al., did not report the total number of meta-analyses or effects included in their examination of effects from meta-analyses, but specified that sourced the included studies from Bloom, Hill, Black, and Lipsey (2007) and M. Lipsey, Bloom, Hill, and Rebeck Black (2007). b Hattie 2009 included 816 meta-analyses, including a total of 52,649 articles, extracted in a non-systematic way.

Table [effect sizes d psychology]. Results of effect size surveys reporting Cohen’s *d* and examining psychology research, blank cells were not reported.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 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) | Heart 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 the Meta-lab project (http://metalab.stanford.edu) | NA | 12 | NA | - | - | - | 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 |

## Categorical effect sizes:

Only a single study was identified that extracted categorical effect sizes from the behavioural sciences literature (Cooper & Findley, 1982b). This study extracted an effect sizes in a unit that is rarely used outside of the context of power analyses; Cohen’s *.* This effect size was proposed by Cohen (1988, 1977) for chi square tests for tests of frequencies or proportions and describes the degree to which the observed relative frequencies deviate from the null hypothesised relative frequency.

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 useful in that it scales to any number of cells, however for 2 by 2 contingency tables more easily interpretable effect sizes are often used (e.g., odds ratios, see supplementary materials 2 for a brief description of this effect size). Our literature review only identified a single study that examined reported effect sizes for contingency table analyses. Cooper and Findley (1982) examined effects reported in social psychology textbooks found a mean of .26 (SD = 0.16), close to Cohen’s “medium” benchmark (.3). However, as this study only examined 15 effect sizes from 15 studies referenced in Social Psychology textbooks, little weight should be placed on this value.

## Effect sizes for association/variance explained:

One of the oldest standardised effect sizes commonly used today, the Pearson correlation coefficient, *r*, measures the degree of linear association between two variables (Pearson, 1903).

Where *x* are the values of x, y are the values of y, and n is the number of pairs of scores. See Supplementary material [conversions] for methods of converting *t*, F and χ2 statistics to a correlation coefficient. A number of papers have extracted Pearson correlations 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. The average values vary considerably again, with the maximum coming again from effects reported in social psychology textbooks (mean = .48, Cooper & Findley, 1982b), and the smallest seen in meta-analyses of social psychology (mean = .21, median = .18, Richard, Bond Jr, & Stokes-Zoota, 2003) and in effects reported in the first correlation table of articles published in the Journal of Applied Psychology and Personnel Psychology from 1980 to 2010 which found a mean of .32 and a median of .16 (Bosco, Aguinis, Singh, Field, & Pierce, 2015). The only sample for which the mean or median was considerably higher than Cohen’s suggested “medium” effect (.3), is Cooper and Findley (1982) who only examined just 23 articles described in social psychology textbooks and found a mean effect size of .48, a value which is unlikely to be representative of social psychology overall.

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) | 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 |

## Multivariate variance explained effect sizes

#### f

Although *f* and *f*2 are now are relatively rarely used outside of sample size determination, these effect size metrics are worth understanding as they are the metrics in which Cohen defined his benchmark values for ANOVA and regression designs, as well as being are the effect sizes requested by most power analysis software (e.g., Faul, Erdfelder, Lang, & Buchner, 2007). *f* is equal to the ratio of the standard deviation of the means of groups compared to the standard deviation of all data. *f*2 is equal to the variance of the means of each group divided by the variance of all included data.

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

with being the degrees of freedom for the effect, the error degrees of freedom and the observed F statistic, Appendix A, (Albers & Lakens, 2018).

#### η2

η2 describes the proportion of variance attributable to an effect standardised by the total variance across the sample. It can be calculated as

where *SSeffect* is The sums of squares between groups, , *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, . η2 summarises the variance explained by one factor within an ANOVA design.

For the purposes of power analysis, it is important to understand two other common effect sizes for variance explained; (partial eta squared) and (Generalised eta squared). partial eta describes the proportion of variance that can be attributed to a particular factor after excluding variance explained by other factors in the model. can be calculated as

where is the sum of squared residuals, , with being the predicted value of i or equivalently the mean of the group under study (Levine & Hullett, 2006; equation 2).

and are equal in one-way ANOVAs as all summed and squared errors are included in the error term, but in multiway or repeated measures ANOVA partial eta squared will be larger as variance explained by the additional factors is not included in the denominator. can be calculated from reported F statistics following Richardson (2011),

with being the degrees of freedom for the effect, the error degrees of freedom and the observed F statistic.

As can be seen in its direct transformation from the F statistic and degrees of freedom, this statistic aligns directly with the significance test of a single factor in a typical multi-way ANOVA design. However, has been criticised in that it will lead to different apparent effect sizes when some factors are measured in some designs but not measured in another (e.g., when a naturally occurring covariate that can account for some variance such as gender is included in some studies but is not included in others). In these cases, an effect sizes will not be comparable with the calculated in another study, as variance explained by the covariate or measured factor will be partialled out of the denominator when the measured variable is included in the model, but included in the error variance when the measured variable is not included in the model (Olejnik & Algina, 2003).

(Generalised eta squared) was developed by Olejnik and Algina (2003) in order to avoid this issue. It is similar to except in that it includes the variance associated with any measured, non-manipulated factors in the denominator. It is calculated as

where 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 but additionally additionally includes any variance from covariates included in the model) (Olejnik & Algina, 2003, equation 5). 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 .

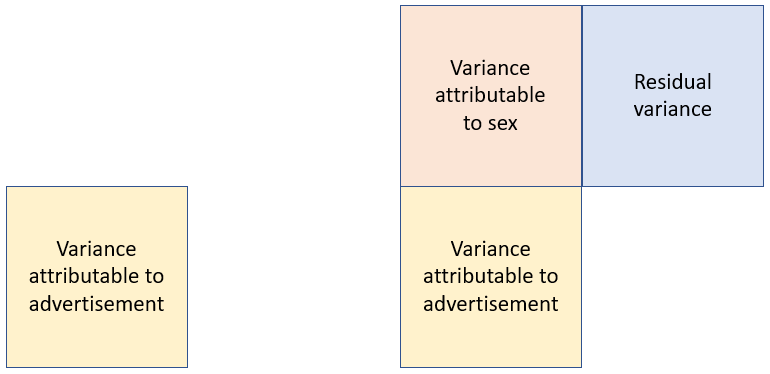
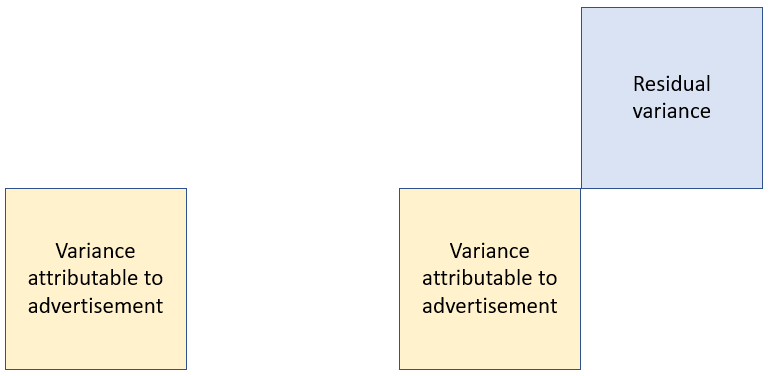
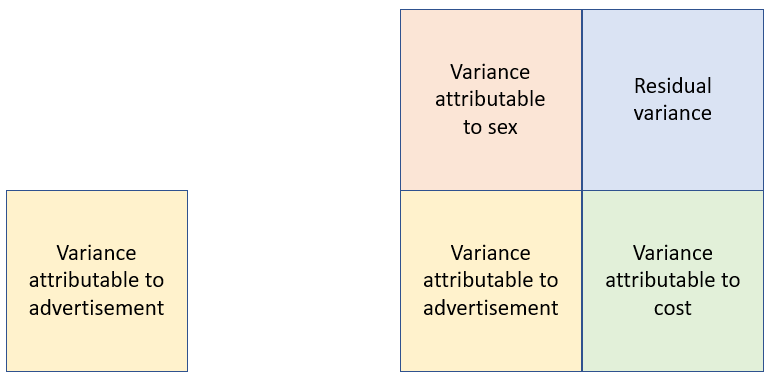
### Deciding between , and

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), all effect sizes are useful in different scenarios. describes the variance attributable to a factor as a proportion of the total variance, describes the proportion of variance attributable to a particular factor after removing variance attributable to other factors in the model, and describes the proportion of variance explained after excluding variance attributable to all manipulated variables but including all non-manipulated factors. The most appropriate effect size to use will be determined by the goals of the researcher and their experimental design.

To make this more concreate we can take the example of an advertising company’s experiment examining the impact of an ad for chocolate and cost on “professed enjoyment of chocolate”. Saying that they used a two by two design, manipulating both price and whether people viewed an ad for chocolate or not. The researchers then run an ANOVA including sex as a factor in their analysis, finding that variance happens to be exactly equally partitioned among the sex, advertisement, cost and residual (see figure [x] for a geometric depiction of this example).

Each version of tells us something different, although all provide interesting and useful information. expresses the amount of variance attributable to the advertisement as compared to the total variability. expresses the impact of the advertisement as a fraction of residual variance, i.e., excluding the variance explained by both sex and cost. expresses the amount of variance explained as a proportion of the total variance excluding any variance attributable to cost, but including variance attributable to sex. This may be appropriate as cost is manipulated factor within the experimenters control and may not be included as a manipulation in other experiments, whereas sex causes variance in non-experimental and experimental settings.

Although the generalised measure () may be preferable for comparing effects across studies, and for planning studies when a previously performed manipulation will not be included, it is often impossible to extract the information required to calculate this value from published papers. Also a consideration when planning research is the fact that , and are upwardly biased (Olejnik & Algina, 2003). Two other estimators have been proposed along with their partial and generalised equivalents,ε2 (Epsilon squared) and ω2 (omega squared).



Generalised eta squared

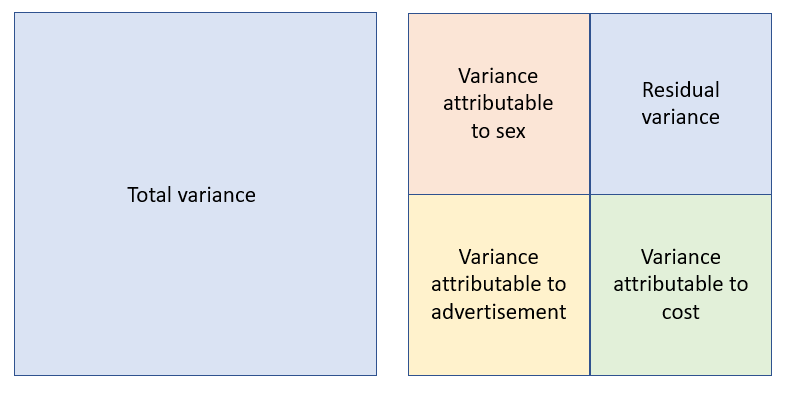
Partial eta squared

Eta squared

/

/

/



Variance partitioning

1

.25

.25

.25

.25

.25

.25

.25

.25

.25

.25

.25

.25

.25

.25

.25

.25

Figure [x]. A geometric explanation of the calculation of eta squared, partial eta squared and generalised eta squared.

#### Epsilon squared ε2

Epsilon squared (usually called adjusted R**2** in the regression context), adjusts the effect size downwards as the number of factors gets larger and as the sample size decreases in order to account ’s upward bias. can be calculated as

following equation 4 (Carroll & Nordholm, 1975), or equivalently as

following equation 7 Carroll and Nordholm (1975). Here, 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.

(Albers & Lakens, 2018, appendix A).

#### Omega squared ω2

A similar less-biased-than estimator is ω2 (omega squared) which is calculated as

where is the sum of squares for the effect, is total sum of squares, and is the error mean squares (Carroll & Nordholm, 1975, equation 6).

There is also a partial version of the omega estimator, , 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

following S. Maxwell, Camp, and Arvey (1981) equation 26.

can also be calculated from reported F statistics using

(Albers & Lakens, 2018, appendix A).

Olejnik and Algina also developed a generalised :

where *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 . This version is adapted from Equation 7[[2]](#footnote-2) of Olejnik and Algina (2003).

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.

Because power analysis software often requires the effect size of interest to be specified in terms of *f*2 or *f*, it’s important to know that any of the above statistics can be converted to using the following formula ( can be replaced with any of the above estimators as appropriate (Cohen, 1988, equation 8.2.19),

Although both and are less biased than , they do have a higher level of variance (i.e., they have a higher mean absolute error but they do not tend to systematically under-estimate the population effect size; Levine & Hullett, 2006), Albers and Lakens (2018) argue that it is preferable to use and for power analysis as simulations show that they lead to better powered studies on average, as they are not upwardly biased, although again sampling variability and publication bias are likely to have more impact on the effect size than the estimator that is used.

The appropriate version of the and type effect sizes to use for planning sample sizes will depend on the research design, as in power analysis you need to choose the effect size that excludes variance so as to reflect your proposed study. In a one-way ANOVA the full, partial and generalied estimators will be equivalent. If a two-way ANOVA is planned and the study is an exact replication, it would be appropriate to use or so as to estimate the proportion of variance appropriate to the statistical test of the impact of the factor of interest. If, on the other hand, a replication is going to be performed without including a variance-causing factor that will nonetheless cause variance (e.g., a replication study where a measured variable such as sex was included in the initial analysis but will not be measured in the replication), the most appropriate effect size for sample size planning will be , the effect size which includes the variance from the unmeasured factor in the denominator.

## Benchmarks

Few efforts to identify empirical benchmarks for measures of variance explained (beyond of correlation coefficients) were found in the current literature survey, possibly because the estimation of these effect sizes is relatively difficult (these values not being reported by typical statistical software and being somewhat difficult to calculate by hand), and rarely reported in the primary research literature (Cumming, Fidler, Kalinowski, & Lai, 2012; Cumming et al., 2007; Fidler et al., 2005). See table [effect sizes not r or d] for all of the identified attempts to find empirical benchmarks for variance explained in the psychology literature. The identified effect size benchmarks for these effect sizes show higher mean values than Cohen estimated with his “medium” or even “large” effect size benchmarks (where Cohen suggested a medium *f* of .25, equivalent to an of .059), although this may be explained by the limited sampling frames for these studies (with the *f* value benchmarks coming from effects reported in social psychology textbooks and the benchmarks coming from univariate statistical tests reported in Journal of Counselling Psychology from 1970 to 1979) .

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) |

### Discussion and conclusion

This chapter provides the definitions and methods of calculation for the most common standardised effect sizes used in power analysis, along with empirical effect size benchmarks for these effect size measures where 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 knowledge of the magnitudes of the effect sizes seen in the literature seems like an essential starting point from which to base effect size estimates for power analysis, be that in estimating a minimum effect size of interest, showing that a study is likely to be adequately power in a grant proposal, or in assessing the results of a sensitivity analysis (see chapter [effect size estimation for PA] for more information on these approaches).

There are large areas of psychology research that have not been surveyed at all, leaving a gap that future research could begin to address. In so far as having a description of the distribution of effect sizes in areas of research is useful in guiding researchers’ intuitions as well as for research planning and educational purposes, it is clear that there is a need for surveys of the magnitudes of effect sizes that are seen across areas of psychology research more broadly. One potentially useful advance would be the development of a searchable databases of effect sizes across areas of psychology, an endeavour that is becoming more tractable given advances in large-scale text mining capabilities and the possibility of automating some of the necessary screening practices using tools like natural language processing.

However, even the small body of existing 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 interventions in clinical psychology (M. W. 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 is also noteworthy that the median observed effect sizes 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.

A note of caution is advisable in interpreting empirical effect size benchmarks from the literature. 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. In fact, none of the included articles attempt to address the issue of publication bias increasing the average effect sizes in the published literature above that which should be expected when planning a study. Given the large differences 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). See chapter [effect size adjustment] for an extended examination of the degree to which effect sizes in even direct replications are likely to be much smaller than reported effect sizes.

**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

Bloom, H., Hill, C., Black, A., & Lipsey, M. (2007). *Using empirical benchmarks for interpreting effect size.* Paper presented at the Presentation to the Interagency Roundtable Meeting on “The Application of Effect Sizes in Research on Children and Families: Understanding Impacts on Academic, Emotional, Behavioral, and Economic Outcomes.

Bonett, D. (2007). *Transforming Odds Ratios Into Correlations for Meta-Analytic Research* (Vol. 62).

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

Bosco, F. A., Aguinis, H., Singh, K., Field, J. G., & Pierce, C. A. (2015). Correlational effect size benchmarks. *Journal of Applied Psychology, 100*, 431-449. doi:10.1037/a0038047

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. (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. (1982a). Expected Effect Sizes. *Personality and Social Psychology Bulletin, 8*, 168-173. doi:10.1177/014616728281026

Cooper, H., & Findley, M. (1982b). Expected Effect Sizes: Estimates for Statistical Power Analysis in Social Psychology. *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., Kalinowski, P., & Lai, J. (2012). The statistical recommendations of the American Psychological Association Publication Manual: Effect sizes, confidence intervals, and meta-analysis. *Australian Journal of Psychology, 64*, 138-146. doi:10.1111/j.1742-9536.2011.00037.x

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

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G\*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175-191. doi:10.3758/bf03193146

Fidler, F., Cumming, G., Thomason, N., Pannuzzo, D., Smith, J., Fyffe, P., . . . Schmitt, R. (2005). Toward Improved Statistical Reporting in the Journal of Consulting and Clinical Psychology. *Journal of Consulting and Clinical Psychology, 73*, 136-143. doi:10.1037/0022-006X.73.1.136

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

Haase, R. F., Waechter, D. M., & Solomon, G. S. (1982). How significant is a significant difference? Average effect size of research in counseling psychology. *Journal of Counseling Psychology, 29*, 58-65. doi:10.1037/0022-0167.29.1.58

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.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical Benchmarks for Interpreting Effect Sizes in Research. *Child Development Perspectives, 2*, 172-177. doi:10.1111/j.1750-8606.2008.00061.x

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., Bloom, H., Hill, C., & Rebeck Black, A. (2007). How big is big enough? Achievement effect sizes in education. *University of Pennsylvania Graduate School of Education*.

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

Morris, J. A., & Gardner, M. J. (1988). Statistics in Medicine: Calculating confidence intervals for relative risks (odds ratios) and standardised ratios and rates. *British Medical Journal (Clinical research ed.), 296*, 1313-1316. doi:10.1136/bmj.296.6632.1313

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

Richard, F. D., Bond Jr, C. F., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology, 7*, 331-363. doi:10.1037/1089-2680.7.4.331

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

Szucs, D., & Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLOS Biology, 15*, e2000797. doi:10.1371/journal.pbio.2000797

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

Ulrich, R., & Wirtz, M. (2004). *On the correlation of a naturally and an artificially dichotomized variable* (Vol. 57).

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 the two formulations of the estimator for the pooled standard deviation is simply a matter of algebraic manipulation, it seems worth demonstrating this equivalence 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 the pooled standard deviation for Hedge’s g is calculated as

as opposed to it being calculated as

However, these formulations are mathematically identical, as becomes clearer if we replace the SDs in the first equation above with the formula for the sample standard deviation.

This then simplifies to

Where “na – 1” and “nb – 1” in the numerator of the fraction cancel out. Simple algebraic manipulation then simplifies this formula to

Which is identical to the second formulation of the estimator for the pooled standard deviation in this Supplementary Materials.

### Supplementary material [conversions]

#### Effect size conversions for mean differences

Cohen’s d can be calculated using the results of an independent samples t tests using the formula

(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 the following equation can be used, although it will be an underestimate if the groups are unequal. However, even if the ratio of samples sizes in each group is as extreme as 7 to 3 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.*

#### Conversions to variance explained measures

r can also be converted from *t*, F, chi square and Z statistics following the Open Science Collaboration (2015) using the following equations, these conversions are used in chapter [effect sizes over time] and chapter [estimating publication bias].

*t* statistics can be transformed using

Where is the observed t statistic and df is the degrees of freedom of the t test.

F statistics were converted using:

Where is the observed F statistic and is the degrees of freedom of the numerator andis the degrees of freedom of the denominator.

Finally, if the only information available is the chi square statistic, this can also be converted into correlation coefficients following (Open Science Collaboration, 2015):

Where is the observed statistic and is the associated degrees of freedom. However, it is important to note that there is no standard method of estimating the standard error of this value.

It is important to note that there is no simple way of estimating valid sampling variances or standard errors for Chi square statistics or for F statistics when the degrees of freedom for the denominator is greater than one after conversion using the transformations above.

#### Conversions for categorical variables

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 point-biserial 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

where r is estimated as above following (Ulrich & Wirtz, 2004), and with the, and where the sign of r is used to assign the sign of *d*. 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.

## Supplementary material effect size descriptions

#### 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, and following Cohen (1977, 1988) I will refer to the repeated measures version as Cohen’s . Cohen’s describes the mean magnitude of the change between measurement occasions (i.e., the mean difference score) standardised by standard deviation of the difference scores. 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 (equation 6, Lakens, 2013)

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

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 (Cohen, 1988, p. 48). The latter equation for highlights an important fact about Cohen’s , that the effect size is dependent upon the correlation between scores on repeated measures. The higher the correlation, the greater the . 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 (S. B. 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 Lakens (2013) Equation 7:

where is the *t* statistic calculated as per a repeated measures *t* test, and *n* is the sample size (i.e., the number of repeated measures taken).

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 (from Gibbons, Hedeker, & Davis, 1993, p. 274):

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

Odds ratios

Given a 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]. Illustration of the calculation of an odds ratio.

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

(J. A. Morris & Gardner, 1988)

When the odds ratio is greater than one, there is a positive association between variables, e.g., that those in the active treatment group had a higher probability of getting a positive outcome. Odds ratios below one indicate the opposite, e.g., that those in the active treatment group had a higher probability of being in another group.

1. See supplementary material [Cohen’s d/Hedges’ g] for a demonstration of the equivalence of these equations. 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, and Ebert-May (2013) misidentifies the latter equation as the equation for Hedge’s *g* and contrasts that with the former, calling it Cohen’s *d*, although they are in fact equivalent.) [↑](#footnote-ref-1)
2. Both the numerator and denominator are divided by N in Olejnik and Algina, 2003, which has been removed for clarity here. [↑](#footnote-ref-2)