**Details of effect sizes used in this dissertation**

**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. Effect sizes are also useful in meta-analysis, allowing for studies’ effects to be compared and collapsed. Standardised effect sizes are useful in power analysis as they are a succinct way of proving a great deal of information about the alternative hypothesis.

In order to perform formal sample size determination, researchers must specify an alternative hypothesis in sufficient detail to determine the sampling distribution of the test statistic under the 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]. This chapter outlines the three main types of effect sizes, outlines some commonly maligned benchmarks that have been proposed for power analysis.

This chapter provides the definitions of the different standardised effect sizes that are often used in power analysis, using the notation and terminology that is followed in this dissertation. 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, hazard ratios) as well as how to extract them from test statistics when possible.

**Effect sizes for Mean differences**

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 computationally complex (although trivial on modern computers), so Hedges also provided 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)

People commonly refer to , and as Hedge's g or Cohen's d interchangeably (Lakens, 2013). They are all virtually identical for most practical purposes when *n* > 30, and all can be interpreted in the same way. For the purposes of power analysis, it is important to realise that Cohen’s d is upwardly biased if estimating a based on a literature that uses the biased estimator, and increasingly so as the sample sizes in the literature become smaller. Practically, for the purposes of power analysis, sampling variability and selective reporting are likely to create greater difficulties than the estimator that has been used.

**Summary statistics conversion for two group scenarios**

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)

**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 . 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.

**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 variables that can be predicted through its linear association with another variable. Multiple R2 will be familiar to people from a regression context, describes this same property 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 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 equivalent 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)

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.

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]

**{Maxwell, 1981 #943} 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**

, 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. Correlation 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, 2003 #933}

**Cohen 1988 uses the Cohen’s F as the main effect size for ANOVA and regression designs. The F ratio**

Cohen defined effect sizes measure for ANOVA (*f*) and regression (*f*2) which, although they were the metrics in which Cohen’s benchmarks for regression or ANOVA designs were first proposed {Cohen, 1977 #789}, are now rarely used. *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 is the variance of the means of each group divided by the variance of of all included data.

*f*2can be calculated from eta squared as

Equation 8.2.19, Cohen (1988) and eta squared can be calculated from *f*2 as

following equation 8.2.20, from Cohen (1988).

It can also be calculated from the F statistic from an ANOVA as

Appendix A, (Albers & Lakens, 2018)

**Categorical 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>

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

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.

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

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

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.

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

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

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

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

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.

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

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

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*, despite their equality.) [↑](#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)