**Understanding Effect Sizes for use in Statistical Power Analysis**

**x.1 Effect size measures in power anlaysis**

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 {Kruschke, 2017 #105;e.g.`, \Cumming, 2013 #158;Wilkinson, 1999 #566;Hedges, 1981 #786}. 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 direct 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.

Standardised effect size measures are also useful in power analysis and meta-analysis. In meta-analysis, the standardised effect sizes are typically the main unit of analysis and allow for a set of studies to be collapsed and assessed together. Effect sizes are useful in power analysis as the analyst needs to provide enough information to adequately specify 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, correlational analysis) the specification of a single standardised effect size characterises the sampling distribution under the alternative hypothesis (given that the assumptions of the statistical testing procedure used hold) adequately for power analysis {Cohen, 1988 #562}. 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 resreachers is that they have trouble developing appropriate effect sizes for use in [cite interviews and survey]. Given that the effect size must be chosen appropriately in order for the power analysis to provide meaningful results,

This chapter details the most common effect size measures used in power analysis, highlights previously proposed benchmarks and the criticism of their use in sample size planning, and performs a systematic review of previous studies which have attempted to provide empirical assessments of bodies of literature in order to show the average effect sizes in different areas of research. Importantly, in most scenarios standardised effect size benchmarks should not be used as the sole basis for a power analysis, as the expected and important effect sizes will differ greatly from context to context. Towards this end, this chapter attempts to provide an intuitive guide to understanding effect sizes in order to facilitate researchers being able to estimate the effect sizes for use in sample size planning. Finally, it briefly discusses one approach to proposing a default variance covariance matrix for use in power analysis in more complex ANOVA designs.

**x.1 Standardized effect size typology**

Most standardized effect sizes can be grouped into several different categories. There are effect sizes for group differences, in either between or within subjects designs. Categorical:

Association - ANOVA / REGRESSION / Correlations

**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 $\delta$, 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.

C:\Users\fsingletonthorn\Downloads\CodeCogsEqn (1).gif x.1

$d = \frac{\bar{x}\_1 - \bar{x}\_2}{s\_p}$

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

Where $\bar{x}\_1$ is the mean of sample 1, and $\bar{x}\_2$ is the mean of sample 2, and $s$ is the pooled standard deviation. The pooled standard deviation is most often calculated for samples as:

https://latex.codecogs.com/gif.latex?%24%24s%20%3D%20%5Csqrt%7B%5Cfrac%7B%5Csum%28x_1%20-%20%5Cbar%7Bx%7D_1%29%5E2%20&plus;%20%28x_2%20-%20%5Cbar%7Bx%7D_2%29%5E2%7D%7Bn_1%20&plus;%20n_2%20-%202%7D%7D%24%24 x.2

$$s\_p = \sqrt{\frac{\sum(x\_1 - \bar{x}\_1)^2 + (x\_2 - \bar{x}\_2)^2}{n\_1 + n\_2 - 2}}$$

(Cohen, 1977, p. 67)

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

(x.3)

$$s = \sqrt{frac{(n\_1-1)s\_1^2 + (n\_2-1)s\_2^2}{n\_1 + n\_2 - 2}}$$

(adapted from Hedges, 1981, p. 110)

Where $s\_j^2$ is the sample variance for each group, calculated as

https://latex.codecogs.com/gif.latex?s%5E2_j%20%3D%20%5Cfrac%7B1%7D%7Bn_j-1%7D%20%5Cdisplaystyle%5Csum_%7Bi%3D1%7D%5E%7Bn%7D%20%28x_%7Bj%2Ci%7D%20-%20%5Cbar%7Bx%7D_j%29%5E2 (x.4)

$$s^2\_j\ =\ \frac{1}{n\_j-1}\ \displaystyle\sum\_{i=1}^{n}\ (x\_{j,i}\ -\ \bar{x}\_j)^2$$

The j subscript indicating the group.

The pooled standard deviation (i.e., “s”) should be calculated for populations (i.e., if the entire population has been sampled) 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, I 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)

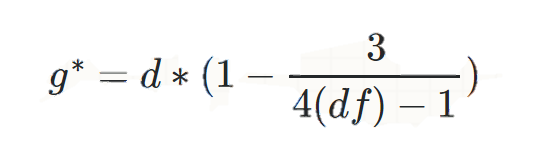
$$g = d \* (\frac{\Gamma(df/2)}{\sqrt{df/2 \,}\,\Gamma((df-1)/2)})$$

Where $df=n1+n2−2$ for an independent groups design, d is calculated as per equation x.1 and $\Gamma(x)$ is the gamma function.

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

However, this correction factor is fairly computationally complex (although trivial on modern computers), so Hedges also provided a computationally simple approximation which performs extremely well in most scenarios. The expectation of $g^\*$ is accurate to within .00033 when the degrees of freedom are greater than 10, and has a maximum error of .007 when there are 2 degrees of freedom (Hedges, 1981) Hedges (1981, p. 114).

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

 (x.6)

$$g^\* = d\*(1 - \frac{3}{4(df)-1})$$

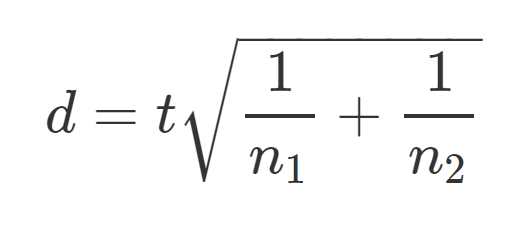
Where $df=n1+n2−2$ for an independent groups design and *d* is Cohen’s d as calculated in equation x.1 using x.2 as the estimator for the variance.

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

People commonly refer to $d$, $g$ and $g^\*$ 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 $\Delta$ based on a literature that uses the biased estimator. Practically, for the purposes of developing summary effect size estimates, sampling variability and selective reporting are likely to create greater difficulties than the estimator that has been used in papers.

**Summary statistics conversion for two group scenarios**

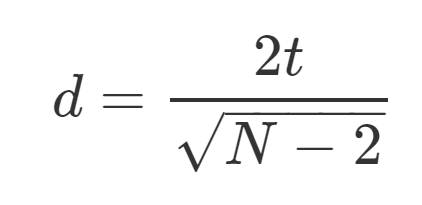
If effect sizes have not been reported, Cohen’s d can be calculated using the results of an independent samples t tests using the formula



$$d = t\sqrt{\frac{1}{n\_1}+\frac{1}{n\_2}} $$

(Lakens, 2013, equation 2)

Where $n\_1$ and $n\_2$ are the sample sizes for groups 1 and two respectively, and *t* is the result of an independent samples t test.



$$d=\frac{2t}{\sqrt{N - 2}}$$

(Rosenthal, 1991, p. 17)

Which is correct if the groups are equal, and will be an underestimate if the groups are unequal. However, although even if the ratio of samples sizes in each group is as extreme as 70/30 the underestimation will be less no more than 8% (Rosenthal, 1991).

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

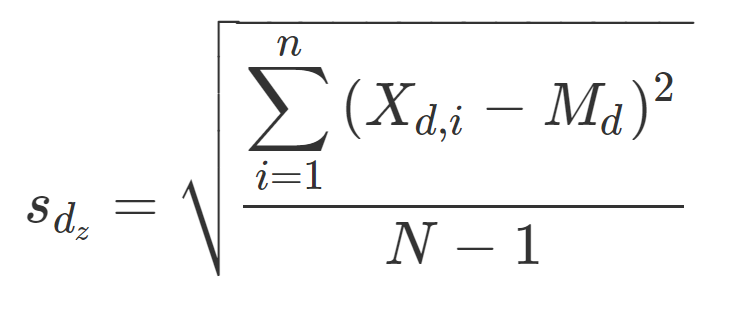
The most common effect size measure for mean difference between repeated measures is also commonly called Cohen’s d, although following Cohen (1977, 1988) I will distinguish between this and the independent samples Cohen’s d by referring to the repeated measures version as Cohen’s $d\_z$. This effect size follows a similar general form to the independent samples Cohen’s d (x.1), except the denominator is the mean difference between measures,

http://www.sciweavers.org/upload/Tex2Img_1522718895/render.png (x.7)

$$d\_z = \frac{M\_d}{s\_d}$$

(Lakens, 2013) equation 6

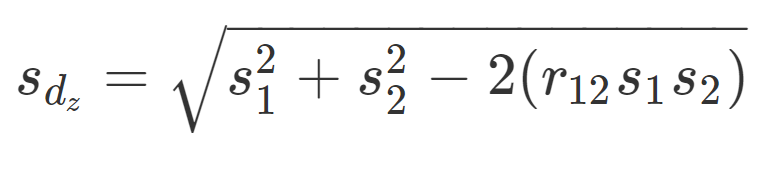
Where $M\_d$ is the mean difference score, and $s\_d$ is the standard deviation of the difference scores calculated as:

 (x.8)

$s\_{d\_z} =\sqrt{\frac{\displaystyle\sum\_{i=1}^{n}{(X\_{d,i} - M\_d)^2}}{N-1}}$

Where $\X\_{d,i}$ is the difference scores for case i, $M\_d$ is the mean difference score, and $ s\_{d\_z}$ is the standard deviation of the difference scores.

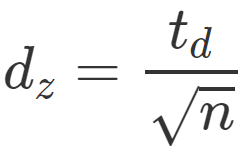
Equivalently, $s\_{d\_z}$ can be calculated as :

 (x.9)

$s\_{d\_z} = \sqrt{s\_1^2 + s\_2^2 - 2(r\_{12}s\_1s\_2)}$

(Cohen 1988 p. 48)

Where $s\_1$ and $s\_2$ are the variances of groups one and two, and $r\_{12}$ 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 $d\_z$, 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 $dz$. This can make a large difference to the $d\_z$, for example, taking the example the two groups have equal variance, and there is a one standard deviation difference between group, $d\_z$ can vary between .707 (when $r\_{12} = 0$) and infinity (when $r\_{12}$ approaches 1). For this reason it has been argued that classical Cohen’s d (equation x.1) should be interpreted in lieu of $d\_z$ for maximum interpretability and comparability across experimental designs {Morris, 2002 #808}. However, for the purposes of power analysis, it is necessary to use $d\_z$, due to the fact that as the correlation between repeated measures increases the standard error of the difference decreases, or equivalently the size of the test statistic increases. In order to calculate $d\_z$ from reported test statistics, equation [Rosen] can be used.

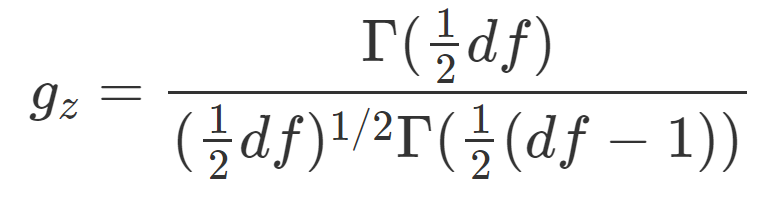
 equation [Rosen]

$$d\_z = {frac{t\_d}{sqrt{n}}} $$

(Lakens, 2013) Equation 7

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

$d\_z$ 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)   
$$g\_z = \frac{\Gamma(\frac{1}{2}df)}{(\frac{1}{2}df)^{1/2} \Gamma(\frac{1}{2}(df-1))}$$

Where $df =$ degrees of freedom (i.e., $n - 1$ as per repeated measures *t*-tests) and $\Gamma (x)$ is the gamma function.

Confidence interval coverage around Cis not robust to non-normality.

BONNETS DELTA – homogeneity of variance is not required – non-normality is an issue for the others

BONNETTE 2009 HERE ????

Categorical effect sizes:

Association:

**Interpreting effect sizes**

Power analysis for a t-test for mean differences between independent groups can either take as input parameters the mean difference along with an estimate of the expected SD of the groups, or a direct estimate of the standardized effect size between groups. The most commonly referenced program for performing power analysis, g\*power [Honours research too, also see later chapter], accepts either as input, although the default input is Cohen’s *d* (Faul, Erdfelder, Lang, & Buchner, 2007)*.*

Proportion overlap (U) -



*Figure [Cohen’s d as population distributions]*. Population distributions with a mean difference of .2, .5, .8 and 1.2 Cohen’s d, along with the percentage overlap between populations (calculated assuming that populations are normally distributed, have equal variance, and equal sample sizes, using equations from (Reiser & Faraggi, 1999)).

**Effect size benchmarks**

The use of standardized effect size benchmarks such as those proposed in Cohen (1962) is often criticized on the grounds that these benchmarks are not empirically validated (i.e., the “medium” effect size benchmark is not the mean or median effect across psychology research), and that using standardized effect size benchmarks are likely to be poor estimates of the actual effect size of any particular experiment. Not only do they not adhere to the average size of effect seen in a particular research subfield, but researchers will often have additional information about the effect sizes that could plausibly be expected from a given experiment.

The original effect size benchmarks were developed by Cohen to reflect approximate …

The “small” and “large” effect size benchmarks examined

To get an intuitive sense of what each effect size is, it is worth examining another effect size, the percentage point group overlap (always calculated assuming that the distribution of the two groups is normal PERCENTAGE OVERLAP FORMULA -> (Reiser & Faraggi, 1999).

In fact, the use of standardized effect sizes has been criticized in of itself, on the grounds that they make it more difficult to extract the main an (e.g., Cohen’s *d*) not only that, but ??

However these benchmarks are useful in the case of power surveys, in which case we are not interested in estimating the power of any individual experiment, but instead can examine the hypothetical power of a body of studies at a variety of plausible effect sizes.

Look at:

Interpreting the magnitudes of correlation coefficients. (Hemphill 2003)

4.1.3 True average effect sizes seen in areas of psychology research

Another approach to developing benchmarks effect sizes in psychological research has been to

DISCUSS ESTIMATES OF THE ACTUAL AVERAGE EFFECT SIZE SEEN IN FIELDS OF PSYCHOLOGICAL RESEARCH

Systematic review

A Google scholar search of “average effect size” found 7 articles.

See C:\Users\fsingletonthorn\Documents\PhD\Dissertation Documents\List of papers providing effect size benchmarks.xlsx for tracking document

Cooper, H., & Findley, M. (1982). Expected Effect Sizes: Estimates for Statistical Power Analysis in Social Psychology

* Probably possible to just double the f statistics to get d for df = 1, maybe for greater ones as well

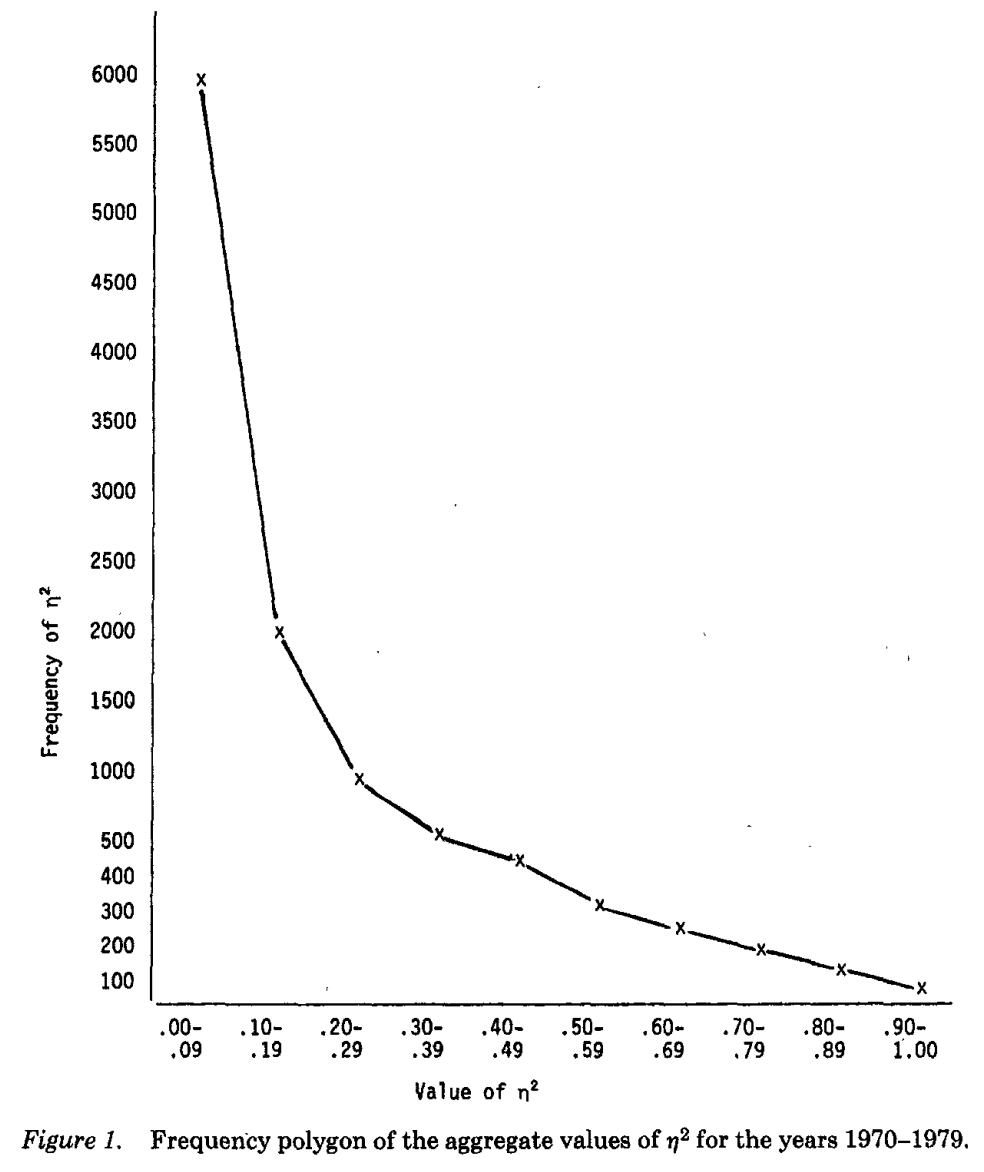


Figure [counselling] Frequency polygon of all univariate inferential statistics reported in the Journal of Counselling Psychology, 1970-1979 Reproduced from {Haase, 1982 #833}

4.1.4

[EffectSizeBenchmarksImage.R] FIX % location

Maybe explain what happens to the sampling distribution for the mean difference as the groups get further apart. Non-central ts, etc.

4.1.5

Attempts to estimate

Richard, Bond Jr., & Stokes-Zoota (2003) One hundred years of social psychology …

Is it reasonable to use this as a guide to determining sample sizes – no, not really, it’s going to be inflated. Attempts to correct for sample size… MAYBE DO THIS ???

MAYBE TRY TO ACCOUNT FOR PUBLICATION BIAS IN SOME WAY??

Other reasonable benchmarks? Minimum clinically significant estimates?

References

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

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

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

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

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

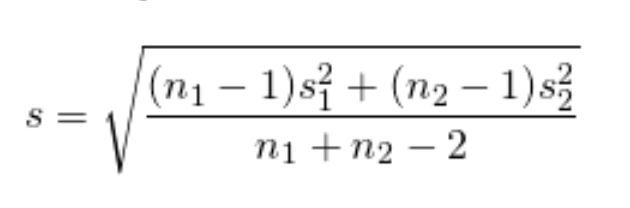
Reiser, B., & Faraggi, D. (1999). Confidence Intervals for the Overlapping Coefficient: the Normal Equal Variance Case. *Journal of the Royal Statistical Society: Series D (The Statistician), 48*, 413-418. doi:10.1111/1467-9884.00199

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

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

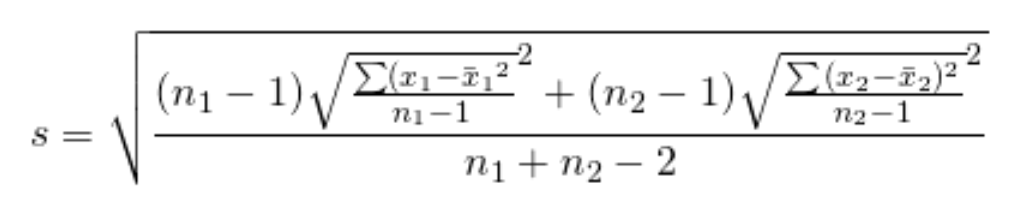
Supplementary material [Cohen’s d/Hedges’ g]

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

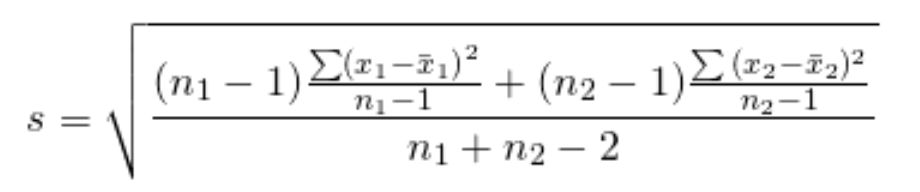
$$s = \sqrt{\frac{(n\_1 -1)s\_1^2 + (n\_2 -1)s\_2^2}{n\_1 + n\_2 - 2}} $$

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]

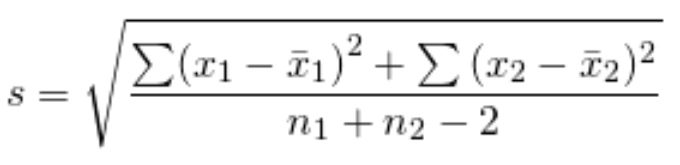
$$s = \sqrt{\frac{(n\_1 -1)\sqrt{\frac{\sum ({x\_1-\bar{x}\_1}^2}{n\_1-1}}^2 + (n\_2 -1)\sqrt{\frac{\sum {(x\_2-\bar{x}\_2)^2}}{n\_2-1}}^2}{n\_1 + n\_2 - 2}}$$

Simple algebra then simplifies this formula to equation [Simplified1].

 [simplified1]

$$ s = \sqrt{\frac{(n\_1 -1)\frac{\sum ({x\_1-\bar{x}\_1)}^2}{n\_1-1} + (n\_2 -1){\frac{\sum {(x\_2-\bar{x}\_2)^2}}{n\_2-1}}}{n\_1 + n\_2 - 2}}$$

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

 [simplified2]

$$s = \sqrt{\frac{\sum ({x\_1-\bar{x}\_1)}^2 + {\sum {(x\_2-\bar{x}\_2)^2}}}{n\_1 + n\_2 - 2}}$$

EVERYTHING TOGETHER FOR COHEN’S D:

$$d = \frac{\bar x\_1 -\bar x\_2}{

\sqrt{\frac{\sum(x\_1 - \bar{x}\_1)^2 + (x\_2 - \bar{x}\_2)^2}{n\_1 + n\_2 - 2}}}$$

Stats exchange answer:

It seems when people say Cohen's d they mostly mean:

$$d = \frac{\bar{x}\_1 - \bar{x}\_2}{s}$$

Where $s$ is the pooled standard deviation,

$$s = \sqrt{\frac{\sum(x\_1 - \bar{x}\_1)^2 + (x\_2 - \bar{x}\_2)^2}{n\_1 + n\_2 - 2}}$$

There are other estimators for the pooled standard deviation, probably the most common apart from the above being:

$$s^\* = \sqrt{\frac{\sum(x\_1 - \bar{x}\_1)^2 + (x\_2 - \bar{x}\_2)^2}{n\_1 + n\_2}}$$

Notation here is remarkably inconsistent, but sometimes people say that the the $s^\*$ (i.e., the $n\_1 + n\_2$ version) version is called Cohen's $d$, and reserve the name Hedge's $g$ for the version that uses $s$ (i.e., with Bessel’s correction, the n1+n2−2 version). This is a bit weird as Cohen outlined both estimators for the pooled standard deviation (e.g., $s$ version on p. 67, Cohen, 1977) before Hedges wrote about them (Hedges, 1981).

Other times Hedge's g is reserved to refer to either of the bias corrected versions of a standardised mean difference that Hedges developed. Hedges (1981) showed that Cohen's d was upwardly biased (i.e., its expected value is higher than the true population parameter value), especially in small samples, and proposed a correction factor to correct for Cohen's d's bias:

Hedges's g (the unbiased estimator):

$$g = d \* (\frac{\Gamma(df/2)}{\sqrt{df/2 \,}\,\Gamma((df-1)/2)})$$

Where $df = n\_1 + n\_2 -2$ for an independent groups design, and $\Gamma$ is the gamma function.

(originally Hedges 1981, this version developed from Hedges and Olkin 1985, p. 104)

However, this correction factor is fairly computationally complex, so Hedges also provided a computationally trivial approximation that, while still biased, is only biased to an extremely small extent:

Hedges' $g^\*$ (the computationally trivial approximation):

$$ g^\* = d\*(1 - \frac{3}{4(df) - 1})$$

Where $df = n\_1 + n\_2 -2$ for an independent groups design.

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

But, as for what people mean when they say Cohen's d vs. Hedges' g vs. g\*,

people seem to refer to any of these three estimators as Hedge's g or Cohen's d interchangeably, although I've never seen someone write "$g^\*$" in a non-methodology/stats research paper. If someone says "unbiased Cohen's d", you're just going to have to take your best guess at either of the last two (and I think there might even be another approximation that has been used for Hedge's $g^\*$ too!).

They are all virtually identical if $n > 20$ or so, and all can be interpreted in the same way. For all practical purposes, unless you're dealing with really small sample sizes, it probably doesn't matter which you use (although if you can pick, you may as well use the one that I've called Hedges' g, as it is unbiased).

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)