Received: 10 November 2024 | Accepted: 12 March 2025

DOI: 10.1111/bmsp.12389

the british psychological society
promoting excellence in psychology

INVITED ARTICLE

Effect sizes for experimental research

Larry V. Hedges

Department of Statistics and Data Science, Northwestern University, Evanston, Illinois, USA

Correspondence
Larry V. Hedges, Department of Statistics and Data Science, Northwestern University, 2046 N. Sheridan Road, Evanston, IL 60611, USA.
Email: l-hedges@northwestern.edu

Abstract

Good scientific practice requires that the reporting of the statistical analysis of experiments should include estimates of effect size as well as the results of tests of statistical significance. Good statistical practice requires that effect size estimates be reported along with some indication of their statistical uncertainty, such as a standard error. This article provides a review of effect sizes for experimental research, including expressions for the standard error of each effect size. It focuses on effect sizes for experiments with treatments having a single degree of freedom but also includes effect sizes for treatments with multiple degrees of freedom having either fixed or random effects.

KEYWORDS

Cohen's *d*, effect size, Hedges' *g*, intraclass correlation, ω^2. η^2

1 | INTRODUCTION

Good scientific practice requires that the reporting of the statistical analysis of experiments should include estimates of effect size as well as the results of tests of statistical significance. Good statistical practice requires that effect size estimates be reported along with some indication of their statistical uncertainty, such as a standard error. This article provides a review of effect sizes for experimental research, including expressions for the standard error and methods to compute confidence intervals for each effect size. It focuses on effect sizes for experiments with treatments having a single degree of freedom but also includes effect sizes for treatments with multiple degrees of freedom having either fixed or random effects.

Effect sizes are numerical representations of the results of empirical research studies. They represent the magnitude of effects on a continuous scale (unlike binary tests of statistical significance). They often use standardization to make their meaning less dependent on the particular measurement tools used

I thank the reviewers whose comments helped to substantially improve this manuscript.

in a particular study that is more broadly interpretable. Because they can improve interpretation of statistical analyses, reporting of effect sizes along with significance tests has been recommended as good statistical practice by the Wilkinson (1999), the American Educational Research Association (2006), the International Committee of Medical Journal Editors (2007) and the American Statistical Association (Wasserstein & Lazar, 2016). Good statistical practice requires that effect sizes, like other descriptive statistics, should be reported with some indication of their uncertainty, such as a standard error or a confidence interval.

Effect sizes are particularly helpful in comparing the results of different studies of the same phenomenon. They are used formally in meta-analysis but also informally to interpret the results of a single study and to understand whether two or more studies seem to have gotten similar answers among other things (see Kelley & Preacher, 2012).

Many different effect size measures have been proposed, but the definition of most effect size measures for continuous outcomes is primarily motivated by one of three principles. One principle is to describe the parameter that determines statistical power of the statistical analysis whose results the effect sizes is intended to describe. The effect size measures used by Cohen in his 1988 book on power analysis are of this type. A second principle is to characterize variance of a dependent variable accounted for by a linear or nonlinear relation with the independent variable in the statistical analysis. The correlation coefficient and correlation ratio are of this type. A third principle is to characterize the overlap between distributions being compared. Cohen's U_3, the What Works Clearinghouse's improvement measure and stochastic dominance are of this type.

This paper deals with effect sizes for experiments and other research designs comparing group means on a continuous outcome variable. It does not deal with effect sizes for discrete outcomes, such as odds ratios or risk ratios, nor does it address effect sizes for studies with continuous independent variables, such as correlation coefficients.

1.1 | What makes a good effect size?

Because effect sizes are intended to aid in interpretation, they should have a 'natural' interpretation to researchers in the field in which they are used. This is partly a social consideration that might differ from field to field. Effect sizes should depend as little as possible on incidental features of the research design and as much as possible on the structure of the populations being studied. For example, *p*-values are poor effect sizes because they depend both on research design choices (i.e., sample size) and population structure. A small *p*-value can be obtained either from a small treatment effect (difference between group means) with a large sample size or a large treatment effect with a small sample size. Thus, the same population structure can yield very different *p*-values.

Because effect sizes are intended to describe the results of statistical analyses, they must be congenial with those statistical analyses and the research designs whose results they represent. For example, it would be peculiar if a smaller effect size in a study was statistically significant but a larger one was not. This requirement leads to effect sizes that are functionally related to noncentrality parameters that determine the statistical power of significance tests. Often more than one effect size is possible for a given statistical analysis, but the mathematical relation to the noncentrality parameter is simpler for one of them, so it is preferable as more mathematically natural.

1.2 | Metrics for effect sizes

Standardized effect sizes in psychology and the social sciences are dimensionless, meaning that they are not defined by external units of measurement like pounds, inches or points on the scale of any particular measurement instrument. Such effect sizes, while technically dimensionless, do have a metric determined by the standard deviation used to define the effect size. The metric is usually chosen to be

the total standard deviation of the outcome within the treatment groups being compared. This metric depends less on the details of the design and makes the effect size more generally interpretable.

In some areas of research, interest may focus less on status (measures of outcome at one point in time) and more on change over time (see, e.g., Dankel & Loenneke, 2021; Gibbons et al., 1993). In such situations, gain score standard deviations may be used to define the effect size.

There are situations where the same metric is used universally in a research area (e.g., dollars, pounds or euros when the outcome is earnings or wealth). In such situations, there may be no need to standardize. That is, the outcome might be better represented as a treatment effect in the units used in that field.

1.3 | The appropriate effect size depends on the statistical analysis used in the experiment

Because effect sizes must be congenial to the statistical analyses they represent, somewhat different effect sizes are appropriate in conjunction with fixed effects versus random effects statistical analyses. In addition, different effect sizes are appropriate to describe effects that arise from analyses where the treatment has a single degree of freedom (e.g., treatment vs. control) versus those that have multiple degrees of freedom (e.g., omnibus tests of differences among three or more means).

2 | EFFECT SIZES FOR COMPARISONS OF TWO GROUPS

The most frequently used effect sizes in experiments that compare two groups or a one degree of freedom contrast among group means. The statistical model in this case is that the (population) means of the outcome in the two groups are μ_1 and μ_2, and the standard deviation within the two groups is σ. If the data are normally distributed, the natural effect size parameter is

$$\delta = \frac{\mu_1 - \mu_2}{\sigma}, \tag{1}$$

sometimes called Cohen's *d*. I use the Greek letter δ to emphasize the distinction between population parameters (represented by Greek letters) and sample estimates (represented by Roman letters). The measure δ predates Cohen. It was used as a measure of the distance between two normal distributions with the same dispersion by Mahalanobis (1936), but he cited even earlier statistical research on it. Similarly, Tilton (1937) described measures of overlap between populations as a conventional way to interpret the standardized mean difference.

The obvious estimate of δ is

$$d = \frac{\overline{Y}_1 - \overline{Y}_2}{S}, \tag{2}$$

where $\overline{Y}_1$ and $\overline{Y}_2$ are the sample means in groups 1 and 2, respectively, and S is the pooled-within-groups sample standard deviation. The exact sampling distribution of d follows from the fact that the test statistic

$$t = d\sqrt{\frac{n_1 n_2}{n_1 + n_2}} \tag{3}$$

has the noncentral t-distribution with $n_1 + n_2 - 2$ degrees of freedom and noncentrality parameter $\lambda = \delta\sqrt{n_1 n_2 / (n_1 + n_2)}$. In fact, d is often calculated in research reviews by using Equation (3) and solving for d, given t (or $F = t^2$), n_1 and n_2.

The estimate d has a slight small sample bias, which can be corrected by multiplying by the correction factor $J(\mathrm{df})$ to obtain the (minimum variance) unbiased estimator (sometimes called Hedges' g)

$$g = J(\mathrm{df})d = J(\mathrm{df})\left(\frac{\overline{Y}_1 - \overline{Y}_2}{S}\right) \tag{4}$$

where the correction factor $J(\mathrm{df})$ is given (to a very good approximation) by

$$J(\mathrm{df}) = 1 - \frac{3}{4\mathrm{df} - 1} \tag{5}$$

and df is the degrees of freedom of the standard deviation S (Hedges, 1981). In a two-group experiment, $\mathrm{df} = n_1 + n_2 - 2$.

The effect size g is also suitable for contrasts among $k > 2$ groups means. In that case, the mean difference is replaced by the contrast; the denominator is the square root of the within-cell mean square; and $\mathrm{df} = N - k$, where N is the total sample size.

The exact variances of d and g were given by Hedges (1981), but because they depend on δ, the variance of any sample estimate d or g must itself be estimated. One approach to estimation of the variance is to insert an estimate of δ (d or g) in place of δ in the exact variance formulas. Extensive simulation studies have suggested that this is not more accurate than using the variance estimator given in Equation (6), which was given by Hedges (1982) and recommended subsequently in Hedges and Olkin (1985) and Cooper et al. (2019).

The most widely used variance estimator is

$$\left[SE(g)\right]^2 = \left[J\left(n_1 + n_2 - 2\right)\right]^2 \left[\frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2\left(n_1 + n_2\right)}\right], \tag{6}$$

which is a large sample approximation based on a model where δ is fixed when the sample size becomes large. The variance estimator

$$\left[SE(g)\right]^2 = \left[J\left(n_1 + n_2 - 2\right)\right]^2 \left[\frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2\left(n_1 + n_2 - 2\right)}\right] \tag{7}$$

arises from a different large sample approximation; this one is based on a model where the noncentrality parameter of the noncentral t-distribution

$$\lambda = \delta\sqrt{\frac{n_1 n_2}{n_1 + n_2}} \tag{8}$$

is fixed as the sample size becomes large. In this second model, because λ is fixed as the sample size becomes large, δ must tend to zero. The difference between the two variance estimators is in the denominator of the second term and is negligible unless the sample size is very small. For example, even if $n_1 = n_2 = 10$, the ratio of Equations (6) and (7) is 0.997, and if $n_1 = n_2 = 30$, the ratio is 0.999.

In computing the variance of the unbiased estimator g, some users ignore the correction factor $J(m)$, which makes little difference if sample sizes are not small.

An approximate 100 $(1 - a)$% confidence interval for δ is given by

$$g - z_{\alpha} SE(g) \le \delta \le g + z_{\alpha} SE(g) \tag{9}$$

where z_{α} is the level α two-tailed critical value of the standard normal distribution. An exact 100 $(1 - \alpha)$% confidence interval for δ

$$\delta_L \le \delta \le \delta_U \tag{10}$$

can be obtained by finding the values of the noncentrality parameter λ_L and λ_U, so that $F(t|\, n_1+n_2-2;\lambda_L)=\alpha/2$ and $F(t|\, n_1+n_2-2;\lambda_U)=1-\alpha/2$, where $t=d\sqrt{n_1n_2/(n_1+n_2)}$ and $F(x|\mathrm{df};\lambda)$ is the cumulative distribution function of the noncentral t-distribution with df degrees of freedom and noncentrality parameter λ, then computing $\delta_L=\lambda_L\sqrt{(n_1+n_2)/n_1n_2}$ and $\delta_U=\lambda_U\sqrt{(n_1+n_2)/n_1n_2}$ (Hedges & Olkin, 1985, p. 91). This method is implemented in an R package by Kelley (2007).

2.1 | Interpretation of δ, d and g

Effect sizes are dimensionless constants and therefore have no intrinsic interpretation. One interpretation of δ is that it is the treatment effect in a scale with unit variance – the standardized treatment effect. Another interpretation of the effect size is what Tilton (1937) called 'overlap': The proportion of group 1 that exceeds the mean of group 2. This is what Cohen (1988) later called U_3, which in the notation of this paper is

$$U_3=\mathrm{Prob}\{Y_1>\mu_2\}, \tag{11}$$

where Prob{X} is the probability of event X. If the data are normally distributed and the within-group variances are equal, then $U_3=\Phi(\delta)$, where $\Phi(x)$ is the standard normal cumulative distribution function. A natural (but slightly biased) estimator of U_3 is

$$\hat{U}_3=\Phi(g), \tag{12}$$

which has approximate variance

$$\left[SE\left(\hat{U}_3\right)\right]^2=\frac{e^{-g^2}}{2\pi}\left[\frac{n_1+n_2}{n_1n_2}+\frac{g^2}{2(n_1+n_2)}\right]. \tag{13}$$

An unbiased estimator is known but is quite complex (see Hedges & Olkin, 2016).

A more general overlap measure is the stochastic dominance, the probability that a randomly chosen observation in group 1 exceeds a randomly chosen member of group 2, that is

$$\mathrm{Prob}\{Y_1>Y_2\}.$$

This parameter was called Δ_2 by Hedges and Olkin (2016) and the common language effect size by McGraw and Wong (1992). If the data are normally distributed and the within-group variances are equal, then

$$\Delta_2=\Phi\left(\delta/\sqrt{2}\right) \tag{14}$$

and a natural (but slightly biased) estimator of Δ_2 is

$$\hat{\Delta}_2=\Phi\left(g/\sqrt{2}\right), \tag{15}$$

which has approximate variance

$$\left[SE\left(\hat{\Delta}_2\right)\right]^2 = \frac{e^{-g^2/2}}{4\pi}\left[\frac{n_1+n_2}{n_1 n_2} + \frac{g^2}{2\left(n_1+n_2\right)}\right]. \tag{16}$$

Confidence intervals for U_3 or Δ_2 can be computed by first computing confidence intervals for g, then transforming the confidence limits to the confidence limits for U_3 or Δ_2. For example, the lower and upper confidence limits, respectively, for U_3 are $U_{3L} = \Phi(\delta_L)$ and $U_{3U} = \Phi(\delta_U)$.

The measures of overlap between distributions that Cohen (1988) describes, and the common language effect size Δ_2, are all functionally related to δ if the populations being compared are normal with identical variances. Note that the relation between δ and measures of overlap depends on normality and equal variances – the interpretation of δ and its relation to overlap between distributions can be wildly different if these assumptions are not met (see Hedges, Hedges, 2025).

2.2 | Effect sizes for omnibus comparisons of more than two means in fixed effects analyses

Statistical analyses of omnibus effects reflecting differences among $k > 2$ means are typically represented by effect sizes that quantify the variation among group means. Let $\mu_1, \mu_2, \ldots, \mu_k$ be the means (parameters) of k groups of n normally distributed observations, each having the same variance σ^2, so that the total sample size is $N = kn$. Then the variance of the means is

$$\tau_F^2 = \frac{1}{k-1}\sum_{i=1}^{k}\left(\mu_i - \overline{\mu}\right)^2 \tag{17}$$

where the subscript F refers to the fact that means are considered fixed effects (so that τ_F^2 is not technically the variance of a random variable). The recommended effect size measure is the standardized effect size variance (called f^2 by Cohen, 1988)

$$f_F^2 = \tau_F^2 / \sigma^2. \tag{18}$$

The standardized effect size variance quantifies the size of variation between groups means compared to the variation within the groups. Alternatively, the effect size variance f_F^2 is the variance of the standardized effects $\left(\mu_i - \overline{\mu}\right)/\sigma$.

The standardized effect variance is related to the distribution of the F-test statistic used to test the null hypothesis that the k group means are identical, which has the noncentral F-distribution with $k - 1$ degrees of freedom in the numerator, $k(n-1) = N - k$ degrees of freedom in the denominator and noncentrality parameter

$$\lambda = \frac{n\sum_{i=1}^{k}\left(\mu_i - \overline{\mu}\right)^2}{\sigma^2} = \left(\frac{n(k-1)}{\sigma^2}\right)\left(\frac{1}{k-1}\sum_{i=1}^{k}\left(\mu_i - \overline{\mu}\right)^2\right) = n(k-1)f_F^2. \tag{19}$$

If the design is balanced so that the sample sizes n_k in the k groups are all equal to n, then it follows from the expected value of the noncentral F-distribution that an unbiased estimate of f_F^2 is

$$\widehat{f}_F^2 = \frac{1}{n}\left(\frac{(N-k-2)F}{N-k} - 1\right), \tag{20}$$

where F is the F-statistic used to test the omnibus null hypothesis that all group means are identical. In the balanced case, the variance of the noncentral F-distribution implies that the variance of $\widehat{f}_F^2$ is

$$\left[SE\left(\hat{f}_F^2\right)\right]^2 = \frac{2\left[(k-1)\left(1+nf_F^2\right)^2 + \left(1+2nf_F^2\right)(N-k-2)\right]}{n^2(k-1)(N-k-4)}. \tag{21}$$

Exact 100(1 – *a*)% confidence interval for f_F^2

$$f_{FL}^2 \le f_F^2 \le f_{FU}^2 \tag{22}$$

can be obtained from the noncentrality parameter in a manner analogous to those for δ. First find the values of the noncentrality parameters λ_L and λ_U, so that $G\left(F|a,b;\lambda_L\right) = \alpha/2$ and $G\left(F|a,b;\lambda_U\right) = 1-\alpha/2$, where $G(x|df_1, df_2; \lambda)$ is the cumulative distribution function of the noncentral *F*-distribution with numerator and denominator degrees of freedom df_1 and df_2, respectively, and noncentrality parameter λ. Then compute $f_{FL}^2 = \lambda_L/n(k-1)$ and $f_{FU}^2 = \lambda_U/n(k-1)$ (see, e.g., Steiger, 2004). This method is implemented in an R package by Kelley (2007).

If the design is unbalanced so that the sample sizes $n_1, \ldots, n_k$ in the k groups are not all equal, then the estimate of f_F^2 and its variance are considerably more complicated, but if the imbalance is not profound, then Equations (20) and (22) can be approximately correct if a compromise sample size (such as the harmonic mean) is substituted in place of *n*.

An alternative effect size is a fixed effects analogue to the intraclass correlation

$$\omega^2 = \frac{\tau_F^2}{\tau_F^2 + \sigma^2} = \frac{f_F^2}{f_F^2 + 1}, \tag{23}$$

which was introduced by Hays (1973, pp. 414, 485). If the design is balanced with one factor, then the usual estimator of ω^2 is

$$\hat{\omega}^2 = \frac{SSB - (k-1)MSW}{SSB + SSW + MSW}, \tag{24}$$

where *SSB* is the between-groups sum of squares, *SSW* is the within-group sum of squares and *MSW* is the within-group mean square. The approximate variance of $\hat{\omega}^2$ obtained from Equation (21) is

$$[SE(\hat{\omega}^2)]^2 = \frac{2(1-\hat{\omega}^2)^2[(k-1)[1+(n-1)\hat{\omega}^2]^2 + (1-\hat{\omega}^2)(N-k-2)[1+(2n-1)\hat{\omega}^2]]}{n^2(k-1)(N-k-4)}. \tag{25}$$

An exact 100(1 – *a*)% confidence interval

$$\omega_L^2 \le \omega^2 \le \omega_U^2 \tag{26}$$

for ω^2 can be calculated by first obtaining confidence limits for f_{FL}^2 and f_{FU}^2 for f_F^2 and then transforming them to confidence limits $\omega_L^2 = f_{FL}^2/\left(f_{FL}^2+1\right)$ and $\omega_U^2 = f_{FU}^2/\left(f_{FU}^2+1\right)$ for ω^2.

For a further discussion of ω^2 and formulas for computing ω^2 from a variety of designs, see Kroes and Finley (2023) and Olejnik and Algina (2003).

2.3 | Effect sizes for omnibus comparisons of more than two means in random effects analyses

Random effects are typically represented by effect sizes that quantify the variation among group means. In random effects analyses, there are also k groups of normally distributed observations, each having the same variance σ^2, but the group means $\mu_1, \mu_2, \ldots, \mu_k$ in the experiment are considered a sample from a population of means. Denote the variance of this population of means by τ^2. The object of the statistical

analysis is to estimate the variance component τ^2 and test the hypothesis that $\tau^2 = 0$. One appealing effect size, called f^2 by Cohen (1988), is the standardized effect variance defined by

$$f^2 = \tau^2 / \sigma^2. \tag{27}$$

The standardized effect variance is related to the distribution of the F-test statistic used to test the null hypothesis that the variance component among the k group means is zero, which has a distribution that is $(1 + nf^2)$ times a (central) F-distribution with $k - 1$ degrees of freedom in the numerator and $k(n - 1) = N - k$ degrees of freedom in the denominator.

If the design is balanced so that the sample sizes in the k groups are equal to n, then the mean of the F-distribution implies that an unbiased estimate of f^2 is

$$\hat{f}^2 = \frac{1}{n}\left(\frac{(N - k - 2)}{N - k}F - 1\right), \tag{28}$$

where $N = kn$ is the total sample size and F is the F-statistic for testing the null hypothesis that $\tau^2 = 0$. Note that this estimate (for the groups' random model) is identical to Equation (20) for the groups' fixed model, but the variance is different.

If the design is balanced, the variance of $\hat{f}^2$ is

$$\left[SE\left(\hat{f}^2\right)\right]^2 = \frac{\left(1 + nf^2\right)^2}{n^2}\frac{2(N - 3)}{(k - 1)(N - k - 4)}. \tag{29}$$

Because Equation (29) is proportional to $f^2/(k - 1)$, the uncertainty of $\hat{f}^2$ is likely to be large unless the number k of groups is large.

Exact $100(1 - a)\%$ confidence interval for f^2

$$f_L^2 \le f^2 \le f_U^2 \tag{30}$$

can be obtained by finding the values of F_L and F_U, so that $G\left(F_L | a, b\ ; 0\right) = \alpha/2$ and $G\left(F_U | a, b; 0\right) = 1 - \alpha/2$, where $G(x|\mathrm{df}_1, \mathrm{df}_2; 0)$ is the cumulative distribution function of the central F-distribution with numerator and denominator degrees of freedom df_1 and df_2, respectively, then computing

$$f_L^2 = \frac{1}{n}\left(\frac{(N - k - 2)F_L}{N - k} - 1\right) \text{and} f_U^2 = \frac{1}{n}\left(\frac{(N - k - 2)F_U}{N - k} - 1\right)$$

If the design is unbalanced (the sample sizes in the k groups are not all equal), then the estimate of f^2 and its variance is considerably more complicated, but if the imbalance is not profound, then Equations (28) and (29) are approximately correct if a compromise sample size (such as the harmonic mean) is substituted in place of n.

An alternative effect size is the intraclass correlation

$$\rho_I = \frac{\tau^2}{\tau^2 + \sigma^2} = \frac{f^2}{f^2 + 1}, \tag{31}$$

which was introduced by Fisher (1925). If the design is balanced, then the usual estimator of ρ_I is

$$\hat{\rho}_I = \frac{MSB - MSW}{MSB + (n - 1)MSW}. \tag{32}$$

This estimator is not unbiased, but has approximate bias $(1-\varrho)^2/n$. An unbiased estimator is known, but it is rather complex (see Olkin & Pratt, 1958). The approximate variance of $\widehat{\rho}_I$ was given by Fisher (1925) as

$$\left[SE\left(\widehat{\rho}_I\right)\right]^2 = \frac{2\left(1-\widehat{\rho}_I^2\right)\left[1+(n-1)\widehat{\rho}_I\right]^2}{n(n-1)(k-1)}. \tag{33}$$

Confidence intervals for ϱ_I can be obtained by transforming confidence intervals for f^2 in a manner analogous to obtain confidence intervals for ω^2 from confidence intervals for f_F^2.

A third effect size that is sometimes used when treatments have random effects is the correlation ratio

$$\eta^2 = \frac{\tau^2+\sigma^2/n}{\tau^2+\sigma^2}, \tag{34}$$

which is the ratio of the variance of the group means $(\tau^2+\sigma^2/n)$ to the (total) variance of the observations $(\tau^2+\sigma^2)$, which was suggested by Pearson (1905). The usual estimate of the correlation ratio in a one-factor analysis of variance (ANOVA) is

$$\widehat{\eta}^2 = \frac{SSB}{SSB+SSW} = \frac{SSB}{SST}, \tag{35}$$

where SSB is the sum of squares between effects and SST = SSB + SSW is the total sum of squares. The general sampling distribution of $\widehat{\eta}^2$ was derived by Wishart (1932), who showed that the bias of $\widehat{\eta}^2$ is of the order if $1/n$, even in large samples. An estimator of η^2 that is claimed to be less biased is

$$E^2 = 1-\frac{N-1}{N-k}\left(1-\widehat{\eta}^2\right) = \frac{MST-MSW}{MST} \tag{36}$$

(see Kelly, 1935). Wishart gave the approximate variance of $\widehat{\eta}^2$ as

$$\left[SE\left(\widehat{\eta}^2\right)\right]^2 = \frac{2(k-1)+[4N-2(k-1)]\widehat{\eta}^2}{(N-k)^2\left(1-\widehat{\eta}^2\right)}. \tag{37}$$

One reason that the correlation ratio is less desirable as an effect size is that it depends strongly on a design parameter (n, the group sample size) and therefore may be quite different in two designs that have identical population parameters σ^2 and τ^2. Fisher was critical of the correlation ratio, saying that 'as a descriptive statistic the utility of the correlation ratio is extremely limited' (Fisher, 1925/1973, p. 258).

3 | CONDITIONAL EFFECT SIZES

In designs that control for independent variables other than the treatment of interest, the effect of interest may be the effect of the treatment *controlling for other independent variables*. For example, in factorial designs, the main effect of a treatment represents its effect while controlling for all other treatments. In situations where the effect of interest is conditional on (controlling for) other independent variables, it is important to consider this conditionality in deciding on an effect size. First, the conditionality defines the meaning of the effect being described by the effect size. The effect is not the global effect, but the effect controlling for a certain set of other independent variables. Second, the nature of the conditional effect may have implications for the most appropriate metric for standardization: Should the standard deviation used to standardize the effect include variation associated with the independent variables being controlled or not?

3.1 | Effect sizes from designs using covariates

Sometimes designs measure a covariate X and an outcome Y on the same individuals in each treatment group. The covariate may be used in the analysis to increase precision in either of the two ways. It can be used in a regression adjustment (e.g., in the analysis of covariance). Alternatively, if X and Y are measured on the same scale (such as when X is a pretest and Y is a posttest), the analysis might involve gain scores ($Y - X$). In either case, the (population) means of the pretest in the two treatment groups on X are μ_{1X} and μ_{2X}; the standard deviation within the two groups on X is σ_X; the means in the two groups on Y are μ_{1Y} and μ_{2Y}; the standard deviation within the two groups on Y is σ_Y; and the pooled-within-group correlation between X and Y is ρ.

3.1.1 | Effect sizes for analyses using statistical adjustment

If covariates are used in the analysis, the test statistic is the ratio of a treatment effect to its standard error, but that standard error depends on the correlation ρ between X and Y. Moreover, there is more than one standard deviation that could be used to construct an effect size. It is conventional to use the standard deviation σ_Y of the outcome score to construct an effect size because this makes the effect size comparable to those constructed from analyses that do not use covariates. Thus, the desired effect size is

$$\delta_A = \frac{\mu_{1A} - \mu_{1A}}{\sigma_Y}, \tag{38}$$

where μ_{1A} and μ_{2A} are the adjusted group means. In randomized experiments, the covariate adjustment should not affect the group means, so that $\mu_{1A}=\mu_1$, $\mu_{2A}=\mu_2$ and $\delta_A=\delta$.

If the statistical analysis uses regression adjustments (such as the analysis of covariance), then μ_{1A} and μ_{2A} are the covariate adjusted means. The statistical test for the treatment effect depends not on S_Y^2 but on the covariate adjusted variance S_{YA}^2, which estimates $\sigma_{YA}^2 = \sigma_Y^2\left(1-\rho^2\right)$, where ρ is the pooled-within-group correlation between X and Y. Therefore, the estimator of δ_A corresponding to g is

$$g_A = J\left(\mathrm{df}_A\right)\left(\frac{\overline{Y}_{1A} - \overline{Y}_{2A}}{S}\right) = J\left(\mathrm{df}_A\right)\sqrt{1-r^2}\left(\frac{\overline{Y}_{1A} - \overline{Y}_{2A}}{S_A}\right), \tag{39}$$

where $\overline{Y}_{1A}$ and $\overline{Y}_{2A}$ are the covariate adjusted means of groups 1 and 2, respectively, S is the pooled-within-groups standard deviation (unadjusted for covariates), S_A is the covariate adjusted (residual) standard deviation, r is the within-group covariate outcome correlation (or multiple correlation if there are $q>1$ covariates) and $\mathrm{df}_A = n_1 + n_2 - 2 - q$.

Because the test statistic is

$$t_A = \sqrt{\frac{n_1 n_2}{n_1 + n_2}}\left(\frac{\overline{Y}_{1A} - \overline{Y}_{2A}}{S_A}\right) = \sqrt{\frac{n_1 n_2}{n_1 + n_2}}\left(\frac{\overline{Y}_{1A} - \overline{Y}_{2A}}{S\sqrt{1-r^2}}\right), \tag{40}$$

it follows that

$$g_A = J(\mathrm{df})t_A\sqrt{\frac{n_1 + n_2}{n_1 n_2}}\sqrt{1-r^2}. \tag{41}$$

The recommended estimate of the variance of g_A is

$$\left[SE\left(g_A\right)\right]^2 = \frac{\left(n_1 + n_2\right)\left(1-r^2\right)}{n_1 n_2} + \frac{d^2}{2\left(n_1 + n_2 - 2 - q\right)} \tag{42}$$

(see Hedges et al., 2023). Because of the presence of $(1 - r^2)$ in the first term in Equation (42), the standard error of g_A differs from the standard error of g given in Equations (6) and (7). As this term typically dominates the standard error, $\mathrm{SE}(g_A) < \mathrm{SE}(g)$ whenever $r^2 > 0$. Also Equation (42) does not include the $J(\mathrm{df})$ term. Confidence intervals for δ_A can be computed using the standard error of g_A as in Equation (9).

A slightly different effect size that may be easier to compute is

$$\delta_T = \frac{\beta}{\sigma_T}, \tag{43}$$

where β is the regression coefficient for the treatment effect parameter (the true value of the regression coefficient for treatment), and σ_T^2 is the variance of the outcome variable (ignoring all the covariates) recommended by Keef and Roberts (2004). A simple estimate of δ_T is

$$d_T = \frac{b}{S_T} = \frac{b\sqrt{\left(n_1 + n_2\right)\left(1 - \rho^2\right)/n_1 n_2}}{\mathrm{SE}(b)}, \tag{44}$$

where b is the sample regression coefficient for the treatment dummy, S_T^2 is the sample variance of the outcome (ignoring all the predictors including the treatment dummy), ρ is the multiple correlation between the predictors (including the treatment dummy) and the outcome and $\mathrm{SE}(b)$ is the standard error of b from the regression analysis.

An estimate of the approximate variance of d_T is

$$\left[SE\left(d_T\right)\right]^2 = \frac{\left(n_1 + n_2\right)\left(1 - r^2\right)}{n_1 n_2} + \frac{d_T^2}{2\left(n_1 + n_2 - 2 - q\right)}. \tag{45}$$

Note that because $d_T < d_A$ unless both are zero, Equations (45) and (42) are similar but not identical.

Note that σ_T^2 is not the same metric as σ^2, the within-treatment groups variance, because it includes between-group variation, and therefore δ_T is not in the same metric as δ. Because σ_T^2 is related to σ^2 via

$$\sigma_T^2 = \sigma^2\left(1 + \frac{n_1 n_2 \delta^2}{\left(n_1 + n_2\right)^2}\right), \tag{46}$$

it follows that an estimate of δ is

$$d_R = \frac{d_T}{\sqrt{1 + n_1 n_2 d_T^2/\left(n_1 + n_2\right)^2}} = c d_T, \tag{47}$$

where $c = 1/\sqrt{1 + n_1 n_2 d_T^2/\left(n_1 + n_2\right)^2}$ is an adjustment factor, which will be relatively small for effect sizes commonly encountered. For example, when $n_1 = n_2$, $c \approx 1.01$ for $d_T = 0.2$, $c \approx 1.03$ for $d_T = 0.5$ and $c \approx 1.09$ for $d_T = 0.8$. The approximate variance of d_R is given by

$$\left[SE\left(d_R\right)\right]^2 = \frac{\left(n_1 + n_2\right)\left(1 - \rho^2\right)/n_1 n_2 + d_T^2/2\mathrm{df}}{\left[1 + n_1 n_2 d_T^2/\left(n_1 + n_2\right)^2\right]^3} = c^6\left[\frac{\left(1 - \rho^2\right)\left(n_1 + n_2\right)}{n_1 n_2} + \frac{d_T^2}{2\mathrm{df}}\right], \tag{48}$$

where $\mathrm{df} = n_1 + n_2 - 2 - q$.

3.1.2 | Effect sizes for analysis using gain scores

If the covariate X and outcome Y are measured on the same scale (e.g., pretest and posttest), the statistical analysis might be carried out using gain scores, then $\mu_{1G} = (\mu_{1Y} - \mu_{1X})$ and $\mu_{2G} = (\mu_{2Y} - \mu_{2X})$ are the mean gains in each group. The statistical test for the treatment effect depends not on S_Y^2 but on the variance of the gains S_G^2. If the standard deviations of the pretest and posttest are equal ($\sigma_Y = \sigma_Y$), then S_G^2 estimates $\sigma_G^2 = 2\sigma_Y^2(1 - \rho)$. In randomized experiments, $\mu_{1G} - \mu_{2G} = \mu_1 - \mu_2$ so that $(\mu_{1G} - \mu_{2G})/\sigma_Y = \delta$.

Therefore, the estimator of δ given in Equation (15) that corresponds to g is

$$g_G = J(\mathrm{df}_G)\left(\frac{\overline{Y}_{1G} - \overline{Y}_{2G}}{S}\right) = J(\mathrm{df}_G)2(1-r)\left(\frac{\overline{Y}_{1G} - \overline{Y}_{2G}}{S_G}\right), \tag{49}$$

where $\overline{Y}_{1G}$ and $\overline{Y}_{2G}$ are the mean gains of groups 1 and 2, respectively, S_G is the pooled-within-groups standard deviation of the gain scores, r is the within-groups covariate outcome correlation and $\mathrm{df}_G = n_1 + n_2 - 2$. The statistic g_G is sometimes called the standardized mean change (see Becker, 1988). Because the test statistic is

$$t_G = \sqrt{\frac{n_1 n_2}{n_1 + n_2}}\left(\frac{\overline{Y}_{1G} - \overline{Y}_{2G}}{S_G}\right) = \sqrt{\frac{n_1 n_2}{n_1 + n_2}}\left(\frac{\overline{Y}_{1G} - \overline{Y}_{2G}}{S\sqrt{2(1-r)}}\right), \tag{50}$$

it follows that

$$g_G = J(\mathrm{df}_G)t_G\sqrt{\frac{n_1 + n_2}{n_1 n_2}}\sqrt{2(1-r)}.. \tag{51}$$

The recommended estimate of the variance of g_G is

$$\left[SE(g_G)\right]^2 = \frac{2(n_1 + n_2)(1-r)}{n_1 n_2} + \frac{d^2}{2(n_1 + n_2 - 2)} \tag{52}$$

(see Hedges et al., 2023). Note that the variance of g_G given in Equation (52) differs from g given in Equation (6) in the first term, which typically dominates the variance, and that $\mathrm{SE}(v_G) < \mathrm{SE}(v)$ whenever $r < ½$ (which implies $2(1 - r) < 1$). Also Equation (52) does not include the $J(\mathrm{df})$ term. Note that Equation (51) is simpler than the expression given in Morris and DeShon (2002) but is as accurate. Confidence intervals for δ can be computed using the standard error of g_G as in Equation (9).

3.2 | Effect sizes in multifactor designs

Factorial designs that include other factors than the treatment of primary interest may include variation in some sources that the investigator may wish to include in the metric for the effect size but other sources whose variation the investigator may not wish to include. Some treatment factors are simply blocking factors that are introduced to increase the precision of the estimate of the treatment effect. Such factors and the variation associated with them have been called 'intrinsic'. The variation due to intrinsic factors is usually pooled with the error term to produce the standard deviation used to compute the effect size. Other treatment factors (e.g., other active treatments) increase variance in ways that are artefactual to the experiment. Such factors and the variation associated with them have been called 'extrinsic'. The variation due to extrinsic factors is usually excluded from the error term to produce the standard deviation used to compute the effect size. A thorough discussion of effect sizes in factorial designs is given by Gillett (2003) and Morris and DeShon (1997).

4 | EFFECT SIZES IN CLUSTER RANDOMIZED DESIGNS

Designs that randomize intact groups or clusters (often called cluster randomized designs) are widely used in field experiments in education and the social sciences. These designs exploit the natural hierarchical structure of many social populations to obtain their samples (e.g., students nested within schools or patients nested within clinics). These designs introduce the complexity that there are several possible standard deviations that might be used to define the effect size: σ_W (the within-cluster standard deviation), σ_B (the between-cluster mean standard deviation) and $\sigma_T = \sqrt{\sigma_B^2 + \sigma_W^2}$ (the total standard deviation ignoring the clusters).

The most natural effect size in such designs uses the metric of the total standard deviation, so that the effect size becomes

$$\delta_C = \frac{\mu_{1\bullet} - \mu_{2\bullet}}{\sigma_T} = \frac{\mu_{1\bullet} - \mu_{2\bullet}}{\sqrt{\sigma_B^2 + \sigma_W^2}}, \tag{53}$$

where $\mu_{1\bullet}$ and $\mu_{2\bullet}$ are the mean outcomes across all clusters in treatment groups 1 and 2, respectively. The standard deviation of all the observations (within treatment groups) S_T does not estimate σ_T (as might be expected) and consequently the estimate standardized by S_T needs a correction. If the design is balanced so that each cluster has the same sample size n, the estimate of δ_T becomes

$$d_C = \left(\frac{\overline{Y}_{1\bullet} - \overline{Y}_{2\bullet}}{S_T}\right)\sqrt{1 - \frac{2(n-1)\rho}{N-2}}, \tag{54}$$

where N is the total sample size and $\rho = \sigma_B^2 / \sigma_T^2 = \sigma_B^2 / \left(\sigma_B^2 + \sigma_W^2\right)$ is the intraclass correlation. The approximate variance of d_T is

$$\left[SE\left(d_C\right)\right]^2 = \frac{\left(N_1 + N_2\right)[1 + (n-1)\rho]}{N_1 N_2} + \frac{d_T^2}{2(M-2)}, \tag{55}$$

where N_1 and N_2 are the total sample sizes within treatment groups, n is the cluster size and $M = N/2n$ is the total number of clusters. Note that the first term of Equation (55) is the same as the first term in Equation (6) but multiplied by the constant $[1 + (n-1)\rho]$. That constant is often called the design effect and represents the penalty (in variance) for using cluster sampling (of persons within clusters) rather than simple random sampling.

If designs are unbalanced, but not profoundly so, Equations (54) and (55) will provide reasonable approximations for the effect size and its variance if the harmonic mean is used in place of n. A thorough discussion of calculating effect sizes in cluster randomized designs is given in Hedges (2007), which includes the considerably more complex expressions for the effect size estimate and its variance in unbalanced designs. A thorough discussion of effect sizes in three-level cluster randomized designs is given in Hedges (2011). Sometimes, designs have nesting in only one group. For example, if the treatment is administered in groups, but the control condition is ungrouped (e.g., small group tutoring vs. a waitlist and group counselling vs. placebo). Methods for effect sizes in such designs are described in Hedges and Citkowicz (2015).

5 | OTHER DESIGNS

5.1 | Factorial designs with random effects

The choice of effect sizes (and the computation of their standard errors) in factorial designs with some random factors can be quite complex. For example, in the multisite individually randomized design

(where the treatment factor is crossed with the sites factor), the effect size might be defined the same way (using the metric of the within-cell standard deviation), but the standard error would be different depending on whether the sites factor was fixed or random. More methodological work on effect sizes in these designs is needed.

5.2 | Longitudinal and repeated-measures designs

The preferred effect sizes in longitudinal designs will usually use the metric of the standard deviation at one point in time, which is straightforward to compute if the raw data are available but more difficult if the effect size is computed only from the summary statistics that may happen to be reported (as in meta-analysis). For a discussion of effect sizes in these designs, see Morris and DeShon (2002) or Kroes and Finley (2023). If the design also includes factors with random effects, this needs to be taken into account when computing the standard errors of effect sizes, but there is little methodological work on effect sizes in such designs.

AUTHOR CONTRIBUTIONS

Larry V. Hedges: Conceptualization; writing – original draft; writing – review and editing.

CONFLICTS OF INTEREST

The author declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

No data was used in this article.

ORCID

Larry V. Hedges https://orcid.org/0000-0002-7531-0631

REFERENCES

American Educational Research Association. (2006). Standards for reporting on empirical social science research in AERA publications. *Educational Researcher*, *35*, 33–40.

Becker, B. J. (1988). Synthesizing standardized mean-change measures. *British Journal of Mathematical and Statistical Psychology*, *41*, 257–278.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum.

Cooper, H., Hedges, L. V., & Valentine, J. (2019). *The handbook of research synthesis and meta-analysis* (3rd ed.). The Russell Sage Foundation.

Dankel, S. J., & Loenneke, J. P. (2021). Effect sizes for paired data should use the change score variability rather than the pre-test variability. *Journal of Strength and Conditioning Research*, *35*(6), 1773–1778. https://doi.org/10.1519/jsc.0000000000002946

Fisher, R. F. (1925/1973). *Statistical methods or research workers*. Oliver and Boyd.

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*(3), 271–279. https://doi.org/10.3102/10769986018003271

Gillett, R. (2003). The metric comparability of meta-analytic effect-size estimators from factorial designs. *Psychological Methods*, *8*, 419–433.

Hays, W. L. (1973). *Statistics* (2nd ed.). Holt, Reinhart, and Winston.

Hedges, L. V. (1981). Distribution theory for Glass' estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*, 6107–6128.

Hedges, L. V. (1982). Estimation of effect size from a series of independent experiments. *Psychological Bulletin*, *92*, 490–499.

Hedges, L. V. (2007). Effect sizes in cluster randomized designs. *Journal of Educational and Behavioral Statistics*, *32*, 341–370.

Hedges, L. V. (2011). Effect sizes in three level designs. *Journal of Educational and Behavioral Statistics*, *36*, 346–380.

Hedges, L. V. (2025). Interpretation of the standardized mean difference effects size when distributions are not normal or have unequal variances. *Educational and Psychological Measurement*, *85*, 245–257.

Hedges, L. V., & Citkowicz, M. (2015). Estimating effect size when there is clustering in one treatment group. *Behavior Research Methods*, *47*, 1295–1308. (erratum, 2018, volume 50, p.450).

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Academic Press.

Hedges, L. V., & Olkin, I. (2016). Overlap between treatment and control group distributions of an experiment as an effect size measure. *Psychological Methods*, *21*, 61–68.
Hedges, L. V., Tipton, E., Zejnullahi, R., & Diaz, K. G. (2023). Effect sizes in ANCOVA and difference-in-differences designs. *British Journal of Mathematical and Statistical Psychology*, *76*(2), 259–282.
International Committee of Medical Journal Editors. (2007). *Uniform requirements for manuscripts submitted to biomedical journals: Writing and editing for biomedical publication*. International Committee of Medical Journal Editors.
Keef, S. P., & Roberts, L. A. (2004). The meta-analysis of partial effect sizes. *British Journal of Mathematical and Statistical Psychology*, *57*, 97–129.
Kelley, K. (2007). *MBESS: Methods for the Behavioral, Educational, and Social Sciences*. R package version 0.0.8. http://CRAN.R-project.org/
Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods*, *17*(2), 137–152.
Kelly, T. L. (1935). An unbiased correlation ratio measure. *Proceedings of the National Academy of Sciences of the United States of America*, *21*, 554–559.
Kroes, A. D. A., & Finley, J. R. (2023). Demystifying omega squared: Practical guidance for effect size in common analysis of variance designs. *Psychological Methods*. https://doi.org/10.1037/met0000581
Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Science of India*, *2*, 49–55.
McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin*, *111*, 361–365.
Morris, S. B., & DeShon, R. P. (1997). Correcting effect sizes computed from factorial analysis of variance for use in meta-analysis. *Psychological Methods*, *2*, 192–199.
Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological 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*(4), 434–447. https://doi.org/10.1037/1082-989X.8.4.434
Olkin, I., & Pratt, J. W. (1958). Unbiased estimation of certain correlation coefficients. *Annals of Mathematical Statistics*, *29*, 201–211.
Pearson, K. (1905). *Mathematical contributions to the theory of evolution, XIV. On the general theory of skew correlation and non-linear correlation. Draper's company research memoirs, biometric series, II*. Dulau and Company.
Steiger, J. H. (2004). Beyond the F test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods*, *9*, 164–182.
Tilton, J. W. (1937). The measurement of overlapping. *Journal of Educational Psychology*, *28*, 656–662.
Wasserstein, R. l., & Lazar, N. L. (2016). The ASA statement on *p*-values: Context, process, and purpose. *American Statistician*, *70*, 129–133.
Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594–604.
Wishart, J. (1932). A note on the distribution of the correlation ratio. *Biometrika*, *24*, 441–446.

How to cite this article: Hedges, L. V. (2026). Effect sizes for experimental research. *British Journal of Mathematical and Statistical Psychology*, *79*, 31–45. https://doi.org/10.1111/bmsp.12389