Behavior Research Methods (2021) 53:2412–2429
https://doi.org/10.3758/s13428-021-01550-4

Expected and empirical coverages of different methods for generating noncentral *t* confidence intervals for a standardized mean difference

Douglas A. Fitts [1]

Accepted: 21 January 2021 / Published online: 12 April 2021

Abstract
Different methods have been suggested for calculating "exact" confidence intervals for a standardized mean difference using the noncentral *t* distributions. Two methods are provided in Hedges and Olkin (1985, "H") and Steiger and Fouladi (1997, "S"). Either method can be used with a biased estimator, *d*, or an unbiased estimator, *g*, of the population standardized mean difference (methods abbreviated H*d*, H*g*, S*d*, and S*g*). Coverages of each method were calculated from theory and estimated from simulations. Average coverages of 95% confidence intervals across a wide range of effect sizes and across sample sizes from 5 to 89 per group were always between 85 and 98% for all methods, and all were between 94 and 96% with sample sizes greater than 40 per group. The best interval estimation was the S*d* method, which always produced confidence intervals close to 95% at all effect sizes and sample sizes. The next best was the H*g* method, which produced consistent coverages across all effect sizes, although coverage was reduced to 93–94% at sample sizes in the range 5–15. The H*d* method was worse with small sample sizes, yielding coverages as low as 86% at $n = 5$. The S*g* method produced widely different coverages as a function of effect size when the sample size was small (93–97%). Researchers using small sample sizes are advised to use either the Steiger & Fouladi method with *d* or the Hedges & Olkin method with *g* as an interval estimation method.

Keywords Effect size · Sample size planning · Simulation

There is increasing interest in the use of confidence intervals for standardized effect sizes either as an adjunct to or as a replacement for null hypothesis statistical tests (Borenstein et al., 2009; Cumming, 2014; Cumming & Finch, 2001; Goulet-Pelletier & Cousineau, 2018; Harlow et al., 1997; Hedges & Olkin, 1985; Kelley, 2007). Confidence intervals for effect sizes are much more useful to meta-analysts than *p* values in the estimation of effect sizes across various studies, and the move away from exclusively using *p* values can help to reduce publication bias (Cumming, 2014; Ferguson & Brannick, 2012).

This article provides a detailed comparison of the coverages of two methods for creating noncentral *t* confidence intervals for a standardized mean difference. The two methods trace to Hedges (1981), Hedges and Olkin (1985) and Steiger and Fouladi (1997). The coverage of a confidence interval is the proportion of the time that confidence intervals derived from random samples will include the population parameter, and it can be calculated from the underlying distribution (expected, theoretical or predicted coverage) or estimated from simulations of large numbers of random experiments with known parameters (empirical or observed coverage). For example, in simulations, a nominal 95% confidence interval should include the population parameter in 95% of simulated experiments. Because the sample standardized mean difference, *d*, is a biased estimator of the population standardized mean difference, δ, confidence intervals have been proposed for the biased *d* itself (Cumming & Finch, 2001; Cumming, 2014; Hedges & Olkin, 1985; Kelley, 2007; Steiger & Fouladi, 1997) or for an unbiased standardized mean difference, *g* (Goulet-Pelletier & Cousineau, 2018; Hedges, 1981). When sample sizes are large, the Hedges & Olkin and Steiger & Fouladi methods generate similar confidence intervals and nominal coverage. However, the limits of the intervals generated by the two methods are different enough at small and moderate sample sizes that wide discrepancies in coverage can result. Many researchers must use small sample sizes when subjects are rare, expensive, or subject to ethical

✉ Douglas A. Fitts
dfitts@uw.edu

[1] University of Washington (Retired), Snohomish, WA 98290, USA

concerns (Fitts, 2011), and those researchers need to understand the differences between the two methods for generating confidence intervals and how the differences could affect the interpretation of their data.

Methods

Biased and unbiased standardized mean differences

If μ_1 and μ_2 are population means on a continuous variable (presumably a control population and an experimental population) and σ is some measure of a common population standard deviation, the formula for the population standardized mean difference, δ, can generally be written as follows (Cohen, 1988).

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

Substitution of sample mean values $\overline{X}_1$ and $\overline{X}_2$ for μ_1 and μ_2 and a sample standard deviation S for σ produces the sample standardized mean difference, d:

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

Although $(\overline{X}_1 - \overline{X}_2)$ is an unbiased estimator of $(\mu_1 - \mu_2)$ and the sample S^2 is an unbiased estimator of σ^2, S is not an unbiased estimator of σ because the square root function is not a linear transformation. The average S will underestimate σ, and this means that the average d calculated from Eq. 2 will overestimate δ. This bias in d can be corrected to an unbiased g (Hedges, 1981) by multiplying d by a correction factor, J (notation from Borenstein et al., 2009, rather than Hedges, 1981, who used c), that depends only on the degrees of freedom, ν:

$$g = dJ(\nu) \tag{3}$$

The formula for $J(\nu)$ uses the gamma function, Γ, as follows:

$$J(\nu) = \frac{\Gamma\left(\frac{\nu}{2}\right)}{\sqrt{\frac{\nu}{2}}\,\Gamma\left(\frac{(\nu-1)}{2}\right)} \tag{4}$$

The value of this bias function is 0.56419 for $\nu = 2$, and it rises rapidly with increasing ν toward an asymptotic upper limit of 1.0. Therefore, g will always be smaller than d in absolute value, but the difference is nontrivial only at small values of ν. For example, the value of J is already 0.95225 with only 16 degrees of freedom. An example of how to calculate $J(16)$ in the free statistical programming language R is "exp(lgamma(16/2)−(log(sqrt(16/2))+lgamma((16−1)/2)))". Using logarithms helps avoid overflow in intermediate calculations with large ν.

The calculation of d depends on the experimental design. In this article, I consider two simple experimental designs, a two-sample experiment such as a control group and an experimental group, and a one-sample experiment with two conditions for each subject, such as a pretest and a posttest. Matched subjects are regarded as a one-sample experiment because there is only one set of difference scores. Two-sample experiments assume normally distributed populations with homogeneous variances. One-sample experiments assume a normally distributed population of difference scores.

Two-sample experiments in this article always have equal sample sizes. A general formula for a pooled standard deviation of the two samples, S_P, that can also be used with unequal sample sizes is:

$$S_P = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1 + n_2 - 2}} \tag{5}$$

The d_P based on this pooled standard deviation and the degrees of freedom ν are:

$$\text{Two samples, } d_P = \frac{\overline{X}_1 - \overline{X}_2}{S_P};\ \nu = n1 + n2 - 2 \tag{6}$$

One-sample experiments are assumed to use the differences between the pretest and posttest scores of n subjects so that the d_D based on the mean, $\overline{D}$, and standard deviation, S_D, of the single set of difference scores is:

$$\text{One sample, } d_D = \frac{\overline{D}}{S_D};\ \nu = n - 1 \tag{7}$$

Either d_P or d_D can be converted to the respective unbiased g using Eq. 3.

Not considered is a one-sample test (paired samples) where the standard deviation is a pooled estimate from the two sets of scores. It is valid to calculate this d value, but trying to convert it to an unbiased estimator g or trying to construct a confidence interval for it requires the use of degrees of freedom. The correct degrees of freedom for that design is an unsettled issue, and currently published methods for confidence intervals using that design (Cumming & Finch, 2001; Goulet-Pelletier & Cousineau, 2018) are incorrect (Fitts, 2020).

Noncentral *t* distributions

The creation of confidence intervals for standardized effect sizes requires the use of noncentral t distributions. Early proposals for generating confidence intervals for standardized mean differences emphasized approximate methods that

avoided the necessity to calculate from noncentral t distributions (Borenstein et al., 2009; Hedges, 1981, 1982; Hedges & Olkin, 1985; Morris, 2000; Viechtbauer, 2007), but modern computers have eliminated the need for approximations. More recent software programs use noncentral t distributions directly (Cumming & Finch, 2001; Goulet-Pelletier & Cousineau, 2018; Kelley, 2007), and those papers review the rationale for the use of the noncentral t. Examples of how to evaluate noncentral t distributions in R and worked examples of confidence intervals are given in Supplement 1. A link to executable programs and their C source code that can generate the data presented in this article is provided in the Software section.

Briefly, the familiar central t distribution is a special case of the noncentral t distributions in which the population mean difference δ is 0 and the t distribution is perfectly symmetrical around 0. When δ is not 0, randomly sampled d values times a constant will be distributed as a noncentral t with degrees of freedom ν and population non-centrality parameter λ. When the sample sizes are unequal, the non-centrality parameter depends on the calculation of the harmonic mean, $\widetilde{n}$,

$$\widetilde{n} = 2\frac{n_1 n_2}{n_1 + n_2} \tag{8}$$

This usage of $\widetilde{n}$ is consistent with Goulet-Pelletier and Cousineau (2018) and Fitts (2020). It differs from the $\widetilde{n}$ defined by Hedges (1981), and accordingly formulas using it are slightly different.

The general formulas for population and sample noncentrality parameters are:

$$\text{Population}, \lambda_\delta = \delta\sqrt{A} \tag{9a}$$

$$\text{Sample Biased } d, \widehat{\lambda}_d = d\sqrt{A} \tag{9b}$$

$$\text{Sample Unbiased } g, \widehat{\lambda}_g = g\sqrt{A} \tag{9c}$$

where the constant A and the degrees of freedom both depend on the experimental design:

$$\text{Two samples}, A = \frac{\widetilde{n}}{2}; \nu = n1 + n2{-}2 \tag{10}$$

$$\text{One sample}, A = n;; \nu = n{-}1 \tag{11}$$

The non-centrality parameter λ_δ represents the noncentral t distribution that is the sampling distribution of $\widehat{\lambda}_d$, and that is the principal distribution of interest. All expected coverages are calculated using that distribution.

The λ_δ will always be 0 when the δ is 0, and it will always have the same sign as δ when δ is not 0. The noncentral t distributions for all λ_δ values other than 0 will always be skewed in the direction of that sign, and the degree of skewness will increase as the absolute value of λ_δ becomes larger. See Fig. 1 for examples. More comprehensive descriptions of noncentral t distributions are given in Cumming and Finch (2001), Goulet-Pelletier and Cousineau (2018), and Kelley (2007).

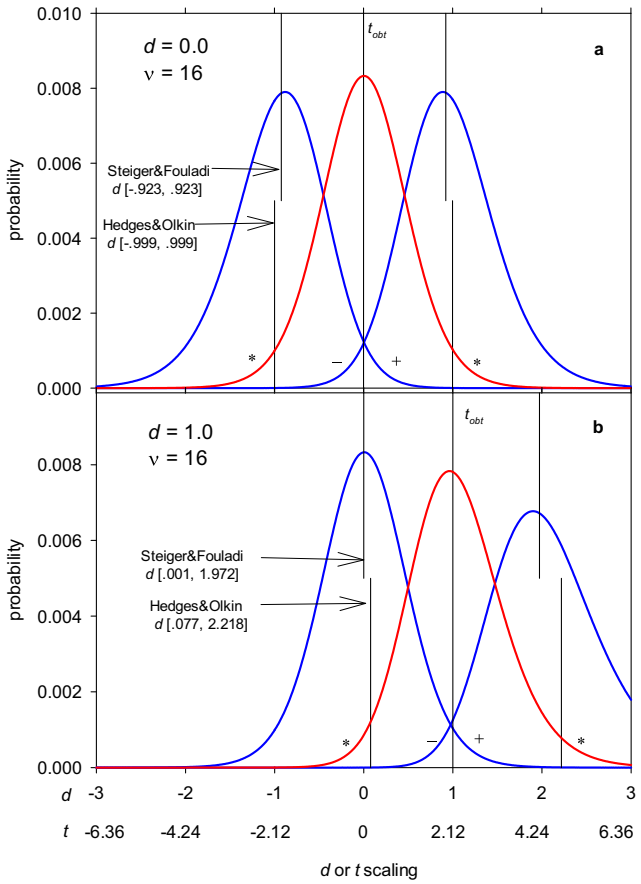

Fig. 1 Comparative illustrations of the Hedges & Olkin and Steiger & Fouladi methods for constructing a 95% noncentral t confidence interval for d. The examples use a two-tailed, two-sample test with 9 per group, ν = 16, and a sample d of either 0.0 (1a, top) or 1.0 (1b, bottom). The curves are the relevant noncentral t distributions, and both units of scaling are indicated on the abscissa (d or t). A Hedges & Olkin interval is the lower and upper 2.5% of the center (red) noncentral t distribution around obtained t (t_{obt}, tails marked with * at lower drop lines). A Steiger & Fouladi interval is the two blue non-centrality parameters of noncentral t distributions at the upper drop lines that leave 2.5% above ("+") or below ("–") t_{obt}

We define α as the complement of the confidence coefficient. For example, if we desire a 95% confidence interval, α = 1 – .95 = .05. In this article, the goal for a confidence interval is always to assign a probability of $\alpha/2$ beyond the lower and upper limits of the interval, thus dividing the uncertainty equally between the two tails. This is a simplifying assumption, not a requirement for all confidence intervals.

Hedges & Olkin and Steiger & Fouladi method confidence intervals

The Hedges & Olkin and Steiger & Fouladi methods of forming confidence intervals differ only in how they use the

noncentral t distributions to set the limits of the interval once the non-centrality parameter for the sample is known. The Hedges & Olkin (H) method (1985) uses the noncentral t distribution corresponding to $\widehat{\lambda}_d$ and its degrees of freedom to identify quantiles in the cumulative probability distribution at the probability values of $\alpha/2$ for the lower limit and $1 - \alpha/2$ for the upper limit of the confidence interval expressed as t values (Eqs. 9b and 9c):

$$\text{H}t \text{ lower and upper limits}, LL_{Ht} = t_{\frac{\alpha}{2},\nu,\widehat{\lambda}_d}; UL_{Ht} = t_{1-\frac{\alpha}{2},\nu,\widehat{\lambda}_d} \quad (12)$$

The Steiger & Fouladi method (1997) uses two different noncentral t distributions to define the lower and upper limits of the confidence interval. The non-centrality parameter of our sample d as calculated in Eq. 9 is t_{obt}. The lower t limit of the Steiger & Fouladi confidence interval, LL_{St}, is defined as the non-centrality parameter of the unique noncentral t distribution with ν degrees of freedom that has t_{obt} as the quantile at a cumulative probability of $1 - \alpha/2$. The upper t limit of the Steiger & Fouladi confidence interval, UL_{St}, is defined as the non-centrality parameter of a different noncentral t distribution with ν degrees of freedom that has t_{obt} as the quantile at a cumulative probability of $\alpha/2$. A computer search routine is recommended, and software is provided.

Both methods are illustrated in Fig. 1 for two d values, 0.0 (1a, no effect) and 1.0 (1b, an effect size equal to 1 standard deviation). The graphs assume a two-sample experiment with equal sample sizes of 9 per group ($\nu = 2(n - 1) = 16$). The desired confidence coefficient is 95%. In each graph there are three probability distributions for d generated from noncentral t distributions that were then linearly re-scaled to standard deviation units like d. Both d and t scales are given on the abscissa.

The equations for converting these lower and upper t limits to the same standardized units as d (LL_{std}, UL_{std}) requires dividing the t-scaled limits by $\sqrt{A}$ unique to the experimental design as defined in Eqs. 10 and 11. For example, LL_t can represent either LL_{Ht} or LL_{St} in Eq. 13.

$$LL_{std} = \frac{LL_t}{\sqrt{A}}; UL_{std} = \frac{UL_t}{\sqrt{A}} \quad (13)$$

Computing intervals using Cover2D.exe

Appendix 1 gives output from the provided software (see Software section at end of Methods) for the example given in Fig. 1b (see Example 1 in the user's notes). The problem is to find 95% confidence intervals for both the Hedges & Olkin and Steiger & Fouladi methods, each based on either d or g (i.e., 4 separate intervals), when $d = 1.0$ and $\nu = 16$. The program reads the instructions from the input file and writes an output text file that can be opened with a text editor or spreadsheet (see Appendix 1). The output contains the mean difference, lower and upper limits, lower and upper segments (the distance between the non-centrality parameter and the respective upper and lower confidence limits), and the full width of the interval for both the standardized biased scaling and the standardized unbiased scaling for both the Hedges & Olkin and Steiger & Fouladi methods. The d-scaled confidence intervals from Fig. 1b are given under "Standardized Biased Scaling" for Hedges & Olkin and Steiger & Fouladi confidence intervals.

Supplement 1 gives detailed example calculations for each of the four methods used in this paper.

Predicting coverages for confidence intervals

It is obvious from Fig. 1 that the Hedges & Olkin and Steiger & Fouladi methods differ, and it is important to understand how they differ in order to predict and compare the coverages of the two methods. When the d is 0.0, both methods produce symmetrical confidence intervals, and in that characteristic, they are similar to central t confidence intervals which can be expressed as a mean value plus and minus a calculated half width.

When the d is nonzero, such as 1.0, all three distributions in Fig. 1a are positively skewed, and the distances from the non-centrality parameter to the confidence limits differ on the right and left sides. I will call these unequal distances segments. With the Hedges & Olkin method, the wider segment is always on the same side as the direction of skewness; that is, a positively skewed distribution will have a wider right segment and a negatively skewed distribution will have a wider left segment. With the Steiger & Fouladi method, the reverse is true: a positively skewed distribution will have a wider left segment and a negatively skewed distribution will have a wider right segment. We can see in the example for $d = 1.0$ in Fig. 1b that the wider segment for the Hedges & Olkin method is obviously in the direction of skewness to the right, and is less obviously in the opposite direction for the Steiger & Fouladi method (i.e., 1.0 – .001 = .999 on the left and 1.972 – 1.0 = .972 on the right, see segment widths in Appendix 1). The confidence interval for a d of –1.0 would simply be the mirror image of this graph in the negative direction.

The difference in the direction of the wider segment for Hedges & Olkin and Steiger & Fouladi methods has implications for the coverages of the intervals. With effect size 0 (Fig. 1a), the symmetrical Hedges & Olkin interval is wider than the symmetrical Steiger & Fouladi interval. This implies greater coverage for the Hedges & Olkin interval with δ at or near 0. As the effect size increases to a value such as 1.0 in Fig. 1b, the Hedges & Olkin interval is wider in the direction of skewness toward the tails, where there is a smaller probability density. The Steiger & Fouladi interval is wider toward the center of the distribution, where there is a larger probability density.

Thus, a small increase in the wider segment of a Steiger & Fouladi interval will add more coverage probability than an equal increase in the wider segment of a Hedges & Olkin interval. Steiger & Fouladi intervals with moderate to large effect sizes can be narrower (more precise) and yet have higher coverage than Hedges & Olkin intervals.

Overview and method of predictions

Before proceeding to the technical details, let us examine the output of the provided software to calculate the expected coverages for all four confidence intervals using *d* or *g* and using the Hedges & Olkin or Steiger & Fouladi methods. The procedure is given in Example 3 in the user's notes and the output for 16 degrees of freedom rounded to three decimal digits is displayed in Table 1. Ignoring the intermediate calculations and examining the expected coverages in the far right column, we see that coverage for Hedges & Olkin using *d* (H*d*) is .930, Hedges & Olkin using *g* (H*g*) is .944, Steiger & Fouladi using *d* (S*d*) is .950, and Steiger & Fouladi using *g* (S*g*) is .957.

The expected coverage in Table 1 is the probability of sampling a *d* value between two boundary *d* values. These boundaries for *d* are called "leftd" and "rightd" in the text output in Table 1. These *d* values are converted to *t* values called $\widehat{\lambda}_{d,left}$ and $\widehat{\lambda}_{d,right}$ using Eq. 9b. These *t* boundaries are selected to represent the most extreme possible *d* value in that direction (left or right) that could have a confidence interval of the given type that still includes δ. That is to say, all *d* values to the left of leftd will have confidence intervals of that type that do not include δ, and all *d* values to the right of rightd will also have confidence intervals of that type that do not include δ. The probability of a $\widehat{\lambda}_d$ value to the left of $\widehat{\lambda}_{d,left}$ (called $\widehat{\alpha}_{left}$) or to the right of $\widehat{\lambda}_{d,right}$ ($\widehat{\alpha}_{right}$) can be determined directly from the cumulative distribution function of the corresponding noncentral *t* distribution with non-centrality parameter $\lambda_\delta = \delta\sqrt{A}$ and ν as given in Eqs. 10 and 11. The sum of these two tail probabilities ("ltail" and "rtail" in Table 1) is the total probability of sampling a *d* that does not include δ ($\widehat{\alpha} = \widehat{\alpha}_{left} + \widehat{\alpha}_{right}$). The complement, $1 - \widehat{\alpha}$, is the expected coverage of the confidence interval as given in the far right column in Table 1.

For H*d* and S*d*, the *d* boundaries (leftd and rightd in Table 1) are determined by a process using H*d* and S*d* confidence intervals directly. For H*g* and S*g*, the *d* boundaries (leftd and rightd) are determined by a two-step process: (1) use H*g* and S*g* confidence intervals to determine *g* boundaries, then (2) convert *g* boundaries to *d* boundaries by dividing the *g* boundaries by $J(\nu)$. For example, rightd for method H*g* is computed as rightd = rightg/$J(\nu)$ = 1.972/.95225 = 2.071. The leftg and rightg values have been calculated but are not needed for methods H*d* and S*d*.

The process for discovering these *d* boundaries involves a computer search for a confidence interval of a given type that has one of the standardized limits equal to the standardized population effect size, δ. The standardized limits of a confidence interval are determined with respect to a corresponding *t*-scaled non-centrality parameter, and the non-centrality parameter to be used depends on whether the confidence interval is a *d* interval ($\widehat{\lambda}_d$, Eq. 9b) or a *g* interval ($\widehat{\lambda}_g$, Eq. 9c). The left *d* boundary for determining the expected coverage of H*d* or S*d* confidence intervals is a *d* whose H*d* or S*d* standardized confidence interval has an upper limit of δ. The left *d* boundary for determining coverage of H*g* or S*g* confidence intervals is a *d* whose associated H*g* or S*g* standardized confidence interval has an upper limit of δ. Note that we do not rescale δ for use with a *g* interval. In either case, there can be no *d* value to the left of the selected boundary that has a confidence interval of the same type that includes δ. For the H*g* and S*g* confidence intervals, this involves converting the *d* to a *g*, searching for the *g* boundary that has an upper limit of δ, and then converting that *g* boundary back to a *d* boundary. Any *d* to the left of that *d* boundary would have a corresponding *g* that is to the left of its *g* boundary. The *d* boundary generated by a H*g* or S*g* confidence interval method will not be the same as the *d* boundary generated by a H*d* or S*d* confidence interval method.

Table 1 Sample output from Cover2D.exe using the data of Fig. 1b for a two-sample test with $\delta = 1.0$ and $\nu = 16$. Expected coverage for each of four methods (M) is listed in the far right-hand column

Samples	δ	M	ν	*n*	leftd	rightd	leftg	rightg	ltail	rtail	$\widehat{\alpha}$	coverage
2	1	H*d*	16	9	**0.001**	**1.972**	0.001	1.878	0.017	0.053	0.070	0.930
2	1	H*g*	16	9	0.001	2.071	**0.001**	**1.972**	0.017	0.039	0.056	0.944
2	1	S*d*	16	9	**0.077**	**2.218**	0.073	2.112	0.025	0.025	0.050	0.950
2	1	S*g*	16	9	0.081	2.330	**0.077**	**2.218**	0.025	0.018	0.043	0.957

The other values are intermediate data in the calculation. Leftd and rightd are the *d* boundaries for the H*d* and S*d* methods that include δ as an upper or lower confidence limit around the *d*. Leftg and rightg are the *g* boundaries for the H*g* and S*g* methods that include δ as an upper or lower confidence limit around *g*. The *g* boundaries must be converted to *d*, and then the expected coverage is the probability between those *d* boundaries calculated using the noncentral *t* sampling distribution of λ_δ. Note that the leftd and rightd values for the *d* methods are the same numbers as the leftg and rightg values for the *g* methods (numbers in **bold**)

Computation of *d* boundaries is done by a computer search algorithm, but the equivalent hand calculation could be done with much effort by trial and error. Suppose we wish to calculate the lower *d* boundary, leftd, for the S*d* method. A standardized confidence interval of the S*d* type around this *d* boundary should have an upper limit δ (i.e., the population standardized effect size), so that is our target *d*. We select an initial educated guess as to the *d* boundary, d_{guess} (it must be less than δ if the right limit will be δ), then search for a S*d* confidence interval for d_{guess} and examine the upper limit. If it is greater than δ, reduce the size of d_{guess} and try again. If it is less than δ, increase the size of d_{guess} and try again. Stop when the value of d_{guess} generates a confidence interval suitably close to δ. The computer does this to eight decimal digits. Note that similar searches are required for repeatedly computing Steiger & Fouladi confidence intervals on the way to determining that boundary, so the searching is onerous indeed. There are shortcuts to this end, discussed later, but the computer always calculates the value from this basic principle.

Pseudo-algorithm used by computer program

The procedure used by my computer program for determining the expected coverage of a *d* or *g* confidence interval for either method follows. Steps beginning "Search for..." are done using specialized search functions written for that purpose.

1. Calculate $\lambda_\delta = \delta\sqrt{A}$ and ν (Eqs. 9a to 11).
2. If finding the boundaries for H*d* or S*d* use Step 2a. If finding the boundaries for H*g* or S*g* use Step 2b. For each instance of the word "[Method]," substitute "Hedges & Olkin Method" if finding the coverage of H*d* or H*g* confidence intervals or substitute "Steiger & Fouladi Method" if finding the coverage of S*d* or S*g* confidence intervals.

2a. H*d* or S*d* procedure

   i. Search for the left *t* boundary, $\widehat{\lambda}_{d,left}$, which is the noncentrality parameter of the [Method] confidence interval to the left of $\lambda_\delta = \delta\sqrt{A}$ that includes λ_δ as the right-hand confidence limit.
   ii. Search for the right *t* boundary, $\widehat{\lambda}_{d,right}$, which is the non-centrality parameter of the [Method] confidence interval to the right of $\lambda_\delta = \delta\sqrt{A}$ that includes λ_δ as the left-hand confidence limit.

2b. H*g* or S*g* procedure

   i. Search for the left *t* boundary, $\widehat{\lambda}_{g,left}$, which is the noncentrality parameter of the [Method] confidence interval to the left of $\lambda_\delta = \delta\sqrt{A}$ that includes λ_δ as the right-hand confidence limit. Note that we do not rescale λ_δ by the bias correction factor because the population value is not biased.
   ii. Search for the right *t* boundary, $\widehat{\lambda}_{g,right}$, which is the non-centrality parameter of the [Method] confidence interval to the right of $\lambda_\delta = \delta\sqrt{A}$ that includes λ_δ as the left-hand confidence limit.
   iii. Calculate $\widehat{\lambda}_{d,left} = \widehat{\lambda}_{g,left}/J(\nu)$ and $\widehat{\lambda}_{d,right} = \widehat{\lambda}_{g,right}/J(\nu)$.

3. Using the noncentral *t* distribution for λ_δ and ν determine the cumulative probability, p_{left}, of $\widehat{\lambda}_{d,left}$ and assign $\widehat{\alpha}_{left} = p_{left}$. Do the same for the other tail to compute p_{right} based on $\widehat{\lambda}_{d,right}$ and assign $\widehat{\alpha}_{right} = 1 - p_{right}$.
4. Calculate $\widehat{\alpha} = \widehat{\alpha}_{left} + \widehat{\alpha}_{right}$.
5. Calculate expected coverage = $1 - \widehat{\alpha}$.

Supplement 1 gives examples of how to calculate expected coverages for both the Hedges & Olkin and Steiger and Fouladi methods used in this paper.

Comments and shortcuts for predicting coverage

These *d* boundaries are not confidence intervals, although they can be interpreted in a fashion similar to confidence intervals. That is, the tail probabilities represent *d* values that have confidence intervals that exclude δ, and the probabilities between the boundaries represent *d* values that have confidence intervals that include δ. However, $\widehat{\alpha}$ is a dependent variable in this process rather that an independent variable. If the process generating the confidence intervals is well behaved and follows the appropriate noncentral *t* distribution exactly, then $\widehat{\alpha}$ will equal the α that was determined a priori for the project, and the expected coverage of the interval will be the nominal coverage. We can see in Table 1 that the methods are not all perfectly well behaved with these parameters because $\widehat{\alpha}$ does not always equal .05.

The values for the boundaries in Table 1 may look familiar because they are the same as the *d*-scaled confidence limits in Fig. 1b, where $d = 1.0$ and $\nu = 16$. The confidence limits in Fig. 1b were determined for a sample *d* of 1.0, whereas the *d* boundaries in Fig. 2b are for a population δ of 1.0, but otherwise, the computation is the same. Notice that the Steiger & Fouladi confidence limits for $d = 1.0$ were [0.001, 1.972], and this is the same as the *d* boundaries in Table 1 for the H*d* method (i.e., not the Steiger & Fouladi method). The Hedges & Olkin confidence limits for $d = 1.0$ were [0.077, 2.218], and this is the same as the *d* boundaries in Table 1 for the S*d* method (i.e., not the Hedges & Olkin method). A *d* confidence interval around a value for one method produces the *d* boundaries around that value for the other method. This does not work for *g*, but see the next paragraph.

In Table 1, notice that the leftg and rightg boundaries for H*g* are the same as the leftd and rightd boundaries for H*d*, and

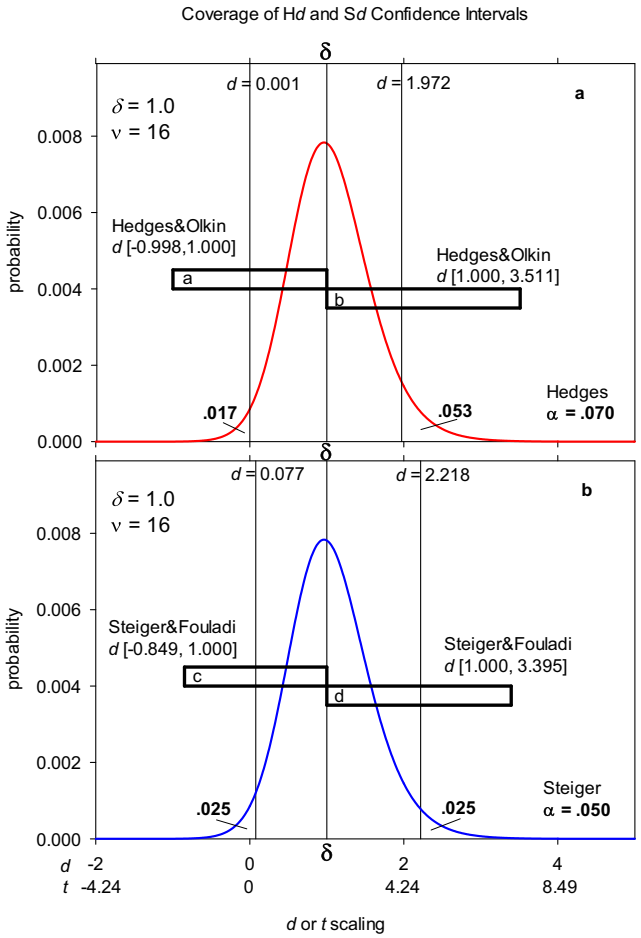

Fig. 2 Method for calculating expected coverage in H*d* and S*d* confidence intervals for a two-sample test as in Fig. 1. The curve in each panel is the sampling distribution of *d*, i.e., a noncentral *t* distribution for $\delta = 1.0$ and $\nu = 16$, specifically, $t_{2.12132,16}$. Hedges & Olkin method (2A, top): Box 'a' represents a Hedges & Olkin interval that includes δ as the upper limit and having a *d*-scaled non-centrality parameter of 0.001. Box 'b' is a Hedges & Olkin interval that includes δ as the lower limit and having a *d*-scaled non-centrality parameter of 1.972. Any *d* sampled below 0.001 or above 1.972 cannot have a confidence interval that contains δ. Steiger & Fouladi method (2B, bottom): Boxes 'c' and 'd' are Steiger & Fouladi confidence intervals that include δ as the upper and lower limits, respectively. The tails of each sampling distribution are marked with the proportion representing sampled *d* values that could not include δ. Expected coverage is the middle portion between tails

the leftg and rightg boundaries for S*g* are the same as the leftd and rightd boundaries for S*d* (values in bold type). Even though we have changed scaling for *g*, and we have consequently changed the non-centrality parameter for the sampling distribution, we have not rescaled δ. The calculations for each follow. The rightd value for method H*d* in Table 1 is 1.97214 (with two extra decimal places). The non-centrality parameter for that *d* is $1.97214\sqrt{9/2} = 4.18355$. The quantiles of the noncentral t distribution can be determined using the qt() function in R (see Supplement 1 for tips). The quantile at a cumulative probability of .025 in a noncentral *t* distribution with non-centrality parameter 4.18355 and $\nu = 16$ is qt(.025, 16, 4.18355) = 2.121321, and the *d* associated with this *t* is $2.121321/\sqrt{9/2} = 1.0000$, which matches our requirement that $\delta = 1.0000$ must be the lower limit in a Hedges & Olkin *d* standardized confidence interval (Eq. 12). In *g* scaling, if the rightg is 1.97214, its non-centrality parameter is also 4.18355 and the quantile at cumulative probability .025 is also 2.121321. The *g* associated with that *t* is $2.121321/\sqrt{9/2} = 1.0000$. With method H*d* the 1.97214 is a *d* value, but with method H*g* it is a *g* value that needs to be transformed to a *d* value as rightd = 1.97214/*J*(ν) = 1.97214/0.95225 = 2.071, which matches the rightd value for method H*g* in Table 1. The value 1.97214 was determined by independent computer searches for *d* and *g*, but they must be the same.

Thus, for a given δ, one can find the *d* boundaries to calculate the expected coverage of a H*d* confidence interval by calculating a S*d* confidence interval for δ. For example, in Appendix 1, where the confidence intervals are printed, the Hedges & Olkin standardized biased scaling confidence limits [0.077, 2.218] are the same as both (1) the *d* boundaries for the *d*-scaled Steiger & Fouladi confidence interval (S*d*), and (2) the *g* boundaries for the *g*-scaled Steiger & Fouladi confidence interval (S*g*). Similarly, one can find the *d* boundaries of a S*d* confidence interval by calculating a H*d* confidence interval for δ. In Appendix 1, the Steiger & Fouladi *d* standardized biased scaling confidence limits [0.001, 1.972] are the same as both (1) the *d* boundaries for the Hedges & Olkin confidence interval (*Hd),* and (2) the *g* boundaries for g-scaled Hedges & Olkin confidence interval (H*g*) in Table 1. Thus, one can build all of the boundaries in Table 1 from knowledge of Appendix 1. Intensive computer searching can be reduced to the computation of one Steiger & Fouladi confidence interval. The rest of Table 1 can be generated by determining the tail probabilities of the boundaries using R as described in Overview and Method of Predictions and Supplement 1. The method in the computer program Cover2D.exe computes all intervals, boundaries, and probabilities from first principles rather than relying on equivalences.

Software

Cover2D.exe (coverage for two designs) is a 32-bit console application for a PC-compatible computer. It is designed to assist researchers to create noncentral *t* confidence intervals using either the method of Hedges and Olkin (1985) or that of Steiger and Fouladi (1997) when applied to either the biased standardized mean difference (Cohen's *d*) or the unbiased standardized mean difference (Hedges' *g*). It also calculates expected coverages for each of these types of confidence intervals for a user-selected effect size over a user-selected range of sample sizes. The data for expected coverages in the figures and tables in this article can be reproduced using this program

(see examples in the user's notes). The program is not limited to any set confidence coefficient, degrees of freedom, or effect size. The data for simulated coverages in the figures and tables in this article can be reproduced using the 64-bit program Coversim.exe. Source code is provided. Files are available through the Open Science Foundation: https://osf.io/5tb7u/

Listing 1 is an R script demonstrating computation of Hedges and Olkin (1985) confidence intervals and Steiger and Fouladi (1997) confidence intervals for two independent groups in a simulation program. At the top of the script the user must set the simulation parameters *n* (number of subjects per group), *d* (population standardized effect size), *n_sim* (the number of experiments to include in the simulation), and *gamma* (the confidence coefficient). Near the bottom of the code, the user selects whether to conduct simulations using either *g* or *d* and using either Hedges & Olkin or Steiger & Fouladi method. One line for each choice must be commented to remove the line from the program (a commented line begins with "#"). Thus, to run a simulation for the Hedges & Olkin method with *g* (method H*g*), one leaves the lines "res[i, 1] <- hedgesg" and "res[i, c(2,3)] <- hedges81(hedgesg, n, gamma)" without comments and comments the lines "#res[i, 1] <- cohend" and "#res[i, c(2,3)] <- steigerfouladi97(hedgesg, n, gamma)" to remove them from the script. Then run the script in R. The run time will depend on the method selected (Steiger & Fouladi is slower) and the number of simulations in the model.

Results

Table 2 contains the expected coverages of 95% noncentral *t* confidence intervals of two-sample tests using either 8 or 16 degrees of freedom for either *d* or *g* for three population standardized effect sizes (0, 0.5, 1.0) using either the Hedges & Olkin method (rows H*d*, H*g*) or Steiger & Fouladi method (rows S*d*, S*g*). In each row, the *d* or *g* prediction boundaries are given for a relevant noncentral *t* distribution. For rows H*d* and S*d*, the prediction *d* boundaries were determined as the non-centrality parameter in standardized *d* scaling for a confidence interval of that type that had δ as the right (top line of each pair) or left (bottom line of each pair) confidence limit. For rows H*g* and S*g*, the *g* boundaries were determined as the non-centrality parameter in standardized *g* scaling for a confidence interval of that type that had δ as the right (top) or left (bottom) confidence limit. The *g* boundaries for H*g* and S*g* were converted to *d* boundaries by dividing by the bias correction factor $J(\nu)$. The *d* boundaries were used for all methods to determine tail probabilities (tailp) for that side using a noncentral *t* distribution with non-centrality parameter λ_δ and degrees of freedom ν. The two tail probabilities were summed as $\widehat{\alpha}$, and the expected coverage (Exp) was determined as $1 - \widehat{\alpha}$. For example, the *d*-scaled boundaries for the top pair of rows for $\nu = 8$ are [−1.240, 1.240]. The area of the tail is given as tailp for both the left and right sides. The two tails are summed together as $\widehat{\alpha} = .043 + .043 = .086$. The expected coverage was $1 - .086 = .914$.

In addition to expected coverages, Table 2 includes results of a Monte Carlo simulation ("Sim") of 100,000 two-sample experiments conducted with the same δ and ν from randomly sampled and normally distributed data. See Supplement 1 for the simulation method. For each experiment, separate empirical Hedges & Olkin confidence intervals for *d* and *g* were constructed (H*d*, H*g*), and separate empirical Steiger & Fouladi confidence intervals for *d* and *g* were constructed (S*d*, S*g*) from the same data. Then, for each method it was noted whether or not the interval in standardized form contained δ. Empirical coverage was calculated as the number of experiments that contained δ divided by 100,000.

Table 2 demonstrates that the expected (Exp) and simulated (Sim) coverages agree quite well. The only method that always produced actual 95% coverage of δ was the Steiger & Fouladi confidence interval for *d*. The Steiger & Fouladi confidence interval for *g* was not constant across effect sizes or sample sizes. The Hedges & Olkin confidence intervals for *d* and *g* were both lower than the nominal 95% and were not constant across effect sizes or sample sizes. In all cases, coverage was as close or closer to the nominal 95% at $\nu = 16$ than at $\nu = 8$. Any coverage that does not equal .950 means that the underlying distribution for the confidence interval of that type is not exactly a noncentral *t* distribution. Because the theoretical prediction method is based on the behavior of confidence intervals of that type, accurate predictions can be made even when the underlying distribution is not exactly a noncentral *t*.

The provided software Cover2D.exe was used to calculate expected coverages for a wide range of effect sizes for each of the four methods across sample sizes ranging from 5 to 89, and the results are displayed in Fig. 3. Simulation results for the same problems are illustrated in Figure 5 in Supplement 1. The results for the Steiger & Fouladi confidence intervals were outstanding both for their perfectly nominal coverage with *d* intervals and for their dramatically inconsistent coverage with *g* intervals. The results for the Hedges & Olkin intervals with *d* were poor because the coverage was sub-nominal with small sample sizes and was not consistent across effect sizes. The results for Hedges & Olkin intervals when used with *g* were much more consistent across effect sizes and had a much smaller window of sub-nominal coverage with small sample sizes.

The good news in Fig. 3 is that the coverage for all methods is in the neighborhood of the nominal 95% at large sample sizes. Additional calculations and simulations, not shown, demonstrate that sample sizes of 90 and higher are incrementally closer to 95% for all methods. Researchers working with small and intermediate sample sizes (5–40 per group) should use either the Steiger & Fouladi method with *d* or the Hedges

Table 2 Two-sample calculations of expected values (Exp) and simulated estimates (Sim) for coverage of the following methods (M) for generating 95% confidence intervals: Hedges & Olkin used with *d* (H*d*) or with *g* (H*g*); Steiger & Fouladi used with *d* (S*d*) or with *g* (S*g*)

		$\nu = 8$						$\nu = 16$					
		Boundaries				Coverage		Boundaries				Coverage	
δ	M	d_P	g	tailp	$\widehat{\alpha}$	Exp	Sim	d_P	g	tailp	$\widehat{\alpha}$	Exp	Sim
0	H*d*	−1.**240**	−1.118	.043	.086	.914	.914	**−0.924**	−0.880	.034	.068	.932	.933
		1.240	1.119	.043				**0.924**	0.880	.034			
	H*g*	−1.373	**−1.240**	.031	.062	.938	.938	−0.970	**−0.924**	.028	.056	.944	.944
		1.373	**1.240**	.031				0.970	**0.924**	.028			
	S*d*	**−1.459**	−1.317	.025	.050	.950	.950	**−0.999**	−0.951	.025	.050	.950	.951
		1.459	1.317	.025				**0.999**	0.951	.025			
	S*g*	−1.616	**−1.459**	.017	.034	.966	.966	−1.049	**−0.999**	.020	.041	.959	.960
		1.616	**1.459**	.017				1.049	**0.999**	.020			
0.5	H*d*	**−0.777**	−0.701	.029	.086	.914	.914	**−0.448**	−0.427	.025	.068	.932	.931
		1.748	1.578	.058				**1.432**	1.364	.044			
	H*g*	−0.861	**−0.777**	.022	.062	.938	.939	−0.470	**−0.448**	.022	.056	.944	.943
		1.936	**1.748**	.039				1.504	**1.432**	.034			
	S*d*	**−0.824**	−0.745	.025	.050	.950	.950	**−0.443**	−0.422	.025	.050	.950	.950
		2.154	1.944	.025				**1.592**	1.516	.025			
	S*g*	−0.913	**−0.824**	.019	.035	.965	.965	−0.465	**−0.443**	.023	.041	.959	.958
		2.387	**2.154**	.016				1.672	**1.592**	.019			
1	H*d*	**−0.356**	−0.321	.018	.089	.911	.912	**0.001**	0.001	.017	.070	.930	.931
		2.303	2.079	.071				**1.972**	1.878	.053			
	H*g*	−0.395	**−0.356**	.015	.061	.939	.939	0.001	**0.001**	.017	.056	.944	.944
		2.552	**2.303**	.046				2.071	**1.972**	.039			
	S*d*	**−0.253**	−0.228	.025	.050	.950	.950	**0.077**	0.073	.025	.050	.950	.950
		2.906	2.623	.025				**2.218**	2.112	.025			
	S*g*	−0.280	**−0.253**	.023	.037	.963	.962	0.081	**0.077**	.026	.043	.957	.957
		3.219	**2.906**	.015				2.330	**2.218**	.018			

Each method was generated for either $\nu = 8$ or 16 and $\delta = 0$, 0.5, or 1.0. The *d* or *g* was found as the most extreme result in that direction that would produce a confidence interval with that method that included δ. The "tailp" is the noncentral *t* probability of a *d* more extreme than the listed value for *d* on that line. The $\widehat{\alpha}$ is the sum of the two theoretical "tailp" values and the expected coverage is $1 - \widehat{\alpha}$. Simulations used 100,000 experiments and reported the empirical coverage for each method as the (number including δ)/100,000. Expected coverage is always calculated from the *d* boundary using the noncentral *t* distribution for the population δ

& Olkin method with *g*, as the other methods may produce unacceptably errant coverage. The Hedges & Olkin method used with *g* should be reported as a 94% confidence interval rather than a 95% confidence interval with the smallest sample sizes.

One-sample tests had expected coverages as listed in Table 3 with the same format as Table 2. The *d* was calculated as in Eq. 7 using the standard deviation of the difference scores, and the non-centrality parameter was calculated as in Eqs. 9 and 11. The degrees of freedom were $\nu = n - 1$. The effect sizes were the same numerically as in Table 2, but these do not represent the same absolute effect size because the standard deviation of the scores and the standard deviation of the differences are rarely the same. Again, the expected and simulated results agreed well, and the Steiger & Fouladi method used with *d* was the only method that consistently gave a 95% confidence interval at all sample sizes and effect sizes.

Expected coverages for one-sample tests with effect sizes of 0, 0.25, 0.5, 1, 1.25, and 1.5 are displayed in Fig. 4 for all sample sizes between 5 and 89. Simulation results for the same problems are illustrated in Figure 6 in Supplement 1. The Steiger & Fouladi method used with *d* remained consistently good for all effect sizes and sample sizes. The Hedges & Olkin method used with *g* had coverages at or above 94% at all sample sizes above 13, but sample sizes as small as 5 had coverages as low as 92.8%. All methods produced confidence intervals with coverages between about 94 and 96% at sample sizes above 40, but the most variable coverage was again the Steiger & Fouladi method used with *g*, which produced

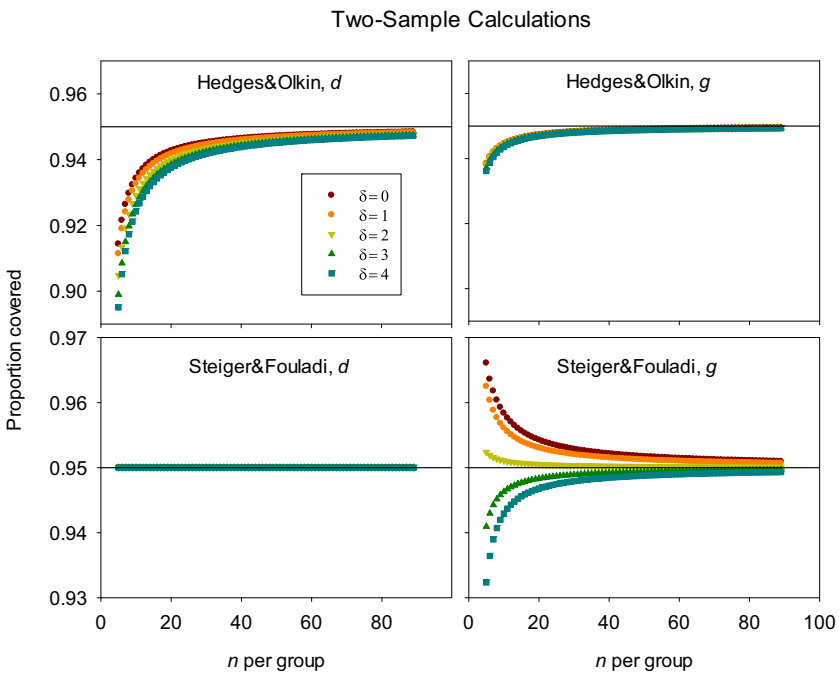

Fig. 3 Two-sample calculated (theoretical or expected) coverages of 95% confidence intervals for *d* or *g* with each of the four methods and with δ values ranging from 0 to 4 across all sample sizes from 5 to 89. Compare to simulations in Supplement 1 Fig. 5

coverages from 93.5 to 97.5% at $n = 5$ depending on the effect size.

The above analysis presents data only for 95% confidence intervals for reasons of space, but the same experiments were conducted with confidence coefficients of .90 and .99 with similar conclusions. The provided software can use other confidence coefficients.

Depending on the software in use, the calculation of probabilities for extremely large *t* values may generate computation faults. This occurs most often with tiny or huge sample sizes or huge effect sizes and can occur more often with the Steiger & Fouladi method because of the necessity to calculate probabilities for upper confidence limits that can occasionally sample a large observed *t* as the critical point at the .025 quantile. If the observed *t* is large, the non-centrality parameter at the upper Steiger & Fouladi limit will be extremely large. Software such as R and my programs rely on an algorithm by Lenth (1989) that requires the evaluation of the C code "exp(–0.5*t*t)", and this can generate faults with large *t* in some programming implementations or environments. In my two-sample tests the largest number of faults was 44 with $\delta = 4$, $n = 5$ (i.e., a huge effect size combined with a tiny sample size). The maximum that empirical coverage could have been increased if all excluded experiments actually had confidence intervals that included δ is 44/100,000 = .00116. The alternative would be to eliminate the excluded experiments from the denominator, which would artificially increase coverage instead of decreasing it. Neither would noticeably affect results in a graph.

Discussion

The Hedges & Olkin method was first proposed by Hedges (1981) and was later explicitly described as an "exact method" by Hedges & Olkin (1985, p. 91), where it was deemed too difficult for general use at the time. Hedges (1981) first presented the method for generating an unbiased standardized mean difference, *g*, but the confidence interval method presented in Hedges and Olkin (1985) was applied to the biased standardized mean difference, *d*. (The notation in these early papers differed from what is generally used today.) Steiger and Fouladi (1997) presented a different method, which they also refer to as "exact," and seemed to consider the methods identical when they refer (p. 236) to the Hedges and Olkin (1985) method as "exact," and comment that the "authors provided nomographs only for some limited cases involving very small samples." The Steiger & Fouladi method has been studied or promoted in various contexts (Algina et al., 2006; Bird, 2002; Chen & Peng, 2013; Cumming, 2014; Fidler & Thompson, 2001; Kelley, 2005; Kelley, 2007; Kelley & Rausch, 2006; Lecoutre, 2007; Smithson, 2001; Steiger, 2004; Steiger & Fouladi, 1997) and is included in the free software packages ESCI (Cumming & Finch, 2001) and MBESS (Kelley, 2007). A recent review (Goulet-Pelletier & Cousineau, 2018) noted differences between the Hedges & Olkin and the Steiger & Fouladi methods in their appendix C, where they "suspect that this [Steiger & Fouladi] method is less appropriate than the [Hedges & Olkin] noncentral method presented in this text in most, if not all, scenarios" and they recommended its abandonment. In an erratum, however, they reported an error in their calculations and rescinded that

Table 3 One-sample calculations of expected values (Exp) and simulated estimates (Sim) for coverage of the following methods (M) for generating 95% confidence intervals: Hedges & Olkin used with *d* (H*d*) or with *g* (H*g*); Steiger & Fouladi used with *d* (S*d*) or with *g* (S*g*)

		$\nu = 8$						$\nu = 16$					
						Coverage						Coverage	
δ	M	d_D	g	tailp	$\widehat{\alpha}$	Exp	Sim	d_D	g	tailp	$\widehat{\alpha}$	Exp	Sim
0	H*d*	**−0.654**	−0.590	.043	.086	.914	.914	**−0.475**	−0.452	.034	.068	.932	.933
		0.653	0.589	.043				**0.475**	0.452	.034			
	H*g*	−0.724	**−0.654**	.031	.062	.938	.938	−0.499	**−0.475**	.028	.056	.944	.945
		0.724	**0.654**	.031				0.499	**0.475**	.028			
	S*d*	**−0.769**	−0.694	.025	.050	.950	.950	**−0.514**	−0.489	.025	.050	.950	.951
		0.769	0.694	.025				**0.514**	0.489	.025			
	S*g*	−0.852	**−0.769**	.017	.034	.966	.965	−0.540	**−0.514**	.020	.041	.959	.960
		0.852	**0.769**	.017				0.540	**0.514**	.020			
0.5	H*d*	**−0.210**	−0.190	.018	.088	.912	.913	**−0.013**	−0.012	.017	.069	.931	.929
		1.183	1.068	.070				**0.999**	0.951	.052			
	H*g*	−0.233	**−0.210**	.016	.061	.939	.940	−0.013	**−0.012**	.017	.056	.944	.943
		1.310	**1.183**	.046				1.049	**0.999**	.039			
	S*d*	**−0.163**	−0.147	.025	.050	.950	.950	**0.025**	0.024	.025	.050	.950	.949
		1.490	1.345	.025				**1.123**	1.069	.025			
	S*g*	−0.181	**−0.163**	.022	.037	.963	.963	0.026	**0.025**	.025	.043	.957	.956
		1.650	**1.490**	.015				1.179	**1.123**	.018			
1	H*d*	**0.169**	0.153	.006	.094	.906	.904	**0.403**	0.384	.009	.073	.927	.926
		1.792	1.618	.088				**1.576**	1.501	.065			
	H*g*	0.187	**0.169**	.007	.062	.938	.938	0.423	**0.403**	.011	.056	.944	.943
		1.985	**1.792**	.054				1.655	**1.576**	.045			
	S*d*	**0.337**	0.304	.025	.050	.950	.950	**0.501**	0.477	.025	.050	.950	.950
		2.300	2.076	.025				**1.787**	1.702	.025			
	S*g*	0.373	**0.337**	.033	.046	.954	.953	0.527	**0.502**	.032	.049	.951	.952
		2.548	**2.300**	.014				1.876	**1.786**	.017			

Each method was generated for either $\nu = 8$ or 16 and $\delta = 0$, 0.5, or 1.0. The *d* or *g* was found as the most extreme result in that direction that would produce a confidence interval with that method that included δ. After converting *g* to *d*, the "tailp" is the noncentral *t* probability of a *d* more extreme than the listed value for *d* in the sampling distribution of *d* with non-centrality parameter $\sqrt{\frac{n}{2}}$. The $\widehat{\alpha}$ is the sum of the two theoretical "tailp" values and the expected coverage is $1 - \widehat{\alpha}$. Simulations used 100,000 experiments and reported the empirical coverage for each method as the (number including δ)/100,000

recommendation (Goulet-Pelletier & Cousineau, 2020). They did not provide extensive data on coverage to compare the methods.

The coverages as calculated and simulated in this study support the use of both methods in different contexts. The Steiger & Fouladi method used with the biased *d* produced excellent coverages for all sample sizes and effect sizes for both one- and two-sample tests. The Hedges & Olkin method, when used with unbiased *g*, also produced highly consistent coverages across all effect sizes, and the coverage was close to nominal for sample sizes greater than or equal to about 20. At sample sizes of 5–20, the coverage of the Hedges & Olkin method with *g* was consistently reduced on average to a nadir of about 94% (two-sample) or 93% (one-sample).

The Hedges & Olkin method with *d* and the Steiger & Fouladi method with *g* produce discrepant coverages with different effect sizes at small sample sizes and probably should not be used at all for sample sizes less than about 40 per group.

Precision in a confidence interval is determined by the width of the interval (Kelley, 2007; Kelley & Rausch, 2006). Estimation of the parameter is more precise if the interval is narrow instead of wide. In general, intervals formed with *g* are narrower than intervals formed with *d*, and intervals formed with the Steiger & Fouladi method are narrower than intervals formed with the Hedges & Olkin method. Of the methods studied here, the Steiger & Fouladi method with *d* creates confidence intervals that are always close to the

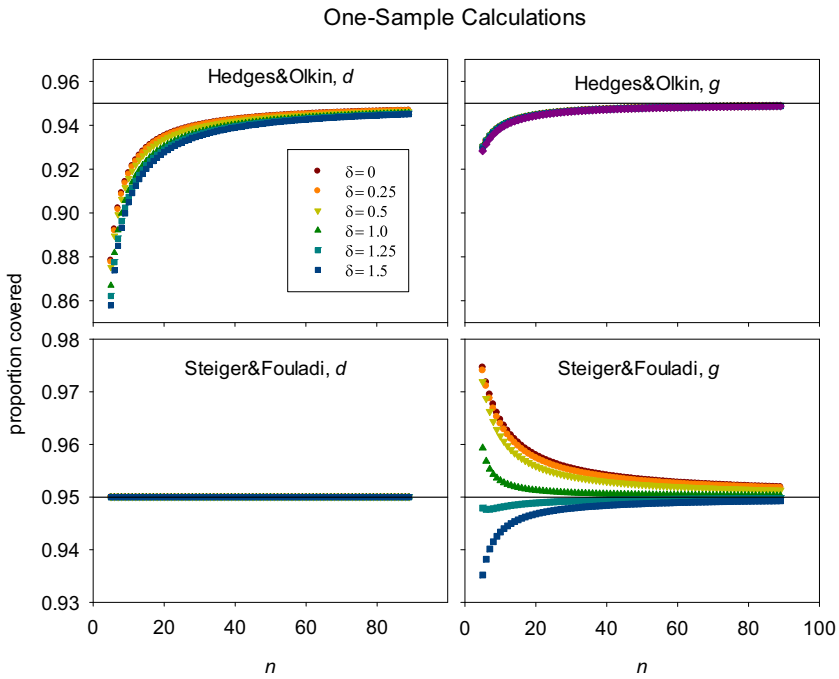

Fig. 4 One-sample calculated (theoretical or expected) coverages of 95% confidence intervals for *d* or *g* with each of the four methods and with δ values ranging from 0 to 1.5 across all sample sizes from 5 to 89. Compare to simulations in Supplement Fig. 6

nominal confidence coefficient and are generally slightly more compact even than those generated with the Hedges & Olkin method with *g*. Although a point estimator such as *g* provides a more accurate estimate of δ than *d*, the most accurate and compact confidence interval is obtained with S*d*, not S*g*. Therefore, it is not possible to use either a *d* or *g* method and have both the most accurate point estimate and also the most compact confidence interval with exact coverage. Those needing an accurate point estimate should report *g*. Those needing the most precise confidence interval should report S*d*. Those who must report both within either a single *d* or *g* context should use *d* and S*d* or *g* and H*g*.

In the original draft of this manuscript I suggested reporting *g* as the best point estimator and also reporting the Steiger & Fouladi *d* confidence interval (S*d*) as the best interval estimator. One reviewer strongly disagreed with that suggestion because of the possibility that using different systems for point and interval estimation could result in intervals being reported that do not contain the point estimate. The issue is debatable, but it raises an interesting question: "What is a biased or unbiased interval estimate?" If we put the method used to generate the interval in a black box and look only at the performance of the method, the Steiger & Fouladi method with *d* clearly is the best choice for generating narrow confidence intervals that include δ a nominal percentage of the time. Is that interval "biased" because it is using *d* instead of *g* to generate the interval? The answer is clearly no, because the interval is not a parameter that can be biased or unbiased. Kelley & Rausch (2006, Footnote 5, p. 365) list several circumstances, including the standardized mean difference, where it is best to report the unbiased point estimate alongside an interval that is based on a biased point estimate.

An example of this is contained in the publication by Viechtbauer (2007), who examined 21 potential approximation formulas for confidence intervals for standardized effect sizes and compared their coverage and compactness with that of exact confidence intervals. All approximations used some version of a biased or unbiased estimate of the effect size and the variance of the noncentral *t* distribution and either a central normal distribution or a central *t* distribution to approximate the noncentral *t* distribution. The closed-form approximations offered the advantage that they are not iterative and are therefore easier to calculate. The paper cites both Hedges and Olkin (1985) and Steiger and Fouladi (1997) for exact confidence intervals, but the author does not state what variation of the "exact" interval he used as the standard for comparison. As seen here, the Hedges & Olkin procedure is not iterative. Judging from example data and results presented on pages 48–49 of Viechtbauer (2007), the "exact" method was the Steiger and Fouladi method with the biased *d*, abbreviated S*d* here. The results presented for that method listed *g* as the unbiased point estimate and the Steiger method calculated with the biased *d* as the exact interval. Given the results of the present paper, the S*d* method was indeed the most exact of the methods, so the choice was appropriate. This is an example of using the unbiased *g* for the point estimate and the biased *d* to compute the interval. My recommendations in this paper are to use this iterative exact and compact S*d* method instead of an approximation. If a non-iterative approach is required for some reason, the H*g* method may be substituted to obtain similar coverage except at the smallest sample sizes

and at the expense of a slightly wider interval. If an approximation technique is preferred for some reason, the reader is referred to the recommendations of Viechtbauer (2007) for the best techniques.

In a similar vein, if we use a confidence interval protocol that employs the number .95 inside the black box, and if that protocol has an expected and empirical coverage of .944, why should we call it a 95% confidence interval? The confidence coefficient should be named for the actual coverage of the method, regardless of how the method is calculated. The method presented in this paper now allows us to predict the actual coverage of the protocol. Thus, it is entirely appropriate to call a Hedges & Olkin *g*-scaled confidence interval (H*g*) with $\nu = 16$ in Tables 2 and 3 a 94.4% confidence interval because the expected coverage is 94.4%. In so doing, a researcher runs a risk of confusing readers into thinking that the α value used in the calculations was .056 instead of .05 unless it is made perfectly clear in the publication that the naming of the confidence coefficient is based on the expected coverage rather than the α used in the calculations. With a small effort, I used Cover2D.exe to identify a nominal confidence coefficient of .957 that generated an actual coverage of .950 with $\nu = 16$ for method H*g* in a two-sample test. This latter should be called a 95% confidence interval, not a 95.7% confidence interval, but the researcher needs to clarify how it was done. This is done in other fields when a confidence interval protocol predictably produces observed coverage other than the nominal confidence coefficient (Singham, 2014, section 5.2 Choosing a Confidence Coefficient). The language for doing this obviously needs to be clarified (Singham used different notation for the confidence coefficient employed in the computations and the confidence coefficient representing the actual coverage).

Given the above paragraph, either the Steiger & Fouladi intervals with *d* (S*d*) or the Hedges & Olkin intervals with *g* (H*g*) produce excellent intervals where the actual coverage of the interval is known a priori and does not vary markedly according to effect size. The Hedges & Olkin method with *d* (H*d*) creates poor intervals because the coverage is sub-nominal to different degrees for different effect sizes. Some researchers may consider the Steiger & Fouladi *g* method (S*g*) to produce excellent intervals if one's definition of excellent coverage is an interval that includes the parameter at least a nominal proportion of the time. A *g* value of 3 or 4 is almost unheard of in practical research in psychology, and those are the circumstances where the coverage was sub-nominal. However, researchers who must use small samples because subjects are rare, expensive, or subject to ethical limitations would consider a confidence interval protocol that generated 97% coverage instead of the nominal 95% coverage to be a waste of subjects. These researchers need a confidence interval protocol that generates a known coverage with a known sample size even if the effect size is unknown. As sample size management matures for these protocols (Kelley & Rausch, 2006), researchers will want to know the minimum sample size that produces a confidence interval with a known precision, and that will be best with the S*d* or H*g* methods.

Goulet-Pelletier and Cousineau (2018) among others have called for researchers to report effect sizes for single sample or paired sample experiments using the pooled standard deviation S_P as the denominator for *d* because the effect size will then be directly comparable to standardized effect sizes for two-sample experiments. In this paper, the standardized effect size of 1.0 was not the same in the two-sample experiments as it was in the one-sample experiments using S_D as the denominator. The problem is that a computation of *g* from *d* or a computation of a confidence interval for *d* when using S_P in a one-sample test requires a computation of the degrees of freedom, and the exact degrees of freedom are some function of the population correlation, ρ, that is currently unknown (Fitts, 2020). I have generated an approximate solution to the degrees of freedom to be published elsewhere, but even with more accurate degrees of freedom the empirical coverage is not perfect because of the necessity to substitute a random variable *r* for ρ in the prediction equation. By contrast, the effect sizes and confidence intervals for a one-sample test with S_D in the denominator are exact, and this is the only way currently to compare confidence intervals for different experiments using one-sample tests.

It is important to mention that the studies presented here employed ideal conditions for generating consistent confidence intervals. Real data may often be drawn from non-normal distributions or from populations with different variances, and this can dramatically affect coverage (Kelley, 2005). When the treatment affects the variance, a standardized effect size may be more appropriately reported with the standard deviation of the control group in the denominator rather than a pooled standard deviation in Eq. 2 (Glass, 1976; Hedges, 1981; Morris, 2000, for repeated measures). The same approach could be used if sample sizes are unequal. The techniques used here with a pooled error term will probably generalize to pooled error terms from unequal sample sizes as long as the sizes are roughly the same. A complete study of unequal sample sizes must also consider the consequences of unequal variances (Fitts, 2010), and that complicated topic is beyond the scope of the current article.

Statistical software is becoming available to generate these confidence intervals, but one must be careful to inspect the notation and the exact method used to be sure the software returns the intended confidence interval. Language such as a "noncentral *t* confidence interval" is ambiguous. Kelley's (2007) MBESS uses the Steiger & Fouladi method with *d*, but only for independent groups analyses. Cumming and Finch (2001) also used the Steiger & Fouladi method with *d* and include repeated measures as well as independent groups; however, the degrees of freedom for the repeated measures

intervals are wrong, as just discussed. The current versions of these software may differ. By contrast, the Hedges & Olkin *g*-scaled confidence intervals are fairly easy to calculate without specialized software other than the qt() function in R and a table of values of the bias coefficient *J* (Hedges, 1981). Whatever the software or method of computation, researchers should report exactly the method used in publications.

Listing 1

Listing 1 R script demonstrating computation of Hedges and Olkin (1985) confidence intervals and Steiger and Fouladi (1997) confidence intervals in a simulation program. The search program for the Steiger & Fouladi method is a part of the MBESS package (Kelley, 2007). Author: D. Cousineau.

```
library(mvtnorm) # for rmvnorm
library(MBESS) # for conf.limits.nct
library(psych) # for t.test
# Simulation parameters
n <- 9 #sample size
d <- 1.00 #population true effect size
n_sim <- 10000
gamma <- 0.95
J <- function(df) {
#compute unbiasing factor
exp(lgamma(df/2)-(log(sqrt(df/2))+lgamma((df-1)/2)))
}
showresults <- function(res, d, n) {
# takes a matrix with estimate in col. 1 and bounds in cols 2&3
cat(
sprintf('True parameters: d = %5.3f, n = %1d\n', d, n),
sprintf('Mean statistic: = %5.3f\n', mean(res[,1]) ),
sprintf('Non-rejection = %.3f\n', mean(d > res[,2] & d < res[,3]) ),
sprintf('Rejection left = %.3f\n', mean(d < res[,2])),
sprintf('Rejection right = %.3f\n', mean(d > res[,3])),
sprintf('Mean lower bound = %.3f\n', mean(res[,2])),
sprintf('Mean upper bound = %.3f\n', mean(res[,3])),
sprintf('Mean width = %.3f\n', mean(res[,3]-res[,2]))
```

```
)
}
steigerfouladi97 <- function(es, n, gamma = .95) {
#compute noncentrality parameter
lambda <- es * sqrt( n/2 )
# METHOD of Steiger and Fouladi, 1997###
tCI <- conf.limits.nct(lambda, 2*(n-1), conf.level = gamma)
tCI.low <- tCI$Lower.Limit
tCI.hig <- tCI$Upper.Limit
limits <- c(tCI.low, tCI.hig) / sqrt(n/2)
limits
}
hedges81 <- function(es, n, gamma = .95) {
#compute noncentrality parameter
lambda <- es * sqrt( n/2 )
#confidence interval of hedges g
tlow <- qt(1/2 - gamma/2, df = 2*(n-1), ncp = lambda )
thig <- qt(1/2 + gamma/2, df = 2*(n-1), ncp = lambda )
limits <- c(tlow,thig) / sqrt(n/2)
limits
}
# Create covariance matrix with no correlation
cov_matrix <- array(c(1, 0, 0, 1), dim=c(2,2))
# Preallocate a result matrix containing left and right boundaries for each sim
res <- matrix(data=NaN, nrow=n_sim, ncol=3)
# Run the simulations
for (i in 1:n_sim) {
# Generate random sample from two normal distribution
X <- rmvnorm(n, mean=c(d/2, -d/2), sigma=cov_matrix)
```

```
# Get descriptives

Mx <- mean(X[,1])

My <- mean(X[,2])

sx <- sd(X[,1])

sy <- sd(X[,2])

#get pairwise statistics Delta means and pooled SD

dmn <- Mx-My

sdp <- sqrt((sx^2 + sy^2)/2)

#compute biased Cohen's d and unbiased hedges' g

cohend <- dmn / sdp

hedgesg <- cohend * J(2*(n-1))

# to your choice, select which function to use on what estimator

res[i, 1] <- hedgesg

#res[i, 1] <- cohend

res[i, c(2,3) ] <- hedges81(hedgesg, n, gamma)

#res[i, c(2,3) ] <- steigerfouladi97(hedgesg, n, gamma)

}

# Shows the results

showresults(res, d, n)
```

Appendix

Appendix 1. Output of Cover2D.exe in CI_MODE. Compare to Fig. 1b.

Mode: CI_MODE	
N of Samples 2	
Two-tailed Alpha	0.050000
SAMPLE_MEANDIFF	1.000000
SAMPLE_SD	1.000000
SAMPLE_n	9.000000
Degrees of freedom (v)	16.000000
Biased d	1.000000
Variance of d	0.294031
Bias coefficient J(v)	0.952254
Unbiased g	0.952254
Variance of g	0.266624

HEDGES&OOLKIN 95.0% NONCENTRAL t CONFIDENCE INTERVAL

Standardized Biased Scaling	
Mean difference d	1.000000
Lower Limit of d	0.076877
Upper Limit of d	2.218295
Lower Segment	0.923123
Upper Segment	1.218295
Full Width	2.141418
Standardized Unbiased Scaling	

Mean difference g	0.952254
Lower Limit of g	0.028710
Upper Limit of g	2.157190
Lower Segment	0.923544
Upper Segment	1.204936
Full Width	2.128480

STEIGER&FOULADI 95.0% NONCENTRAL t CONFIDENCE INTERVAL

Standardized Biased Scaling	
Mean difference d	1.000000
Lower Limit of d	0.000580
Upper Limit of d	1.972144
Lower Segment	0.999420
Upper Segment	0.972144
Full Width	1.971564
Standardized Unbiased Scaling	
Mean difference g	0.952254
Lower Limit of g	–0.041062
Upper Limit of g	1.919303
Lower Segment	0.993316
Upper Segment	0.967049
Full Width	1.960365

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13428-021-01550-4.

Acknowledgements Listing 1 was generously provided by Denis Cousineau. The only change was to insert the log equation for *J* so an error would not be generated with large degrees of freedom.

Declaration

Conflicts of interest The author has no conflicts of interest to report.

References

Algina, J., Keselman, H.J., & Penfield, R.D. (2006). Confidence interval coverage for Cohen's effect size statistic. *Educational and Psychological Measurement*, 66, 945–960. https://doi.org/10.1177/0013164406288161.

Bird, K.D. (2002) Confidence intervals for effect sizes in analysis of variance. *Educational and Psychological Measurement*, 62, 197–226

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. John Wiley & Sons

Chen, L.-T. & Peng, C.-Y. J. (2013) Constructing Confidence Intervals for Effect Sizes in ANOVA Designs. *Journal of Modern Applied Statistical Methods*, 12(2), Article 5. https://doi.org/10.22237/jmasm/1383278640.

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

Cumming, G. (2014) The new statistics: Why and how. *Psychological Science*, 25, 7–29. https://doi.org/10.1177/0956797613504966.

Cumming, G., & Finch, S. (2001). A primer on the understanding, use and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement*, 61, 532–574. https://doi.org/10.1177/0013164401614002.

Ferguson, C.J., & Brannick, M.T. (2012). Publication bias in psychological science: Prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychological Methods*, 17, 120–128. https://doi.org/10.1037/a0024445.

Fidler, F., & Thompson, B. (2001). Computing correct confidence intervals for ANOVA fixed- and random-effects effect sizes. *Educational and Psychological Measurement*, 61, 575–604

Fitts, D.A. (2010). The variable-criteria sequential stopping rule: Generality to unequal sample sizes, unequal variances, or to large ANOVAs. *Behavior Research Methods*, 42, 918–929

Fitts, D.A. (2011). Ethics and animal numbers: informal analyses, uncertain sample sizes, inefficient replications, and type I errors. *Journal of the American Association of Laboratory Animal Science*, 50, 445–453

Fitts, D.A. (2020). Commentary on "A review of effect sizes and their confidence intervals, part I: The Cohen's d family"**:** The degrees of freedom for a paired samples design. *The Quantitative Methods for Psychology*, 16, 281–294. https://doi.org/10.20982/tqmp.16.4.p281.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, 5, 3–8

Goulet-Pelletier, J.-C., & Cousineau, D. (2018). A review of effect sizes and their confidence intervals, part I: The Cohen's d family. *The Quantitative Methods for Psychology*, 14, 242–265. https://doi.org/10.20982/tqmp.14.4.p242.

Goulet-Pelletier, J.-C., & Cousineau, D. (2020). Erratum to Appendix C of "A review of effect sizes and their confidence intervals, Part I: The Cohen's d family". *The Quantitative Methods for Psychology*, 16(4), 422–423. https://doi.org/10.20982/tqmp.16.4.p422.

Harlow, L.L., Mulaik, S.A, & Steiger, J. H. (Eds.), (1997). *What if There Were no Significance Tests?* Lawrence Erlbaum Associates

Hedges, L.V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, 6, 107–128

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

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

Kelley, K. (2005). The effects of nonnormal distributions on confidence intervals around the standardized mean difference: bootstrap and parametric confidence intervals. *Educational and Psychological Measurement*, 65, 51–69. https://doi.org/10.1177/0013164404264850.

Kelley, K. (2007). Confidence intervals for standardized effect sizes: Theory, application, and implementation. *Journal of Statistical Software*, 20, 1–24

Kelley, K., & Rausch, J. R. (2006). Sample size planning for the standardized mean difference: Accuracy in parameter estimation via narrow confidence intervals. *Psychological Methods*, 11, 363–385. https://doi.org/10.1037/1082-989X.11.4.363.

Lecoutre, B. (2007) Another look at the confidence intervals for the noncentral *t* distribution. *Journal of Applied Statistical Methods*, 6(1), Article 11. https://doi.org/10.22237/jmasm/1177992600.

Lenth, R. (1989). Algorithm AS 243: Cumulative Distribution Function of the Non-Central T Distribution. *Applied Statistics*, 38, 185–189

Morris, S.B. (2000). Distribution of the standardized mean change effect size for meta-analysis on repeated measures. *British Journal of Mathematical and Statistical Psychology*, 53, 17–29

Singham, D. I. (2014). Selecting stopping rules for confidence interval procedures. *ACM Transactions on Modeling and Computer*

Simulation, 24(3) Article 18, 18 pages. https://doi.org/10.1145/2627734.

Smithson, M. (2001) Correct confidence intervals for various regression effect sizes and parameters: The importance of noncentral distributions in computing intervals. *Educational and Psychological Measurement*, 61, 605–632

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. https://doi.org/10.1037/1082-989X.9.2.164.

Steiger, J.H, & Fouladi, R.T. (1997). Noncentrality interval estimation and the evaluation of statistical methods. In L.L. Harlow, S.A, Mulaik, & J.H. Steiger (Eds.), *What if There Were no Significance Tests?* (pp. 221–257). Mahwah: Lawrence Erlbaum Associates.

Viechtbauer, W. (2007). Approximate confidence intervals for standardized effect sizes in the two-independent and two-dependent samples design. *Journal of Educational and Behavioral Statistics*, 32, 39–60. https://doi.org/10.3102/1076998606298034.