2020 ■ Vol. 16 ■ no. 4

Commentary on "A review of effect sizes and their confidence intervals, Part I: The Cohen's d family": The degrees of freedom for paired samples designs.

Douglas A. Fitts [a]

[a]University of Washington

Abstract ■ In their review of effect sizes of the Cohen's d family, Goulet-Pelletier and Cousineau (2018) proposed several methods for generating confidence intervals for the unbiased standardized mean difference, g. Among them they proposed using degrees of freedom $\nu = 2(n-1)$ instead of $\nu = (n-1)$ for all paired samples designs that use a pooled standard deviation to standardize the mean difference (pooled paired samples) when calculating g and its confidence limits from a noncentral t distribution. Simulations demonstrate that the exact ν for a pooled paired samples design vary as a function of the population correlation ρ between $2(n-1)$ at $\rho = .0$ and $(n-1)$ at $\rho = 1.0$. This affects the calculation of g and the selection of the appropriate noncentral t distribution for calculating the confidence limits. Using a sample r to estimate the unknown ρ causes a further deviation from the presumed noncentral t distribution even when the ν are known. These facts adversely affect the coverage of the confidence intervals computed as recommended by the authors. These methods for calculating noncentral t confidence intervals should not be used as described with pooled paired samples designs. Noncentral t confidence intervals for either a two sample design or a paired samples design where the mean difference is standardized by the standard deviation of the difference scores are unaffected by this problem. An R script and C source code are provided.

Keywords ■ noncentral t distributions, parameter estimation, simulation. **Tools** ■ R, C.

✉ dfitts@uw.edu

10.20982/tqmp.16.4.p281

Acting Editor ■ Roland Pfister (Universität Würzburg)

Reviewers
■ Daniel Lakens (Technische Universiteit Eindhoven, Universiteit Leiden)
■ Denis Cousineau (Université d'Ottawa)
■ Jean-Christophe Goulet-Pelletier (Université d'Ottawa)

Introduction

Goulet-Pelletier and Cousineau (2018) proposed several methods for generating confidence intervals for the unbiased standardized mean difference, g. This paper critically evaluates some of those methods. As described in their review, a confidence interval for an unbiased standardized effect size should use a noncentral t distribution to determine the limits. Their paper states that the degrees of freedom, ν, for any paired samples design using a pooled error term (pooled paired samples design) should be $\nu = 2(n-1)$ to calculate g and its variance, $\mathrm{Var}(g)$, where n is the number of pairs of scores. This contrasts with the $\nu = (n-1)$ used with a paired t test. The present simulations followed their methods as proposed and test whether $2(n-1)$ or $(n-1)$ provide the proper coverage for a 95% confidence interval across a range of effect sizes, sample sizes, and correlations.

A paired samples design can be any design with two repeated measures such as a pretest and posttest in a single group of subjects or two groups of subjects that have been matched into pairs on a different but positively correlated variable. I have not considered negative values of ρ because few researchers would set out to use a paired measures design when the correlation is not expected to be positive. For example, subjects are never matched based on a negatively correlated matching variable. For convenience below, formulas are described as a pretest and posttest in a single group of subjects.

Methods and Results

Simulations employed a C language program that generated pseudorandom data from a normal-bivariate distribution (see Fitts, 2018, C code and documentation available at: https://osf.io/jdtcz/). The paired samples model was: x_{pre}, $i \sim N(\mu_{pre}, \sigma^2)$, x_{post}, $i \sim N(\mu_{post}, \sigma^2)$, $i = 1, \ldots, n$, Correlation(x_{pre}, x_{post}) = ρ. The population standard deviation was $\sigma = 1.0$, and the size of the sample, n, was varied. A general standardized mean difference, d, is a mean difference divided by a standard deviation. In the present simulations, the population (δ_P) and sample (d_{P}) standardized mean differences were, respectively:

$$\delta_P = \frac{\mu_{pre} - \mu_{post}}{\sigma_P}; \tag{1a}$$

$$d_{\mathrm{P}} = \frac{\bar{X}_{pre} - \bar{X}_{post}}{S_P} \tag{1b}$$

with the pooled standard deviation as proposed by Goulet-Pelletier and Cousineau (2018) as:

$$S_P = \frac{S_{pre} + S_{post}}{2} \tag{2}$$

and degrees of freedom for calculating g and its variance as recommended by Goulet-Pelletier and Cousineau (2018) as:

$$\nu = n_{pre} + n_{post} - 2 = 2(n-1). \tag{3}$$

In separate sets of simulations I also used $\nu = (n-1)$ for comparison with the recommended method.

Hedges (1981) demonstrated that the sampling distribution of d times a constant is a noncentral t with appropriate degrees of freedom and non-centrality parameter. The non-centrality parameter λ for a general noncentral t distribution corresponding to δ is:

$$\lambda = \delta\sqrt{A}; \hat{\lambda} = d\sqrt{A} \tag{4}$$

and the general equation for the variance of d (Var(d)) is:

$$Var\,(d) = \left(\frac{1}{A}\right)\frac{\nu}{\nu-2}\left(1 + (A)\,\delta^2\right) - \frac{\delta^2}{J(\nu)^2}. \tag{5}$$

In these equations, δ can be estimated by d, and A varies according to experimental design:

$$\text{Pooled paired samples: } A = \left(\frac{n}{2\,(1-\rho)}\right) \tag{6a}$$

$$\text{Difference paired samples: } A = (n) \tag{6b}$$

$$\text{Two groups: } A = (\frac{\tilde{n}}{2}); \tilde{n} = 2\frac{n_1 n_2}{n_1 + n_2} \tag{6c}$$

and $J(\nu)$ is the correction for bias:

$$J\,(\nu) = \frac{\Gamma\left(\frac{\nu}{2}\right)}{\sqrt{\frac{\nu}{2}}\,\Gamma\left(\frac{(\nu-1)}{2}\right)} \tag{7}$$

owing to Hedges (1981).

The unbiased standardized mean difference, g, is calculated as:

$$g = d \times J(\nu); \tag{8a}$$

$$Var(g) = Var(d)J(\nu)^2 \tag{8b}$$

Empirical Coverage of Confidence Interval and Variance

A 95% confidence interval was calculated for each of 10,000 simulated experiments, and for each it was noted whether or not the interval included the population standardized mean difference δ_P. Coverage was calculated as the number of intervals that included δ_P divided by 10,000. Independent simulations were conducted for $\delta = 0$, 0.5, and 1.0 and for $\rho = 0$, 0.45, and 0.90.

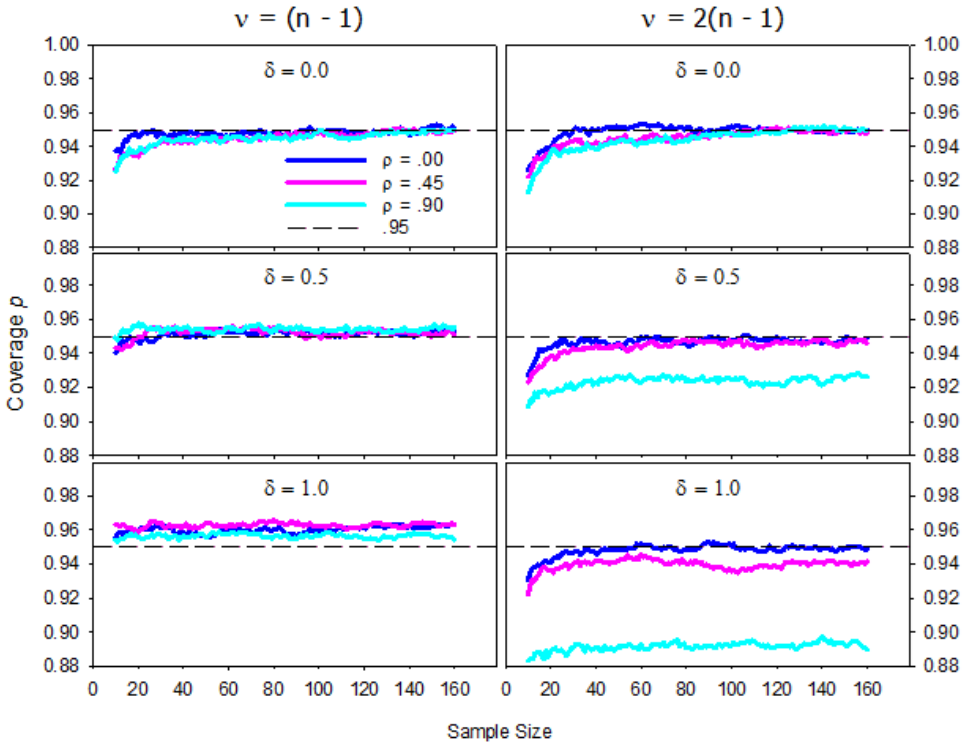
I simulated sequential experiments beginning with $n = 10$, added 1 sample at each iteration, and summarized results at all sample sizes to a fairly large number. The code was borrowed from a sequential sampling simulator (Fitts, 2018). See Appendix A for reliability information.

These simulations employed a noncentral t cumulative probability algorithm (ASA243) written originally in `Fortran77` by Lenth (Lenth, 1989) and converted to C by Burkhardt. This algorithm is used by R. I wrote a binary search function that performs the same task as the qt() function in R. This function performs a binary search using ASA243 with different t values until a quantile is found that matches a target probability value to eight decimal digits.

Figure 1 illustrates the coverages for pooled paired samples tests in two parts, $\nu = (n-1)$ (left) and $\nu = 2(n-1)$ (right). Coverages with $(n-1)$ were not far from the expected .95 although coverage did increase toward .96 as the effect size increased. Coverage was low at small sample sizes with $\delta = 0$. Coverage with $2(n-1)$ was reduced with increasing values of δ and ρ and was unacceptable with $\rho = .9$. It was also too low with small sample sizes. Please note in the bottom panels of Figure 1 for $\delta = 1.0$, where the variability between curves is greatest, that the curve for $\rho = .9$ is closest to .95 on the left side (with $\nu = (n-1)$) and the curve for $\rho = 0$ is closest to .95 on the right side (for $\nu = 2(n-1)$).

I ran additional simulations to explore the deviance of the empirical variance from the theoretical variance with pooled paired samples tests. If a noncentral t distribution with $\nu = n_{pre} + n_{post} - 2$ is a good model of the sampling distribution of $d_{\mathrm{P}}\sqrt{n/(2\,(1-\rho))}$, then the variance of empirical d_{P} values from a simulation should match the theoretical variance calculated from Equation 5. The d_{P}

Figure 1 ■ Pooled paired samples design. Coverage of 95% noncentral t confidence intervals for the unbiased standardized mean difference in pooled paired samples experiments across a variety of sample sizes and effect sizes. The g, $\hat{\lambda}$, and the limits at the t quantiles were calculated using $\nu = (n-1)$ (left side) or $\nu = 2(n-1)$ (right side). Severe deficits of coverage according to n, ρ and δ emerged when using $\nu = 2(n-1)$. The r used in calculations for this figure was the uncorrected Pearson correlation coefficient.



was calculated according to the pooled formula in Equations 1b and 2. Figure 2 plots the empirical variance of 10,000 d_{P} values at sample sizes of 10, 30, 50, and 100 in a paired samples design with $\delta_P = 1.0$ and with ρ set at .0 to .9 in increments of .1. These empirical variance values are plotted as circles. The different sample sizes were taken from separate simulation runs and are independent.

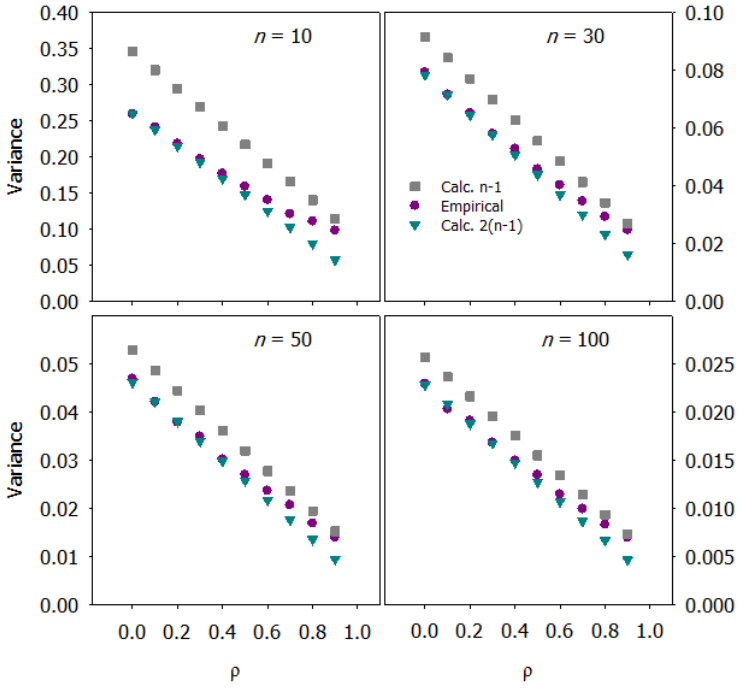
Plotted as guides in Figure 2 are the calculated population variance values (Equation 5) for $\delta = 1.0$ at each ρ when using either $\nu = 2(n-1)$ (triangles Goulet-Pelletier & Cousineau, 2018) or $\nu = (n-1)$ (squares Morris, 2000; Borenstein, Hedges, Higgins, & Rothstein, 2009). The calculations with $\nu = 2(n-1)$ are a reasonable fit for the empirical variance below $\rho \approx 0.4$, but the empirical curve begins to flatten out at higher values of ρ until the $\nu = (n-1)$ calculation is a better fit at $\rho = .9$. At each sample size, a quadratic equation was found for the empirical variance as a function of ρ that accounted for 99.9% of the variance.

Fit of $\hat{\lambda}$ to Noncentral t Distribution

Presuming that the correct degrees of freedom could be calculated for any statistic $d_{\mathrm{P}}\sqrt{n/(2(1-\rho))}$, the distribution of the statistic would be a noncentral t because the factor $\sqrt{n/(2(1-\rho))}$ is a constant. However, if one assumes the degrees of freedom are either $\nu = 2(n-1)$ or $\nu = (n-1)$ for all values of ρ, the fit of the empirical data with that noncentral t will be incorrect when ρ is anything other than 0 or 1.0. Furthermore, we rarely know ρ in an experiment and must estimate it using r. The statistic $d_{\mathrm{P}}\sqrt{n/(2(1-r))}$ will almost never be exactly a noncentral t with $\hat{\lambda} = d_{\mathrm{P}}\sqrt{n/(2(1-\rho))}$ because the factor $\sqrt{n/(2(1-r))}$ is a random variable that changes with the sample value of r.

These issues are illustrated in Figure 3 through Figure 5, which compares calculated noncentral t cumulative probability distributions (green triangles) with empirical cumulative relative frequency distributions from a simulation of $d_{\mathrm{P}}\sqrt{A}$ values (red circles). The tested effect

Figure 2 ■ Simulations of the empirical variance of d_P for the pooled paired samples experiment (circles) at various levels of ρ and n with $\delta = 1.0$. The d_P was calculated according to Equation 1b and 10,000 iterations were used. Plotted as guides are the calculated population variance values (Equation 5) using either $\nu = 2(n-1)$ (triangles) or $\nu = (n-1)$ (squares). Neither set of theoretical calculations explains the empirical variance of d_P at all levels of ρ.

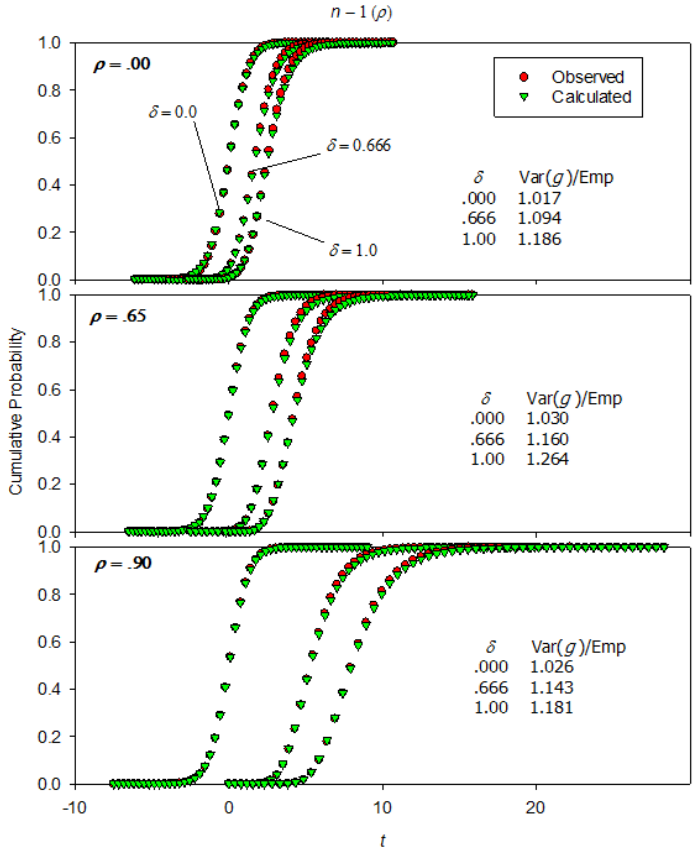


sizes were 0.0, 0.666, and 1.0, ρ-values were .0, .65, and .90, and the relative frequencies were determined from 500,000 simulated experiments for excellent stability. Sample size was $n = 12$. The values $\delta = 0.666$, $\rho = .65$, and $n = 12$ were selected to match the example given in Goulet-Pelletier and Cousineau (2018, Figure 8). A cumulative relative frequency distribution of the statistic was constructed with approximately 50 bins. The three panels are the different levels of ρ. Figures 3 through 5 represent different degrees of freedom and methods used for the experiment. Figure 3 used $\nu = (n-1)$, and the empirical statistic plotted was $d_P\sqrt{n/(2\,(1-\rho))}$ (titled "$n-1$ (ρ)"). Figure 4 used $\nu = 2(n-1)$ and the empirical statistic plotted was also $d_P\sqrt{n/(2\,(1-\rho))}$ (titled "$2(n-1)$ (ρ)". Figure 5 used $\nu = 2(n-1)$, but the statistic plotted was $d_P\sqrt{n/(2\,(1-r))}$ (i.e., using the observed r instead of a constant ρ in the calculation of factor A, titled "$2(n-1)$ (r)"). In each figure, separate curves are drawn for the three effect sizes increasing from left to right.

For "$n-1$ (ρ)" (Figure 3), the fit with the noncentral t was poor at $\rho = .0$ or .45 and a little better at $\rho = 0.9$. For "$2(n-1)$ (ρ)" (Figure 4), the fit with the noncentral t was clearly worst at $\rho = 0.9$ and best (exact) at $\rho = .0$. Neither Figure 3 nor Figure 4 is possible in practice unless ρ is known, which is a rare circumstance. Instead, we have to substitute the sample r for ρ. When the d_P was multiplied by the random variable $\sqrt{n/(2\,(1-r))}$ ("$2(n-1)$ (r)", Figure 5) instead of the constant $\sqrt{n/(2\,(1-\rho))}$, even the case where the fit should be exact, $\rho = .0$, was erroneous, and the fit was wildly inaccurate at $\rho = .9$.

The mean and standard deviation of 1,500,000 values of $A = n/(2\,(1-r))$ are printed in Figure 5. Note the large standard deviations. The constant calculated values of A in the population are: $\rho = .0$, A = 6; $\rho = .65$, A = 17.14; $\rho = .9$, $A = 60$. The mean value of r from the 500,000 simulations and the mean value of $A = n/(2\,(1-r))$ calculated using $n = 12$ for $\delta = 0.0$, 0.5, and 1.0, respectively, were $r = 0.00024$, $A = 5.998507$; $r = 0.631611$, $A = 16.28713$; $r = 0.891007$, $A = 55.04942$. Because r was an underestimate of ρ (Zimmerman, Zumbo, & Williams, 2003) this A calculated from the average r was an underestimate of the population A. Because the distribution of A is skewed with increasing ρ, the actual mean values printed in Figure 5 were even more discrepant from

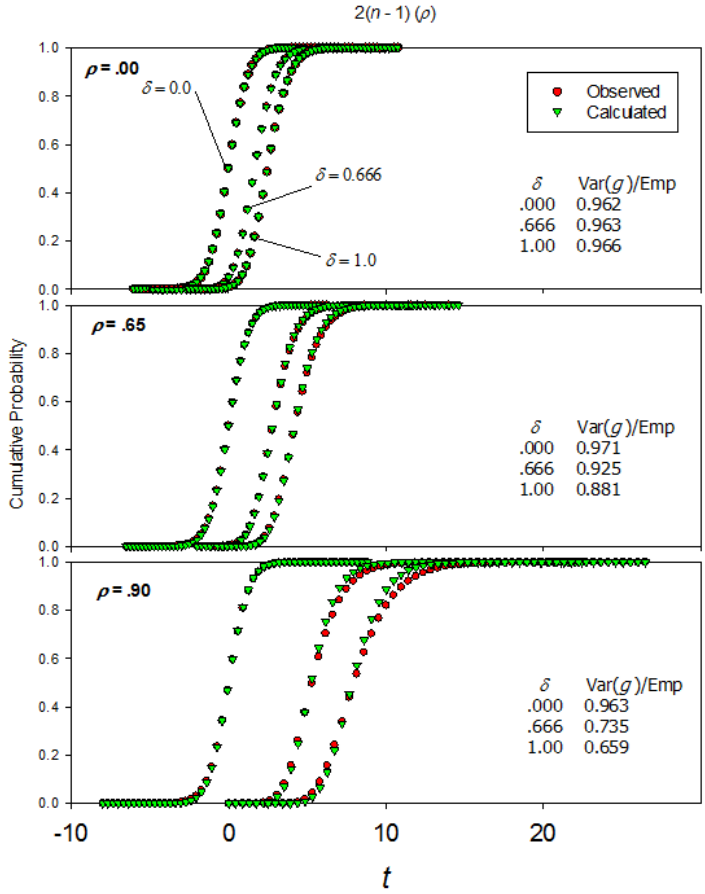
Figure 3 ■ Fit using ρ and $\nu = (n-1)$. Cumulative relative frequency distribution of 500,000 randomly sampled values of $d_{\mathrm{P}}\sqrt{n/(2\,(1-\rho))}$ in pooled paired samples tests (red circles, "Observed") at different population effect sizes δ and correlations ρ, with $n = 12$ and $\nu = (n-1)$. The cumulative noncentral t distribution with population non-centrality parameter λ and $\nu = (n-1)$ is plotted as green triangles. The fit is not unreasonable, but it is certainly not as perfect as would be expected. The ratio of the mean of 500,000 estimates of the population variance, Var(g), to the empirical variance of 500,000 d_{P} values ("Var$(g)/Emp$") was greater than 1.0, meaning that Var(g) is not a good estimate of the actual variance of the d_{P} values.



the population value with increasing ρ. However, there was a perfect predictive relationship ($r^2(7) = .9996$) between the 9 estimates using the mean value of r and the direct mean value of 500,000 values of A given in the Figure 6 ($A_{\text{direct}} = 1.510 \times A_{\text{mean estimate}} - 2.52$). The A is not variable in two-sample or one-sample-difference tests.

The inset tables in Figures 3 and 4 for the constants "$n(\rho)$" and "$2(n-1)(\rho)$" display the ratios of the mean of 500,000 calculated Var(g) values to the empirical variance of 500,000 d_{P} values for different effect sizes and for each level of ρ. The mean Var(g) was calculated for each experiment from Equations 5 and 8b using d_{P} to estimate δ and using the degrees of freedom given in the titles of Figures 3 and 4, respectively. This Var(g) was the unbiased estimate of the population variance of d_{P} for each experiment assuming that the sample $d_{\mathrm{P}}\sqrt{A}$ is distributed exactly as a noncentral t. The empirical variance was the actual calculated variance of simulated d_{P} values and made no assumption about the distribution. When each Var(g) was calculated with $\nu = (n-1)$ the mean Var(g) often overestimated the empirical variance (Figure 3). When Var(g) was calculated with $\nu = 2(n-1)$ it underestimated the empirical variance badly as effect size and correlation increased (Figure 4). The empirical variance was the same in both tests because d_{P} was calculated identically, and the only difference was the degrees of freedom used in calculating Var(g) from Equations 5, 6a, and 8b. In these inset tables in Figures 3 and 4, note that the fit between the circles and tri-

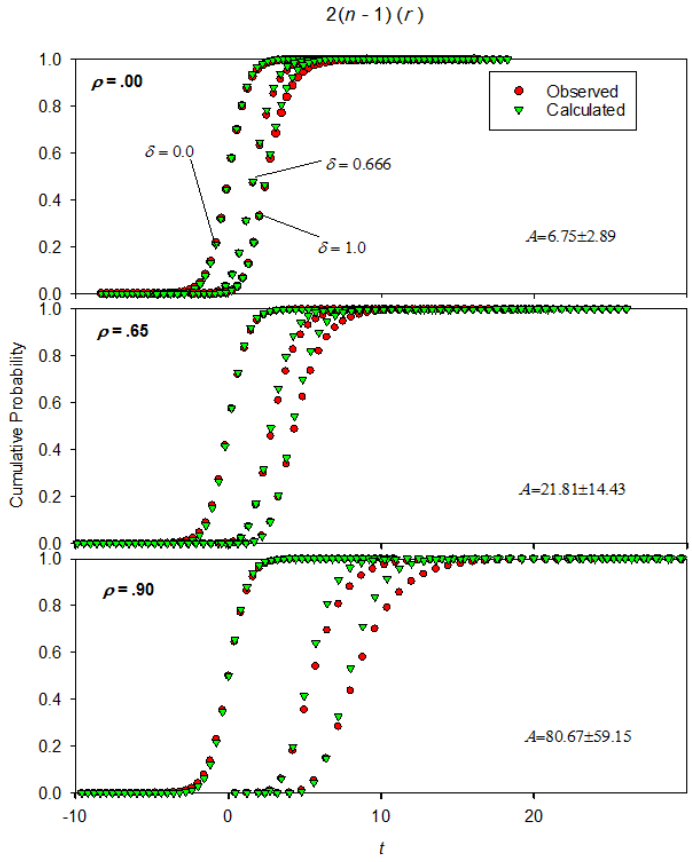
Figure 4 ■ Fit using ρ and $\nu = 2(n-1)$. Cumulative distribution of 500,000 randomly sampled values of $d_P\sqrt{n/(2\,(1-\rho))}$ in pooled paired samples tests (red circles, "Observed") at different population effect sizes δ and correlations ρ, with $n = 12$ and $\nu = 2(n-1)$. The cumulative noncentral t distribution with population non-centrality parameter λ and $\nu = 2(n-1)$ is plotted as green triangles. The fit is excellent with $\rho = .0$, but it is poor with $\rho = .9$ and nonzero effect sizes. The ratio of the mean of 500,000 estimates of the population variance, Var(g), to the empirical variance of 500,000 d_P values was less than 1.0, especially where the fit was worst with $\rho = .9$ and large effect sizes.



angles is best when the ratio of "Var$(g)/Emp$" is closest to 1.0. The instances of poor fit imply that $d_P\sqrt{n/(2\,(1-\rho))}$ was not distributed as a noncentral t with the given degrees of freedom. Note in the top panel of Figure 4, where the fit of the observed distribution and the theoretical distribution were exact, that the Var$(g)/Emp$ ratio is slightly below 1.0, indicating that Var(g) is slightly underestimating the empirical variance. This is typical of small sample sizes such as $n = 12$ (see tables). Figures 2 through 4 suggest that the exact degrees of freedom for a pooled paired samples design are $2(n-1)$ when $\rho = 0$ (Figure 4, top), $(n-1)$ when $\rho = 1.0$ (Figure 3, bottom is $\rho = 0.9$ instead of 1.0), and some function of ρ when the ρ is between 0 and 1.0. Presumably, if the proper degrees of freedom were known, all plots using the constant $\sqrt{n/(2\,(1-\rho))}$ would fit as well as the top panel in Figure 4. This is how the distributions look for a two independent groups design or a difference paired samples design (graphs withheld to conserve space). However, when ρ is unknown and must be estimated from the random variable r, the sampling distribution of the statistic does not fit a noncentral t with $\lambda = \delta\sqrt{n/(2\,(1-\rho))}$ even when the correct degrees of freedom are known (Figure 5, top panel, $\rho = 0$, $\nu = 2(n-1)$).

The fit in Figure 5, bottom panel, $\rho = .90$, $\nu = (n-1)$, was close to a noncentral t distribution, and the remaining discrepancy was probably because the correlation was only .90 instead of 1.0. We can test this by using a correla-

Figure 5 ■ Fit using empirical r rather than the constant ρ and $\nu = 2(n-1)$. Cumulative distribution of 500,000 randomly sampled values of $d_P\sqrt{n/(2\,(1-r))}$ in pooled paired samples tests (red circles, "Observed") at different population effect sizes δ and correlations ρ, with $n = 12$ and $\nu = 2(n-1)$. The cumulative noncentral t distribution with population non-centrality parameter λ and $\nu = 2(n-1)$ is plotted as green triangles. The fit is poor with $\rho = .0$ and even worse with $\rho = .9$ at nonzero effect sizes. The mean and standard deviation of 1,500,000 values of $A = n/(2\,(1-r))$ are given for each value of ρ. Note the progressive inflation of the standard deviation with increases in ρ. The r used in this figure was the uncorrected Pearson correlation coefficient.



tion closer to 1.0 than the .90 given in Figure 3. Because the constant $\sqrt{n/(2\,(1-\rho))}$ is undefined when $\rho = 1.0$, I ran a simulation using $\rho = .999$, $\delta = 1.0$, and $n = 12$ (Figure 6). With the one set of 500,000 simulated d_P values I calculated both the constant $d_P\sqrt{n/(2\,(1-\rho))}$ and the variable $d_P\sqrt{n/(2\,(1-r))}$. Because r is a biased estimator of ρ, as noted above, I calculated $d_P\sqrt{n/(2\,(1-r))}$ using both the usual biased r and an unbiased r (Olkin & Pratt, 1958), as marked in Figure 6. The calculation is:

$$\hat{\rho} = r\left[1 + \frac{(1-r^2)}{2n}\right] \tag{9}$$

The noncentral t distribution is drawn as a curved line for $\nu = (n-1) = 11$. The fit of the noncentral t with the distribution calculated using the constant ρ was better than the fit with $\rho = .90$ in Figure 3 (bottom). This supports the notion that the correct degrees of freedom for $\rho = 1.0$ is $\nu = (n-1)$. The fit using the biased r was poor, and the fit using the unbiased r was even worse. The mean biased r in all simulations was .9989, and the mean unbiased r was .9990. Consider what happens when the value $d_P\sqrt{n/(2\,(1-r))}$ is recalculated with an unbiased r that is slightly larger than the biased r: the smaller denominator makes the value larger. This shifts the entire distribution to higher t values, which is what we see for the unbiased r in Figure 6. Examples for A calculated with each mean r value are: Using biased r, $12/(2\,(1-.9989))$ = 5,454.55 and Using unbiased r, $12/(2\,(1-.9990))$ = 6000.

The tiny difference in r makes a large difference in A. Clearly, the poor fit of the statistic $d_{\mathrm{P}}\sqrt{n/(2\,(1-r))}$ with this noncentral t distribution, even with the correct degrees of freedom, does not result from using the biased r.

Please note that the r used in Figure 5 was the biased Pearson r. As demonstrated here, correcting for that bias in r would shift the distribution of observed values to the right.

Calculation of g and Estimation of Correct ν

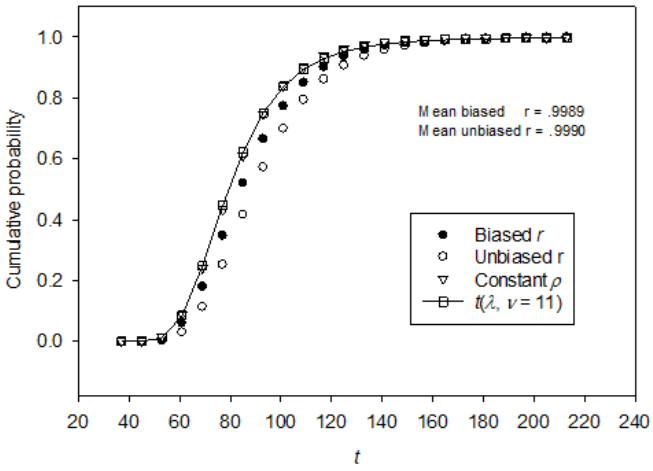
Next, we consider what happens to the calculation of g itself when used with a pooled paired samples design. According to Equation 8a, g is calculated as $g = (d)J(\nu)$, where d is the biased standardized mean difference and J is the bias coefficient with ν degrees of freedom. At the smallest sample size used in Figure 1, $n = 10$, a paired samples test would have $\nu = (n-1) = 9$ for a difference method and, according to Goulet-Pelletier and Cousineau (2018), $\nu = 2(n-1) = 18$ for a pooled method. From Equation 7, $J(9) = 0.91387$ and $J(18) = 0.95765$. Table 1 lists the mean d values of 10,000 pooled paired samples experiments at $n = 10$. This was calculated as d_{P} for pooled paired samples and two independent groups designs (i.e., using the pooled standard deviation as the divisor) and as d_D for difference paired samples designs (i.e., using the standard deviation of difference scores as the divisor). The degrees of freedom are not used in these calculations of mean d. Note that the mean d becomes biased above δ as the effect size increases. To the right are the g values calculated by multiplying the mean d by either $J(9)$ or $J(18)$. If the degrees of freedom are correct, the g should remove the bias in mean d and become closer to δ. The research designs are indicated by "Design" (pairs or 2Gps) and "SD-type" (pool or diff). The top section of the table gives data for pooled paired samples (pairs, pool) at different values of δ and ρ as given in Figure 1. With this design, $g(\nu = 18)$ is closer to δ than $g(\nu = 9)$ when ρ is .0; and $g(\nu = 9)$ is closer to δ than $g(\nu = 18)$ when ρ is .9, although the fit is not perfect. If the degrees of freedom for $\rho = 1.0$ are $(n-1) = 9$, we should expect that the true degrees of freedom for $\rho = .90$ would be slightly larger than 9.

The lower parts of Table 1 give the same information for d values using either a difference paired samples design or a two independent groups design. The g is a good estimate of δ in all cases. The ρ has no effect on g in these two designs.

Table 2 summarizes the problems of estimation when using the pooled paired samples method compared with a two independent samples method or a difference paired samples method. Six simulations of 500,000 iterations were conducted for the "worst case" scenario in Figure 1 of $n = 10$, $\delta = 1.0$ and $\rho = .9$, including four simulations for the pooled paired samples method with either $\nu = 2(n-1)$ or $\nu = (n-1)$ and using either the biased r (Sim) or the Olkin-Pratt-corrected unbiased r (OPcor, Equation 9 Olkin & Pratt, 1958) in calculating $\mathrm{Var}(d)$ or in converting between d- or g-scaled values and the corresponding t-scaled values. In each, the data from the simulation were compared with expected population values (Pop). The expected value of d, $(E(d))$, was calculated as $\delta/J(\nu)$ (Hedges, 1981). Results for the examples with designs using either two samples or difference paired samples were: (1) the empirical variance (Emp var) of d matched the theoretical population value; (2) g matched δ; (3) the mean value of d matched $E(d)$; and (4) the $\mathrm{Var}(g)$ was a slight underestimate because of small sample size (see Figure 4, $\rho = 0$ top). Results for the example with designs using pooled paired samples were: (1) the empirical variance was different from Pop (asterisks); (2) g did not match δ; (3) the mean observed d did not match $E(d)$; and (4) $\mathrm{Var}(g)$ overestimated Pop. The A in this case was calculated from $A = n/(2\,(1-r))$ using the mean value of r (biased or unbiased) rather than averaging 500,000 values of A directly, but we saw in the results for Figure 5 that there is a highly predictive relationship between the values calculated from the means and the values calculated directly. The mean bias of r and A was nearly eliminated by OPcor, but this did not eliminate the variability of r and A. The empirical variance for pooled paired samples, 0.096, is between 0.057 with $\nu = 2(n-1)$ and 0.114 with $\nu = (n-1)$, so the proper ν for $\rho = .9$ is between $2(n-1)$ and $(n-1)$ for the pooled paired samples model.

Putting Tables 1 and 2 together we can roughly estimate the correct degrees of freedom for the pooled paired samples problem with $\delta = 1.0$, $\rho = .9$ and $n = 10$. From Table 1, the mean d in 500,000 simulations (done 4 times) was 1.082, which is between the $E(d)$ values given in Table 2 as 1.044 with $\nu = 18$ and 1.094 with $\nu = 9$. The corrected ν is therefore between 9 and 18. Through some trial and error we observe that, when calculated with $\nu = 10.2$, $E(d) = 1.0/J(10.2) = 1/.0.9243049 = 1.08194$, which is close to our observed mean d of 1.082. This implies necessarily that the observed g value will be 1.0 as we would expect if the ν were correct. Using this new $\nu = 10.2$ we can now calculate $\mathrm{Var}(g) = 0.091$ using Equations 5, 6a, and 8b. This 0.091 is between the $\mathrm{Var}(g)$ values in Table 2 for pooled paired samples with either 9 or 18 degrees of freedom, but it is a slight underestimate of the empirical variance of .096. However, we note that the $\mathrm{Var}(g)$ for the two independent samples and difference paired samples methods were also a slight underestimate of the empirical variance at $n = 10$, so that is not a difference between designs. This method of finding the best ν for $E(d)$ is probably preferred to a method of finding the best ν for $\mathrm{Var}(g)$ because the mean d is a better predictor of $E(d)$ than the

Figure 6 ■ Fit of noncentral t with $\nu = (n-1)$ to simulations with $\delta = 1.0$, $n = 12$ and $\rho = .999$, where the statistic $d_{\mathrm{P}}\sqrt{A}$ is calculated using either the constant ρ, the variable biased r, or the variable unbiased r. The slight bias in the mean r was corrected in the mean unbiased r. The fit of the statistic $d_{\mathrm{P}}\sqrt{n/(2\,(1-\rho))}$ with the noncentral t was almost perfect. The fit of $d_{\mathrm{P}}\sqrt{n/(2\,(1-r))}$ using the biased r was bad and the fit with the unbiased r was worse even though $(n-1)$ is presumably the correct degrees of freedom when ρ is this close to 1.0.



mean Var(g) is a predictor of the actual variance at small sample sizes.

Discussion

The research designs that employ Cohen's d include a single group design, a two independent groups design, a two repeated (paired) measures design, and a comparison to baseline design. The effect size for a paired measures design can be calculated either as a difference between the two means divided by the pooled standard deviation of the two sets of scores, d_{P}, or as a mean of a single set of difference scores divided by the standard deviation of those scores, d_D. If the difference scores are not available, the mean of the difference scores will be identical to the difference between the means of the two sets of scores, and the standard deviation of the difference scores can be calculated exactly from the standard deviations of the two sets of scores and their correlation using this formula,

$$S_D = \sqrt{S_1^2 + S_2^2 - 2\,r_{12}\,S_1 S_2} \tag{10}$$

The problems with degrees of freedom identified in this article apply only to the pooled paired difference design, d_{P}. Calculating d_{P} for a paired samples design is a legitimate way to compare the effect size of a paired measurements design to an effect size from a different experiment with two independent groups. The d_D cannot be compared directly with the d_{P} of the two groups design. If one conducts a paired measures experiment and calculates d_D , one may be able to convert d_D into a d_{P} so that the results can be compared with a two independent groups test. Goulet-Pelletier and Cousineau (2018) give a formula in their Table 1 that is incorrect, and this was corrected in their corrigendum (2019). The correct interconversion formulas are:

$$d_D = \frac{d_{\mathrm{P}}}{\sqrt{2\,(1-r)}} \tag{11a}$$

$$d_{\mathrm{P}} = d_D\sqrt{2\,(1-r)} \tag{11b}$$

Unfortunately, this formula works well only when the sample variances are homogeneous (Lakens, 2013). The difference paired samples design, d_D, requires a normal distribution of difference scores, but, unlike the d_{P} design, it does not require that the two set of scores are equally distributed. Suppose, for example, that a meta-analyst is given a properly conducted, paired samples t test in a target article and wants to calculate d_{P} from the t (Lakens, 2013). The meta-analyst should not assume without other evidence that the variances of the two sets of scores were homogeneous. Trying to find d_{P} from d_D is not always straightforward, and d_{P} should be calculated directly if possible.

The problem is more serious if one instead calculates g_D and then tries to convert the unbiased g_D into a g_P. The formula for calculating a g is given in Equation 8a, and it involves multiplying a d times a correction factor $J(\nu)$ that depends on the degrees of freedom, ν. The correct degrees

Table 1 ■ Effect of research design and degrees of freedom on the calculation of the unbiased standardized mean difference g. The design was either paired samples (pairs) or two independent groups (2 Gps) and the standard deviation for standardizing d was either a pooled standard deviation (pool) or the standard deviation of the difference scores (diff). The mean d value of 10,000 simulated experiments is given for each combination of δ and ρ with $n = 10$. A g value was calculated as $g(\nu = 9)$ using $\nu = (n - 1)$ or as $g(\nu = 18)$ using $\nu = 2(n - 1)$. The $g(\nu = 9)$ was a better estimate of δ with the pooled paired samples design when ρ was large (0.9), and the $g(\nu = 18)$ was a good estimate of δ when ρ was .0. The other research designs are unaffected by ρ, and g was always a good estimate with the proper degrees of freedom.

Design	SD type	δ	ρ	Mean d	$g(\nu = 9)$	$g(\nu = 18)$
pairs	pool	0	0	-0.002	-0.002	-0.002
pairs	pool	0	0.45	-0.002	-0.002	-0.002
pairs	pool	0	0.9	0.001	0.001	0.001
pairs	pool	0.5	0	0.517	0.473	0.495
pairs	pool	0.5	0.45	0.528	0.483	0.506
pairs	pool	0.5	0.9	0.543	0.496	0.520
pairs	pool	1	0	1.045	0.955	1.001
pairs	pool	1	0.45	1.056	0.965	1.011
pairs	pool	1	0.9	1.082	0.990	1.037
pairs	diff	0	0	0.001	0.001	–
pairs	diff	0.5	0	0.540	0.494	–
pairs	diff	1	0	1.089	0.995	–
2 Gps	pool	0	0	-0.002	–	-0.001
2 Gps	pool	0.5	0	0.530	–	0.508
2 Gps	pool	1	0	1.043	–	0.999

of freedom for g_D is $(n - 1)$, but we have seen that the correct ν for g_P varies between $(n-1)$ and $2(n-1)$ depending on the value of ρ. Converting a g_D based on $\nu = (n - 1)$ into a g_P based on some other degrees of freedom and using a sample r instead of the proper ρ is problematic and will not be attempted here. Caution is advised, and the actual g_P should be calculated directly if possible, using the correct degrees of freedom once they are known.

Of course, ρ is rarely known in an experiment like it is in a simulation. Presumably, one might estimate ρ using the sample r and then invent a formula that estimates ν for calculating g and for identifying a noncentral t distribution from the estimate of ν. The problem with this is the fact that the d_P values must be multiplied by a constant in order for them to be scaled and distributed exactly as a known noncentral t variate (Hedges, 1981). In this article I call that constant $\sqrt{A}$, such that $\lambda = \delta_P\sqrt{A}$ (Equation 4), and A is calculated differently for different designs (Equation 6a, 6b, 6c). For the difference paired samples test and the two independent groups test A is a constant in every experiment ($A = n$ and $A = \tilde{n}/2$, Equations 6b and 6c, respectively) and the distribution of $\hat{\lambda}$ is always a unique noncentral t distribution with known degrees of freedom. Replacing the ρ in $A = (n/(2\,(1 - \rho)))$ with r as $A = (n/(2\,(1 - r)))$ replaces the required constant with a random variable, and the distribution of $\hat{\lambda}$ calculated with r is rarely the correct noncentral t distribution corresponding to ρ.

For example, suppose we have three experiments with different experimental designs that all have the result $d = 0.5$ and $n = 10$ per group. Experiment 1: For a two independent groups design, $\nu = 2(n - 1) = 18$ and $\hat{\lambda} = d_P\sqrt{\tilde{n}/2} = 1.118$, and the distribution $t_{\lambda,\nu}$ for computing a confidence interval with $d = 0.5$ is always $t_{1.118,18}$. Experiment 2: For a difference paired samples design, $\nu = (n-1) = 9$ and $\hat{\lambda} = d_D\sqrt{n} = 1.581$, so the distribution $t_{\lambda,\nu}$ for computing a confidence interval with $d = 0.5$ is always $t_{1.581,9}$. Experiment 3: For a pooled paired samples design with unknown ρ, the ν and λ are both unknown even though $d = 0.5$ as in Experiments 1 and 2, and the distribution $t_{\lambda,\nu}$ for computing a confidence interval is unknown. Unlike the other two designs, we must estimate both ν and $\hat{\lambda}$ based on an r that will vary from experiment to experiment even if d remains the same. This adds error to the pooled paired samples design that is not present in the other designs. Presumably, as the sample size grows larger, r will vary less and less from ρ, and the error will be less, but this is a qualification we do not have to place on Experiment 1 or Experiment 2. Whether the error is acceptable for practical purposes at a given sample size must

Table 2 ■ Summary of problems with a pooled paired samples design. Mean values of 500,000 simulations for the variance of d with worst-case parameters $n = 10$, $\delta = 1.0$ and $\rho = .9$. Simulations include designs for two independent samples, difference paired sample, and pooled paired samples. The latter was repeated for $(n-1)$ or $2(n-1)$ degrees of freedom, and r was either corrected for bias (Olkin-Pratt correction, "OPcor") or not ("Sim") in simulations. "Pop", theoretical population value; "$E(d)$" $= \delta/J(\nu)$, expected value of d; "calc A", the value of A calculated using the mean r (either biased or unbiased) in $n/(2\,(1-r))$.

	TWO SAMPLES		PAIRED SAMPLES (difference)		PAIRED SAMPLES (pooled)					
	$2(n-1)$		$n-1$		$2(n-1)$			$n-1$		
	Sim	Pop	Sim	Pop	Sim	OPcor	Pop	Sim	OPCor	Pop
n	10	10	10	10	10	10	10	10	10	10
ν	18	18	9	9	18	18	18	9	9	9
r	0	0	0.889	0.9	0.889	0.901	0.9	0.889	0.901	0.9
S	0.986	1	0.973	1	0.975	0.975	1	0.976	0.976	1
d	1.045	1	1.094	1	1.082	1.082	1	1.082	1.082	1
$E(d)$	1.044	–	1.094	–	1.044	1.044	–	1.094	1.094	–
$J(\nu)$	0.958	–	0.914	–	0.958	0.958	–	0.914	0.914	–
g	1.001	1	1.000	1	1.036	1.037	1	0.989	0.989	1
Var(d)	0.272	0.260	0.254	0.217	0.069	0.066	0.057	0.141	0.137	0.114
Var(g)	0.249	0.260	0.212	0.217	0.063	0.061	0.057	0.117	0.115	0.114
Emp var	0.261	0.260	0.218	0.217	*0.096	*0.097	0.057	*0.096	*0.096	0.114
calc A	5	5	10	10	45.045	50.505	50.000	45.045	50.505	50.000

Note. *The empirical variance ("Emp Var"), 0.096, is between .057 and .114, so the proper ν for $\rho = .9$ is between $2(n-1)$ and $(n-1)$ for the pooled paired samples model.

await further studies of the coverage of confidence intervals based on some novel formula to estimate the degrees of freedom for g_P from r.

In their Table 3, Goulet-Pelletier and Cousineau (2018) list the true formula and six different approximation formulas for calculating the variance of Cohen's d. Their true formula is equivalent to the current Equation 5, which I used in conjunction with Equation 8b to calculate Var(g) in Figures 3 and 4 and Table 2. This variance is not required for calculating noncentral t confidence intervals for d or g. However, the variance of d and its square root have been used in calculating central normal approximations to the noncentral t distribution (Hedges, 1981, 1982; Morris, 2000). These normal approximations to the noncentral t confidence intervals using approximations to the true formula for the variance of d leave a lot of room for error even when the sample size is large. Goulet-Pelletier and Cousineau (2018) replace the normal approximations of Hedges and Morris with central t approximations for their tests, which will always produce wider confidence intervals, especially at small sample sizes. These authors and I agree that noncentral t confidence intervals are greatly preferred to these approximations. The approximations were necessary before fast computers and appropriate software made computations from the gamma function and the noncentral t distributions easy.

Goulet-Pelletier and Cousineau (2018) give a listing in the R statistical programming language at the end of their paper that calculates noncentral t confidence intervals using the method of Hedges and Olkin (alluded to by Hedges, 1981; first fully described for d in Hedges and Olkin, 1985, p. 91) for two of the discussed experimental designs, the two independent groups design and the pooled paired samples design. Because of the problems with degrees of freedom identified in this paper, their pooled paired samples design should not be used. They did not provide a script to calculate the difference paired samples design. I am appending code in Listing 1 that demonstrates how to calculate noncentral t confidence intervals using the method of Hedges and Olkin with two independent groups or with difference scores in a paired samples design.

An executable program and its source code are available to demonstrate the functions used in these simulations (see paragraph 1 of the Methods and Results section). It allows calculations similar to the R listing but also does simulations like Figure 1 and demonstrates the limits of the functions used. Users can simulate a two groups design, a difference paired samples design, or the problematic pooled paired samples design with the erroneous degrees of freedom recommended by Goulet-Pelletier and Cousineau (2018) using the software. The problem with degrees of freedom for the pooled paired samples design af-

fects the performance of any method, approximate or exact, for calculating its confidence intervals. For the exact noncentral t methods this includes both the Hedges and Olkin (1985) method and the Steiger and Fouladi (1997) method.

Goulet-Pelletier and Cousineau (2018) recommend abandoning the interval estimation approach of Steiger and Fouladi (1997) in their Appendix C. In an erratum Goulet-Pelletier and Cousineau (2020), they retracted that recommendation. This method generates different confidence intervals from the method of Hedges and Olkin (1985). The true tests of a confidence interval are its performance in covering the parameter with the nominal confidence and its compactness (precision). Goulet-Pelletier and Cousineau (2018) do not demonstrate which method provides better coverage although the Steiger and Fouladi method gives slightly more compact confidence intervals. Software is freely available to calculate confidence intervals with the Steiger and Fouladi approach (Cumming & Finch, 2001; Kelley, 2007), although neither handles paired samples based on the difference scores (d_D). The Cumming and Finch software computes intervals based on d_{P} with $\nu = (n-1)$ instead of the correct degrees of freedom. In my opinion, we should await a comparative study of the coverages and compactness of the two approaches with various experimental designs, effect sizes and sample sizes before recommending the abandonment of either approach.

Goulet-Pelletier and Cousineau (2018) cite three groups who promote the use of the pooled paired samples design over the difference paired samples design. Dunlap, Cortina, Vaslow, and Burke (1996) compare the size of effect generated by the two methods in the context of meta-analysis. They convert between d_D to d_{P} using a formula that is equivalent to Equation 11a. None of their calculations involve the degrees of freedom, such as calculating g_P or a noncentral t confidence interval, so their conclusions are not affected by the present results. Lakens (2013) demonstrates how to calculate d_{P} and also how to calculate $g_P = d_{\mathrm{P}}(J(n-1))$ using $\nu = (n-1)$, and refers to Cumming's (2012) lament that the g_P is not completely unbiased. Goulet-Pelletier and Cousineau (2018) thought (p. 253) that this bias problem was solved by using $\nu = 2(n-1)$, but the present simulations show that the correct degrees of freedom, and therefore the correct g_P, both depend on a knowledge of ρ.

References

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

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York, NY: Routledge.

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. doi:10.1177/0013164401614002

Dunlap, W., Cortina, J., Vaslow, J., & Burke, M. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods*, *1*, 170–177. doi:10.1037/1082-989X.1.2.170

Fitts, D. A. (2018). Variable criteria sequential stopping rule: Validity and power with repeated measures anova, multiple correlation, manova and relation to chi-square distribution. *Behavior Research Methods*, *50*, 1988–2003. doi:10.3758/s13428-017-0968-5

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. doi:10.20982/tqmp.14.4.p242

Goulet-Pelletier, J.-C., & Cousineau, D. (2019). Corrigendum to "a review of effect sizes and their confidence intervals, part i: The cohen's d family". *The Quantitative Methods for Psychology*, *15*, 54–55. doi:10.20982/tqmp.15.1.p054

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*.

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*. San Diego, CA: Academic Press.

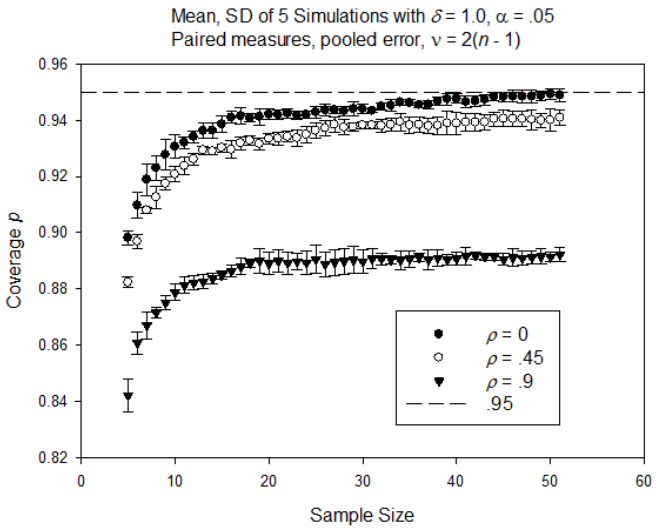
Kelley, K. (2007). Confidence intervals for standardized effect sizes: Theory, application, and implementation. *Journal of Statistical Software*, *20*, 1–24.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t tests and anovas. *Frontiers in Psychology*, *4*, 863–875. doi:10.3389/fpsyg.2013.00863

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.

Figure 7 ■ Reliability of coverage curves for sequential samples from $n = 5$ through 50 for 95% noncentral t confidence intervals for pooled paired samples that were each calculated as recommended by Goulet-Pelletier and Cousineau (2018).



Olkin, I., & Pratt, J. W. (1958). Unbiased estimation of certain correlation coefficients. *Annals of Mathematical Statistics*, *29*, 201–211.

Steiger, J. H., & Fouladi, R. T. (1997). Noncentrality interval estimation and the evaluation of statistical models. *What if there were no significance tests*, 197–229.

Zimmerman, D. W., Zumbo, B. D., & Williams, R. H. (2003). Bias in estimation and hypothesis testing of correlation. *Psicologica*, *24*, 133–158.

Appendix A

Because experiments in Figure 1 were sequential, each sample size in a single experiment is dependent on those before it. The simulation was the mean of 10,000 such experiments, and comparisons within a curve must be made with caution. Each curve was independent of the others, and comparisons between curves is the point of the simulation. Figure 7 shows the mean $\pm$ 1 standard deviation for coverage probabilities of 5 independent simulations of 10,000 sequential experiments each for $n = 5$ through 50. Independent simulations were conducted with $\delta = 1.0$ and $\rho = 0$, .45, and .90, and $\nu = 2(n\check{}1)$. Compare to Figure 1, bottom right panel, for $n = 10$ through 50.

Open practices

The *Open Material* badge was earned because supplementary material(s) are available on the journal's web site.

Citation

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 paired samples designs. *The Quantitative Methods for Psychology*, *16*(4), 281–294. doi:10.20982/tqmp.16.4.p281

Received: 01/01/2016 ~ Accepted: 20/02/2016

Listing 1 follows.

Listing 1 ■ Calculation of noncentral t con1dence intervals using the method of Hedges and Olkin with two independent groups or with difference scores in a paired samples design

```
#Noncentral t confidence interval generic calculator Hedges&Olkin (1985) method.
#Edit values between the rows of dotted lines
#.....................................................
Samples <- 1
   # must be 1 for paired samples or 2 for independent samples, no error checking!
   # For two samples, d is calculated as the mean difference divided by the pooled standard deviation.
   # For one sample (paired, matched) the standard deviation of the difference scores is used.
   #
alpha    <- .05
   # 1 minus confidence coefficient; e.g., for 95% interval alpha = 1 - .95.
   #
Meandiff <- -0.6
   #sample unstandardized mean difference used to form d; positive, negative, or 0
   #
SD    <- 2
   # SD for independent samples is the pooled standard deviation of the two groups
   # SD for paired, correlated, matched is the standard deviation of the difference scores
   #
n  <- 9
   #equal n, sample size per group or number of differences
#.....................................................
#Do not change anything below here

Harmmean <- 2*(n*n)/(n+n)
if (Samples == 1) df <- n - 1
if (Samples == 2) df <- 2*(n - 1)
#Calculating A
if (Samples == 1) {
    A <- n                          # if difference scores
} else {
    A <- Harmmean/2                 # if 2 samples
}
sqrtA <- sqrt(A)
J <- exp(lgamma(df/2)-(log(sqrt(df/2))+(lgamma((df-1)/2))))
d <- Meandiff/SD
varD <- (1/A)*(df/(df-2))*(1+A*d*d)-(d*d)/(J*J)
g <- d*J
varG <- varD*J*J
ncpD <- d * sqrtA                   #non-centrality parameters
ncpG <- g * sqrtA
lldt <- qt(alpha/2, df, ncpD)       #lower limit biased, t-scaling
uldt <- qt(1-alpha/2, df, ncpD)     #upper limit biased, t-scaling
lld  <- lldt/sqrtA                  #lower limit biased, d-scaling
uld  <- uldt/sqrtA                  #upper limit biased, d-scaling
llgt <- qt(alpha/2, df, ncpG)       #lower limit unbiased, t-scaling
ulgt <- qt(1-alpha/2, df, ncpG)     #upper limit unbiased, t-scaling
llg  <- llgt/sqrtA                  #lower limit unbiased, g-scaling
ulg  <- ulgt/sqrtA                  #upper limit unbiased, g-scaling

#Show calculated values
cat("Samples = ",Samples, "  Mean difference =",Meandiff," SD =",SD, " n =",n,"\n")
if (Samples == 2) {cat("SD is pooled standard deviation of the two equal groups.\n")
} else {cat("SD is standard deviation of the differences between the paired scores.\n") }
cat("degrees of freedom = ",df, "  A =",A,"  J =",J, "\n")
cat("Effect size\nBiased d", d, " Var(d)", varD,"\nUnbiased g", g, "Var(g)", varG, "\n")
cat (100*(1-alpha),"% noncentral t confidence interval\n")
cat("Standardized Biased scaling: d =",d," Interval [", lld,",", uld, "]","\n")
cat("Standardized Unbiased scaling: g =",g," Interval [", llg,",", ulg, "]\n")
```