2022 ■ Vol. 18 ■ no. 2

Point and interval estimates for a standardized mean difference in paired-samples designs using a pooled standard deviation.

Douglas A. Fitts [a]

[a]University of Washington

Abstract ■ A standardized mean difference using a pooled standard deviation with paired samples (d_p; paired-pooled design) can be compared directly to a d_p from an independent samples design, but the unbiased point estimate g_p and confidence interval (CI) for d_p cannot unless the population correlation ρ between the scores is known in the paired-pooled design, which it rarely is. The ρ is required to calculate the degrees of freedom ν for the design, and ν is necessary to calculate the g_p and CI. If a variable sample correlation is substituted for ρ the ν is only approximate and the sampling distribution for d_p is unknown. This article uses simulations to compare the characteristics of the unknown distribution to the noncentral t distribution as an approximation and provides empirically-derived regression equations to compensate for the bias in the approximated CI computed using the noncentral t distribution. The result is an approximate but much more accurate coverage of the CI than previously available. Tables are supplied to assist sample size planning and computer programs are provided for computations. These results are experimental and tentative until the actual distribution can be discovered. The regularity of the deviation in coverage that allows the compensation to work encourages that search.

Keywords ■ confidence interval, Cohen's d, Hedges' g, simulation, noncentral t distribution.

Acting Editor ■ Roland Pfister (Universität Würzburg)

Reviewers ■ Denis Cousineau (Université d'Ottawa)

■ and one more.

✉ dfitts@uw.edu

doi 10.20982/tqmp.18.2.p207

Introduction

A common experiment in psychology and science in general involves a comparison of two sample means. These means can be from two independent samples of participants or subjects (two-sample test), or they can be from two dependent or correlated samples (paired-samples test) such as repeated measures on one group of subjects or measures from two different groups of subjects that have been matched into pairs on a positively correlated matching variable. This article presents the most accurate method to date for computing confidence intervals (CIs) for a standardized mean difference in paired samples tests when the pooled standard deviation is the denominator for standardization. The study also explores the effect of using the method on the point estimate of effect size. The method is based on regression equations determined from the results of the simulations, where a "practical confidence coefficient" replaces the nominal confidence coefficient in the calculation of the CI. For many uses, an investigator need only look up the practical coefficient in a table and use that instead of the nominal coefficient when calculating a CI. Some computing using supplied software may be required for coefficients that do not appear in the table.

I begin by describing the way such means have been compared historically, and the problems with the proposed solution. I then present simulations that demonstrate the bias in coverage of approximated CIs and a simple method to compensate for the bias in many cases.

An unstandardized (raw) mean difference D can be standardized by dividing by a standard deviation (Becker, 1988; Cohen, 1988; Hedges, 1981). For two independent samples with equal population variances, the best estimate of the population standard deviation σ is a pooled estimate from the two standard deviations. This article assumes equal sample sizes in the two groups or conditions, so the pooled standard deviation, the standardized mean difference, and the degrees of freedom ν can be calculated as:

$$S_p = \sqrt{\frac{S_1^2 + S_2^2}{2}} \tag{1}$$

$$d_p = D/S_p \tag{2}$$

$$\nu = 2(n-1) \tag{3}$$

For a paired samples design, the denominator is often the standard deviation of the differences calculated directly from the difference scores, or, if the standard deviations and Pearson r are available:

$$S_D = \sqrt{S_1^2 + S_2^2 - 2rS_1S_2} \tag{4}$$

$$d_D = D/S_D \tag{5}$$

$$\nu = (n-1) \tag{6}$$

The standardized mean difference d is a positively biased estimator of the population δ. Hedges (1981) applied a correction to produce an unbiased estimate g from the biased estimate d based on the degrees of freedom ν and the gamma function $\Gamma()$ (the nomenclature here conforms to current usage, not the usage in Hedges' original article).

$$\text{Hedges' unbiased } g = d \times J(\nu);$$

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

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) ) )
```

The log function helps prevent overflow with large n. Thus, the unbiased versions of the standardized mean differences for the two designs are $g_p = d_p \times J(\nu)$ and $g_D = d_D \times J(\nu)$.

The unstandardized mean difference D is calculated the same in two-sample and paired-sample designs and can be compared directly between studies that use different designs. Confidence intervals can easily be computed for an unstandardized difference (Fitts, 2022a; Kelley, Maxwell, & Rausch, 2003) and in some contexts it is the preferred difference for reporting (Bond, Wiitala, & Richard, 2003; Wilkinson & the Task Force on Statistical Inference, 1999). Standardized units are preferred when comparing studies that use units of measurement that are more arbitrary or not compatible with other measuring instruments. For example, effect sizes from two studies of anxiety that use different testing scales can be compared in standardized units when comparing unstandardized units makes no sense.

The standardized estimates of effect size cannot be compared directly between these two experimental designs because the standardizer S_P is usually quite different from S_D. In order to facilitate comparisons of effect sizes from these different designs for purposes of gaining more accurate estimates and for combining results using meta-analysis, methodologists recommended that the same standardizer S_P should be used in both designs as calculated in Equation 1 (e. g. Borenstein, Hedges, Higgins, & Rothstein, 2009, p. 29). In this article, I refer to the paired samples design used with a pooled error term as a "paired-pooled design" to distinguish it from the paired design based on difference scores. This worked well for comparing d_p from the two designs, but the g_p computed from the paired samples design was biased when using $\nu = (n-1)$ as recommended by Borenstein et al. (2009) or when using $\nu = 2(n-1)$ as recommended by Goulet-Pelletier and Cousineau (2018). Fitts (2020) demonstrated that the degrees of freedom in a paired-pooled test, as required for the calculation of $J(\nu)$, varied according to the population correlation ρ from $\nu = n-1$ when ρ = 1.0 to $\nu = 2(n-1)$ when $\rho = 0$. Cousineau (Cousineau, 2020; Cousineau & Goulet-Pelletier, 2021) identified the correct degrees of freedom for the paired-pooled design when ρ is known,

$$\text{Paired-pooled - population: } \nu = 2(n-1)/(1+\rho^2), \tag{8}$$

and presented the approximate distribution of the standardized mean difference d_p when ρ is known. Cousineau and Goulet-Pelletier (2021) then published a study of eight different protocols for constructing approximate CIs for d_p, some of which used the noncentral t distribution as an approximation to the unknown distribution of d_p when calculated using the sample r instead of ρ.

The noncentral t distribution requires the calculation of a noncentrality parameter λ, and in general terms the definitions for the sample estimate and the population parameter are:

$$\text{Sample: } \hat{\lambda} = d\sqrt{A}; \tag{9a}$$

$$\text{Population: } \lambda = \delta\sqrt{A} \tag{9b}$$

where A is a scaling factor based on the experimental design. For the two-sample test or a paired samples difference test the value of A is:

$$\text{Two-samples pooled error term:} A = n/2 \tag{10}$$

$$\text{Paired-samples differences:} A = n. \tag{11}$$

and the associated degrees of freedom are given in Equations 3 and 6. For the paired-pooled design, however, the value of A, like the degrees of freedom, depends on the value of ρ,

$$\text{Paired- pooled - population:}A = \frac{n}{2\left(1-\rho\right)} \tag{12}$$

Because ρ is unknown in an experiment, it must be approximated using some variant of the sample correlation coefficient. For the degrees of freedom, the variant is the approximate unbiased correlation coefficient r_{OP} (Olkin & Pratt, 1958),

$$r_{OP} = r\left[1+\frac{\left(1-r^2\right)}{2\left(n-3\right)}\right] \tag{13}$$

and for the noncentrality parameter the variant is the rectified correlation coefficient r_W (Cousineau, 2020)

$$r_W = r\frac{S_1 S_2}{S_p^2} \tag{14}$$

Thus, the approximate degrees of freedom and A for a sample using the paired-pooled design are,

$$\text{Paired- pooled - sample:}\nu = 2(n-1)/(1+r_{OP}^2) \tag{15}$$

$$\text{Paired- pooled - sample:}A = \frac{n}{2\left(1-r_W\right)}. \tag{16}$$

Because r is a random variable, the degrees of freedom ν and noncentrality parameter $\hat{\lambda}$ from a sample are only approximations and will rarely equal the population values of ν and λ as calculated using ρ. It is known that using the estimated ν and a variable r often produce CIs with other than the expected nominal coverage with the paired-pooled design when using CI protocols employing a noncentral t distribution (Cousineau & Goulet-Pelletier, 2021). What is not known is whether a correction exists to make the coverages of the CIs more accurate.

In studying CIs for standardized mean differences, one frequently refers to the expected coverage of the CI and to its complement the exclusion rate (or Type I error rate). “Coverage” means the proportion of randomly formed CIs using a particular CI protocol that includes the population standardized mean difference δ. This expected coverage can often be calculated directly if the distribution of scores is known to be normal (Fitts, 2021), and I abbreviate this expected or desired coverage as η^0. The exclusion rate for this coverage is therefore $\alpha^0 = 1-\eta^0$. The η^0 and α^0 are inputs to the CI protocol that control coverage and exclusion rate in a simulation. In addition to this expected coverage, there is also an empirical or observed coverage that is an output of a simulation and calculated as the proportion of times that the CIs from random samples in the simulation included the population δ. This observed or empirical coverage is abbreviated $\eta*$. Ideally, the value of $\eta*$ in a simulation will be close to the expected value η^0 if all assumptions have been met. The variability of $\eta*$ around η^0 grows smaller as n grows larger. However, if we apply a CI protocol with a nominal η^0 to a distribution that is skewed or otherwise does not meet the assumptions, we may find that the observed $\eta*$ from the simulation consistently fails to equal η^0. This is important for the upcoming simulations because we will be using the noncentral t distribution based on a population noncentrality parameter λ to estimate the noisier and unknown distribution based on $\hat{\lambda} = d_{\mathrm{P}}\sqrt{n/(2\left(1-r_W\right)}$ (combining Equations 9a and 16).

The exploration of the CI for the paired-pooled design requires a brief digression to decide what method to use and how to calculate it. The statistic d is not normally distributed, but a transformation of d ($\hat{\lambda}$ in equation 9a) creates a variable that is distributed as a noncentral t with noncentrality parameter λ and degrees of freedom as given in Equations 3, 9a, and 10 for the two-sample design or in Equations 6, 9a, and 11 for the paired-sample difference design. For the d_p calculated using ρ with paired-pooled design the distribution is also approximately a noncentral t (Equations 8, 9b and 12). Therefore, an approximate but highly accurate CI can be constructed using the noncentral t distribution for d_p from two independent samples, for d_D from a paired-sample test using S_D, or from a paired-pooled test using d_p if ρ and ν are both known. The subject of this article is the application of the same method to a CI for d_p using the paired-pooled method when using a variable r instead of ρ as would happen in most experiments.

In addition to the point estimate g_p and the CI, it will be useful to calculate the variance of g assuming the sample standardized mean difference is drawn from a noncentral t distribution. The equation for δ is adapted from Hedges (1981) where the factor A is defined in Equations 10, 11, and 12, for each experimental design.

$$Var\left(d\right) = \left(\frac{1}{A}\right)\frac{\nu}{\nu-2}\left(1+\left(A\right)\delta^2\right) - \frac{\delta^2}{J(\nu)^2}. \tag{17}$$

If a biased d is substituted for δ, the unbiased $Var(g)$ (Hedges, 1981) can be calculated as:

$$Var(g) = Var(d)J(\nu)^2 \tag{18}$$

The following experiments will explore whether the $Var(g)$ is an unbiased estimate of the population variance of d when used with a paired-pooled design using $A = \frac{n}{2(1-r_W)}$ (Equation 16) as a substitute for $A = \frac{n}{2(1-\rho)}$ (Equation 12) and using $\nu = 2(n-1)/(1+r_{OP}^2)$ (Equation 15) as a substitute for $\nu = 2(n-1)/(1+\rho^2)$ (Equation 8).

Table 1 ■ Summary of all definitions of sample statistics (a) and population parameters or ideal values (b) for mean difference test between Sample 1 and Sample 2 arms as given throughout the text. The d can be computed using either the standard deviation of the difference scores S_D or the pooled standard deviation of the two sets of scores as in the two-sample test, S_P.

Symbol	Sample statistic (a)	Population parameter or ideal (b)
Raw mean difference	$D = M_1 - M_1$	$\Delta = \mu_1 - \mu_2$
Raw CI full width	W	Ω
Standardized d (two sample scores)	$d_p = D/S_p$	$\delta_p = \Delta/\sigma_p$
A (two samples)	$A = n/2$	$A = n/2$
df (two samples)	$\nu = 2(n-1)$	$\nu = 2(n-1)$
CI width	$w = W/S_p$	$\omega = \Omega/\sigma_p$
Standardized d (difference scores)	$d_D = D/S_D$	$\delta_D = \Delta/\sigma_D$
A (difference scores)	$A = n$	$A = n$
df (difference scores)	$\nu = n - 1$	$\nu = n - 1$
CI width	$w = W/S_D$	$\omega = \Omega/\sigma_D$
Standardized d (paired-pooled scores)	$d_p = D/S_p$	$\delta_p = \Delta/\sigma_p$
A (paired-pooled scores)	$A = \frac{n}{2(1-r_W)}$	$A = \frac{n}{2(1-\rho)}$
df (paired-pooled scores)	$\nu = 2(n-1)/(1+r_{OP}^2)$	$\nu = 2(n-1)/(1+\rho^2)$
CI width	$w = W/S_p$	$\omega = \Omega/\sigma_p$
d to t conversion	$\hat{\lambda} = d\sqrt{A}$	$\lambda = \delta\sqrt{A}$
t to d conversion	$d = \hat{\lambda}\sqrt{\frac{1}{A}}$	$\delta = \lambda\sqrt{\frac{1}{A}}$

Note. λ, the population noncentrality parameter of the relevant noncentral t distribution with degrees of freedom ν in column b. $\hat{\lambda}$, an estimate of that noncentrality parameter with degrees of freedom ν in column a; r_W, the rectified Pearson r, Equation 14; r_{OP}, the Pearson r corrected for bias, Equation 13. Assumes $n_1 = n_2 = n$ per group for two independent samples or $n =$ number of pairs for paired samples. A is a scaling factor unique to each design.

Noncentral t Confidence Intervals

Methods for constructing "exact" CIs for either d or g in a two independent-samples test have been proposed (Goulet-Pelletier & Cousineau, 2018; Hedges, 1981; Hedges & Olkin, 1985; Steiger & Fouladi, 1997). Unfortunately, the CI generated by Hedges and Olkin method is not the same as that generated by the Steiger and Fouladi method. Fitts (2021) provided a method for calculating the exact coverages of these four different CI methods: Hedges and Olkin with d or with g, and Steiger and Fouladi with d or with g. The most accurate was the Steiger and Fouladi method used with d, which always produced coverage that matched the desired coverage η^0. Thus, when it comes to standardized mean differences, we are in the awkward position of preferring the unbiased g as a point estimate of effect size and the CI for the biased d as the best interval estimate. Kelley and Rausch (2006) give other examples where best point and interval estimates are based on a combination of biased and unbiased estimators.

Kelley and Rausch (2006) used the Steiger and Fouladi's (1997) method with d to determine sample sizes for two-sample tests across a wide range of standardized CI full widths, w, that should yield an observed coverage equal to the desired coverage on average. This aids sample size planning in experiments so that a sample size can be selected to get the desired width and coverage on average. A complicating feature of standardized CIs is the fact that the sample size required to produce the desired coverage varies with effect size: larger standardized effect sizes require larger sample sizes to achieve the same desired confidence level for a fixed width of CI than smaller or null standardized effect sizes. Tables 1 to 3 in Kelley and Rausch (2006) are two dimensional tables with the standardized effect size $\delta = \Delta/\sigma$ as the columns and the standardized widths $\omega = \Omega/\sigma$ as rows (where Ω is the desired full width of the CI in raw units, see Table 1).

The exact distribution of $\hat{\lambda} = d_{\text{P}}\sqrt{n/(2\,(1-r_W)}$ is unknown (Cousineau & Goulet-Pelletier, 2021). However, one can use the noncentral t distribution to approximate the unknown distribution with the paired-pooled method to see the conditions under which g_p provides an unbiased point estimate of δ and the Steiger and Fouladi's (1997) method with d_p provides a close enough approximation to the nominal coverage to be useful. One paired-pooled protocol tested by Cousineau and Goulet-Pelletier (2021), which they called the "Pivotal of t'", involves the same calculation of the CI by the Steiger and Fouladi (1997) protocol except that the noncentrality parameter and degrees of freedom were as given in Equations 9a, 15, and 16 for the paired-pooled scores instead of for the two-sample scores. This article uses that protocol.

Although the CI method for d_D assumes only a normal distribution of difference scores, the method for d_p must also require homogeneous variances for the use of S_p to be valid. Therefore, the populations in the simulations had equal variances, and a correlation between sampled scores was induced as a variable in different simulations using a common method as previously described (Fitts, 2018, 2020).

Calculating the CI Limits

Observed CI. The calculation of the CI for the standardized mean difference d uses the algorithm of Steiger and Fouladi (1997) and involves a computer search for convergence rather than a closed-form equation. The d for the experiment is calculated first, and it is converted to a t value, $\hat{\lambda}$, according to Equations 9a and 16. The lower and upper limits, LL_t and UL_t, of a two-sided CI around $\hat{\lambda}$ are the noncentrality parameters of two unique noncentral t distributions with ν degrees of freedom. The LL_t has $\hat{\lambda}$ as the t quantile corresponding to probability $1-\alpha^0/2$, and the UL_t has $\hat{\lambda}$ as the t quantile corresponding to the probability $\alpha^0/2$. The algorithm for this uses a computer search for the desired noncentrality parameters of the two noncentral t distributions with ν degrees of freedom that have $\hat{\lambda}$ as the quantile suitably close to probabilities $1-\alpha^0/2$ and $\alpha^0/2$ (current approximations are within 8 decimal digits).

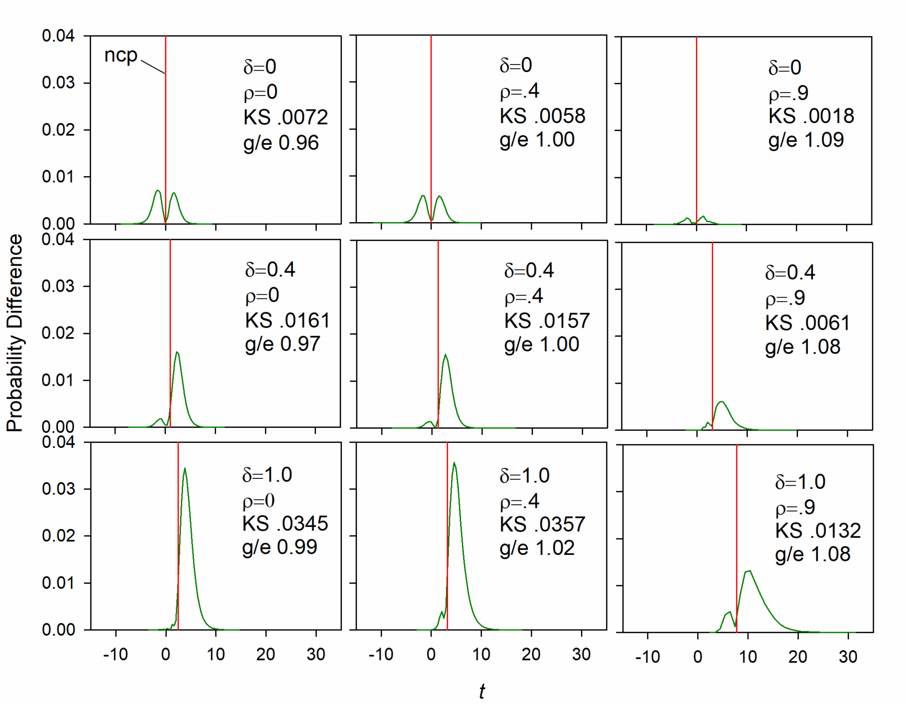
For any given ν, the noncentral t distribution at $\lambda = 0.0$ is identical to the central t distribution for that ν and is symmetrical. All other noncentral t distributions have a degree of skewness in proportion to the absolute value of λ and a direction according to the sign of λ (for graphical illustrations see Cumming & Finch, 2001; Fitts, 2021; Kelley & Rausch, 2006). Because of the skewness, the quantiles at symmetrical probabilities such as $1-\alpha^0/2$ and $\alpha^0/2$ are not equal distances from the noncentrality parameter λ. To put the limits in the same units as d we convert LL_t and UL_t to standard score equivalents LL_d and UL_d according to an algebraic manipulation of Equations 9a and 16:

$$LL_d = LL_t\sqrt{\frac{1}{\frac{n}{2(1-r_W)}}};\qquad UL_d = UL_t\sqrt{\frac{1}{\frac{n}{2(1-r_W)}}} \tag{19}$$

Steiger and Fouladi-Compliant Fixed-width CI. The sample size necessary for a CI having nominal coverage for an interval of exact width ω with a paired-pooled design can be calculated when ρ is known (Fitts, 2021), but it must be estimated from simulations when using r. In the search for the correct sample size, one sets a trial n and a fixed width ω but does not know η for that n. The challenge is to create an interval like a Steiger and Fouladi interval with equal probabilities in the two tails of a noncentral t distribution, and from these tail probabilities one can calculate $\eta = 1 - \alpha$ for that n. The interval of exact width ω must be identical to a Steiger and Fouladi interval created for that n and its now known η. That is, we need to calculate the CI limits of an interval of a fixed width and sample size without knowledge of its confidence coefficient such that the limits are identical to a CI calculated from the normal Steiger and Fouladi algorithm based on the now known confidence coefficient and sample size without knowledge of the width. I call this a Steiger and Fouladi-compliant fixed-width CI. The algorithm calculates a trial value for the lower limit, determines the probability in the tail below that limit, then calculates an upper limit exactly ω standardized units above the lower limit and examines the probability in the upper tail. It adjusts these limits up and down until it finds the unique pair of limits that have equal probabilities (within 8 decimal digits) in the two tails. The sample size at which the expected coverage of this fixed-width interval equals η^0 is the sample size tabled in Kelley and Rausch (2006) for two-sample tests. This is also the sample size where the average standardized width w of the observed noncentral t CI reaches width ω in a two-sample test. In a paired-pooled test where r is used in place of ρ, however, the η must be estimated from a simulation using the real, unknown distribution instead of the noncentral t distribution and the average width w will not equal ω. The fixed-width CI in no way uses η^0 or α^0 in calculations.

Table 1 summarizes all definitions and equations for both sample statistics and population or ideal values for two-samples, paired-samples differences, and paired-samples pooled designs for a mean difference test between Sample 1 and Sample 2.

Figure 1 ■ Difference between the cumulative relative frequency of 500,000 simulated values of $\hat{\lambda} = d_p\sqrt{\frac{n}{2(1-r_W)}}$ and the calculated cumulative probability of $\lambda = \delta_p\sqrt{\frac{n}{2(1-\rho)}}$ with $\nu = 2(n-1)/(1+\rho^2)$ with $n = 12$. The peak value is the Kolmogorov-Smirnov (KS) statistic. g/e is the ratio of the mean of 500,000 computed estimates of the variance of the standardized unbiased g statistic to the empirical variance of 500,000 d values. ncp: non-centrality parameter.

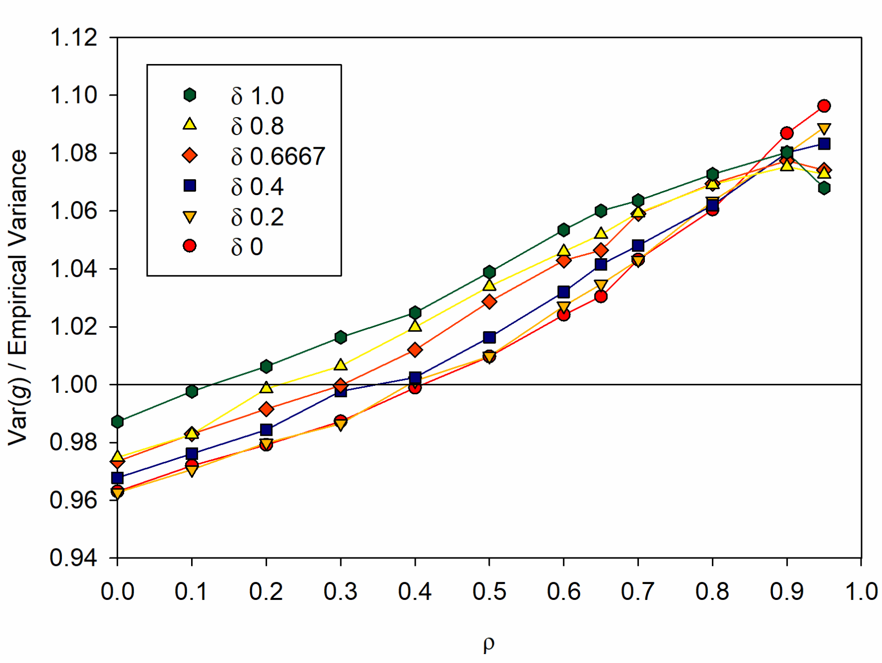


Simulations Using the Paired-Pooled Design

Empirical Sampling Distribution of $d_p\sqrt{\frac{n}{2(1-r_W)}}$

The CI protocol for d_p is to use the basic Steiger and Fouladi (1997) noncentral t procedure with $\hat{\lambda}$ for standardized mean differences with S_p as the basis for standardization of d (Equations 9a and 16). The first step in this exploratory work is to conduct simulations to demonstrate how much error is introduced into the calculations by using r_W instead of ρ. The method was to sample 500,000 statistics of the form $\hat{\lambda} = d_p\sqrt{\frac{n}{2(1-r_W)}}$ with a relatively small sample size of $n = 12$ pairs. In independent simulations the effect size was set at $\delta_p = 0, 0.2, 0.4, 0.6667$, and 1.0 and the population correlation $\rho = .0, .2, .4, .65$, and $.9$, for 25 total simulations. For each, the 500,000 values of $\hat{\lambda}$ were analyzed as cumulative relative frequencies in approximately 50 bins. The approximate noncentral t distribution cumulative probability (Cousineau, 2020) was then computed for each bin using ρ instead of r to calculate λ and ν: $\lambda = \delta_p\sqrt{\frac{n}{2(1-\rho)}}$, and $\nu = 2(n-1)/(1+\rho^2)$. The Kolmogorov-Smirnov (KS) test is a rough but simple way to compare cumulative probability distributions by computing the maximum absolute difference between the two distributions among all 50 bins with larger KS values indicating greater discrepancies between the distributions. Figure 1 plots these differences (empirical minus theoretical) across the 50 bins in a subset of nine of these distributions. The KS statistic is the largest absolute difference in each graph, and these include both extremes observed in 25 combinations. In addition, the empirical variance of 500,000 d_p values was calculated for each combination of simulation parameters. For the same combinations, the mean of 500,000 estimates of $Var(g)$ was calculated from Equations 17 and 18, and the ratio of the $Var(g)$ to the empirical variance is listed in Figure 1 as "g/e". Numbers less than 1.0 indicate that the best estimate of the variance of d given the noncentral t distribution was less than the actual variance of the unknown distribution of d with r_W being a variable. The extremes of the 25 analyses are included in Figure 1 as 0.96 to 1.09. The fit between distributions is

Figure 2 ■ Ratio of the mean of 500,000 $Var(g)$ values to the variance of 500,000 simulated d values in paired-pooled experiments with $n = 12$ (augments data from Figure 1). The variance of g is a best estimator of the variance of δ if the $\hat{\lambda}$ statistics are distributed as a noncentral t with $\lambda = \delta_p\sqrt{\frac{n}{2(1-\rho)}}$ and $\nu = 2(n-1)/(1+\rho^2)$. The empirical variance of g is calculated from each observed d value using Equations 17 and 18. The positive slope is one indicator of the severity of the departure of the $\hat{\lambda}$ statistics from a noncentral t distribution. Loss of experiments to increasing computation faults account for the decline with very high ρ and large δ. See Table 1 for notation.



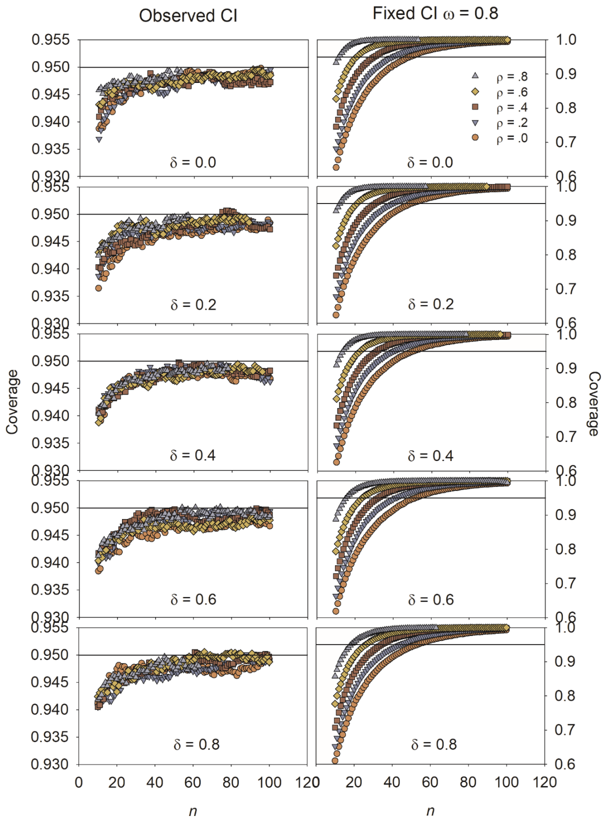
worst where the differences in the figure are greatest, and a perfect fit would be a flat line at 0 with no differences. The noncentrality parameter is the vertical red line, and it is clear that the differences between the distributions occur mainly in one or both tails.

Figure 2 is an extension of the previous stimulations to display the g/e ratio at ρ values of 0, .2, .3, .4, .5, .6, .6667, .7, .8, .9, and .95 with 500,000 iterations each. If the empirical distribution of $\hat{\lambda} = d_p\sqrt{\frac{n}{2(1-r_W)}}$ were a noncentral t, the lines would track a ratio of 1.0 across the graph. Instead, the empirical variance of d is wider than $Var(g)$ at $\rho = 0$ and increases monotonically relative to the $Var(g)$ as ρ increases. The effect size δ is also a source of variance in the ratio, with larger δ values exceeding smaller ones at most values of ρ but then converging near the limit of $\rho = 1.0$. The apparent crossover on the right resulted from experiments lost to computation faults, which increased exponentially with very high correlation and very large effect size. The regularity of the changes as a function of δ and ρ bode well for the future invention of corrections for these biases.

Sample Size and Coverage of Fixed-Width Intervals in a Paired-Pooled Design

Calculated Sample Sizes Using ρ. The sample size to generate a CI of a known fixed width and confidence coefficient is available for a two-sample test using function ss.aipe.smd in the R package `MBESS` (Kelley, 2007). A similar sample size tool for a paired-pooled design is not available, so it was necessary to generate one (see Software). Tables S1, S2 and S3 in the online Supplement list sample sizes for experiments using a paired-pooled experimental design with coverage coefficients .90, .95, and .99, respectively. The method was to calculate the Steiger and Fouladi (1997) approximate CI for d_p using $\lambda = \delta_p\sqrt{\frac{n}{2(1-\rho)}}$ and $\nu = 2(n-1)/(1+\rho^2)$ with successive sample sizes to find the largest sample size that yields a result $w \leq \omega$. This calculation gives the average stopping sample size when ρ is not a random variable and the noncentral t distribution is a very close approximation to the sampling distribution

Figure 3 ■ Simulated coverages of 95% observed CIs and fixed-width CIs for $\omega = 0.8$ at various values of δ, ρ and n with a paired-pooled design using the sample r instead of ρ [$\delta = 0.0$ to 0.8 by 0.2; $\rho = .0, .2, .4, .6, .8$; iterations = 50,000]. See Table 1 for notation.



(Cousineau, 2020).

In practice, most investigators will not know ρ and must use the noisy, unknown distribution of $d_p\sqrt{\frac{n}{2(1-r_W)}}$. For that reason, I simulated all experiments with each combination of ω, δ, and ρ using Equations 9a, 15, and 16 to calculate the noncentrality parameter and degrees of freedom using a CI with a fixed width of ω in independent simulations of 50,000 experiments at each of a large number of sample sizes until I identified the sample size where the coverage of the fixed-width CI was η^0. The observed CI was also calculated at each n as explained below. These sample sizes determined by simulations using r required a slightly larger n to achieve a fixed width of ω, and this was affected by η^0 as follows: η^0 = .90 required +0.8±1.1; η^0 = .95 required +1.3±1.4; and η^0 = .99 required +2.7±3.9 extra subjects. Thus, it is wise to add 1, 2, or 3 extra subjects to the numbers given in Tables S1, S2 and S3 respectively.

The values in Tables S1, S2, and S3 are useful for planning sample size with a paired-pooled design. For each protocol I also calculated the expected coverage of the observed CI using the algorithm of Fitts (2021). These expected coverages using the population ρ were always the nominal η^0. The expected coverages of the observed CI cannot be calculated for the experiments using the sample r because the sampling distribution is not a noncentral t. They are best estimated from the simulations below.

Simulated Sample Sizes Using r Instead of ρ. I simulated the experiments implied by Tables S1, S2 and S3 using a paired-pooled design at different values of n, δ, ω, and ρ using η^0 = .90, .95, or .99 with $\hat{\lambda}$ and ν given by Equations 9a, 15 and 16. The designation is [η^0 = .90, .95, .99; $\omega = 0.25, 0.4, 0.6, 0.8, 1.0, 1.2$; $\delta = 0.0$ to 1.0 by 0.2; $\rho = .0, .2, .4, .6, .8$; iterations = 50,000]. For each combination of ω, δ, and ρ, the output variables were determined independently at each sample size beginning at a small number and ending at a large number relative to the value of ω. For example, the n for $\omega = 0.25$ ranged from 50 to 700, whereas the n for $\omega = 1.2$ ranged from

10 to 34. For $\omega = 0.25$, this means that there were 50,000 independent experiments at each tested sample size from 50 to 700. See supplementary material concerning variability between identical simulations of 50,000 iterations. This simulation gives the average coverage and stopping sample size when r is a random variable that contributes noise to the experiment. Using the same procedure, which they call the "Pivotal of t"", Cousineau and Goulet-Pelletier (2021) found coverages of observed CIs when averaged over 10 correlations and 5 effect size scenarios to be unacceptably low at very small sample sizes but within 1% of nominal .95 (i.e., $\geq .94$) at sample sizes of 15 or more.

Representative example simulations for $\omega = 0.8$ are illustrated in Figure 3 for the limited set of values $\delta = 0.0, 0.2, 0.4, 0.6$, and 0.8 and $\rho = .0, .2, .4, .6$, and $.8$ across a wide range of n. The output observations were the 95% CI coverages of the observed CI (Figure 3, left side) and of the standardized fixed-width CI (right side) with $\omega = 0.8$ for a paired-pooled test. When the sampling distribution is known, such as with a two-sample test, an observed CI has a width that is adjusted using the noncentral t distribution so that the coverage for each n is always the nominal η^0. The same is true with the paired-pooled method when ρ is known rather than approximated. However, the coverage data in Figure 3 are from the approximation with r instead of ρ. See "Calculating the CI Limits".

I did not detect differences in coverage with the observed CI that were consistent across values of ρ and δ in Figure 3, but small differences according to δ and ρ cannot be ruled out (see supplementary material). What was consistent for all ρ and δ for the observed CI was coverage below the intended .95 at small sample sizes. This agrees with the notion that the estimation of ρ from r will be much more variable at small sample sizes. Furthermore, the coverages were within 1% of nominal except at the smallest sample sizes as observed by Cousineau and Goulet-Pelletier (2021). Overall, the coverage of the observed CI had a small negative bias even after the coverages seemed to stabilize at larger sample sizes. This bias is attributable to the use of r in place of ρ because the expected coverage using ρ was always the nominal .950. This small bias was also observed by Cousineau and Goulet-Pelletier (2021) at sample sizes near 100. With a two-sample test or with a paired samples test using S_D, the average coverage of this observed CI is always the nominal .950 (Fitts, 2021).

The coverages of the fixed-width CIs at $\omega = 0.8$ (right side of Figure 3) had very strong and clear effects of both ρ and δ as a function of n. Note that the scales on the ordinates for the left (observed) and right (fixed) graphs are very different, which accounts for much of the apparent smoothness of the data on the right. For a given ρ and δ the fixed-width interval had low coverage at small sample size and increased regularly with increasingly larger sample sizes until the coverage eventually reached and then exceeded the desired η^0 = .95. The abscissa point where the coverage curve intersects the line representing .95 is the sample size where the coverage of a CI protocol with this ω, δ, and ρ is exactly η^0 = .95 on average. These are the points that tended to be slightly larger than the calculated n listed in Table S2 for η^0 = .95. For example, the listing in Table S2 for the protocol [$\delta = 0.8$; $\rho = .8$; $\omega = 0.8$] is 17. The estimate from the simulations with r instead of ρ in the lower right panel of Figure 3 where the far-left curve for $\rho = .8$ intersects the line for .950 was 19, i.e., 2 more than the tabled value.

The foregoing demonstrates that a Steiger and Fouladi-compliant fixed-width CI exists that has a desired width and desired coverage η^0 at some n. With a two-sample test where the sampling distribution is a noncentral t, this interval would be identical to the observed CI at that n, but with the paired-pooled test it is not. In fact, the observed CI calculated using η^0 will have neither that width nor exact coverage at any reasonable n.

Compensating for Biased Coverage of the Observed CI in a Paired-Pooled Design

Here I introduce a new input to the CI protocol, a practical coefficient η'. The confidence coefficient is the value of η that is used to generate a value of $\alpha^0 = (1 - \eta^0)$ for use in the calculation of the limits for the standardized CI (See "Calculating the CI Limits"). Now we will replace it with the practical coefficient $\alpha' = (1 - \eta')$. When a CI protocol with a nominal α^0 consistently produces an observed coverage that is either greater than or less than the desired nominal η^0, perhaps the use of some different value for α' will generate a CI with a consistent average coverage of η^0, which is what we want. This trial α' can be varied in simulation trials until a value is identified that consistently produces the desired overall nominal coverage of η^0. Although a practical coefficient, such as .96, is used in the calculation phase instead of the nominal coefficient, .95, the protocol is still nominally a 95% CI because that is the actual coverage of the procedure (Fitts, 2021).

Figure 4 illustrates both the best η' for each n (open circles and regression curve) and the coverage of an observed CI in a reliability simulation when using the best η' instead of η^0 to calculate the CI at each n (blue filled circles with standard deviation). The best η' was found in simulations with sample sizes between 10 and 160 using the above method for η^0 values of .90, .95, and .99. An inverse first or second order polynomial equation that provided a good fit to the data (adjusted $R^2 > .95$) is drawn as a solid curve (SigmaPlot 14.0) through the plotted empirical best η' results averaged over 30 combinations of δ and

Table 2 ■ Regressed values of the practical coefficient α' for calculating a two-tailed CI for a paired-pooled design at 90, 95, or 99% confidence as a function of n. Values were determined from best-fit inverse first or second order polynomial regression to yield an adjusted $R^2 > .95$. Equations at bottom. See Figure 4.

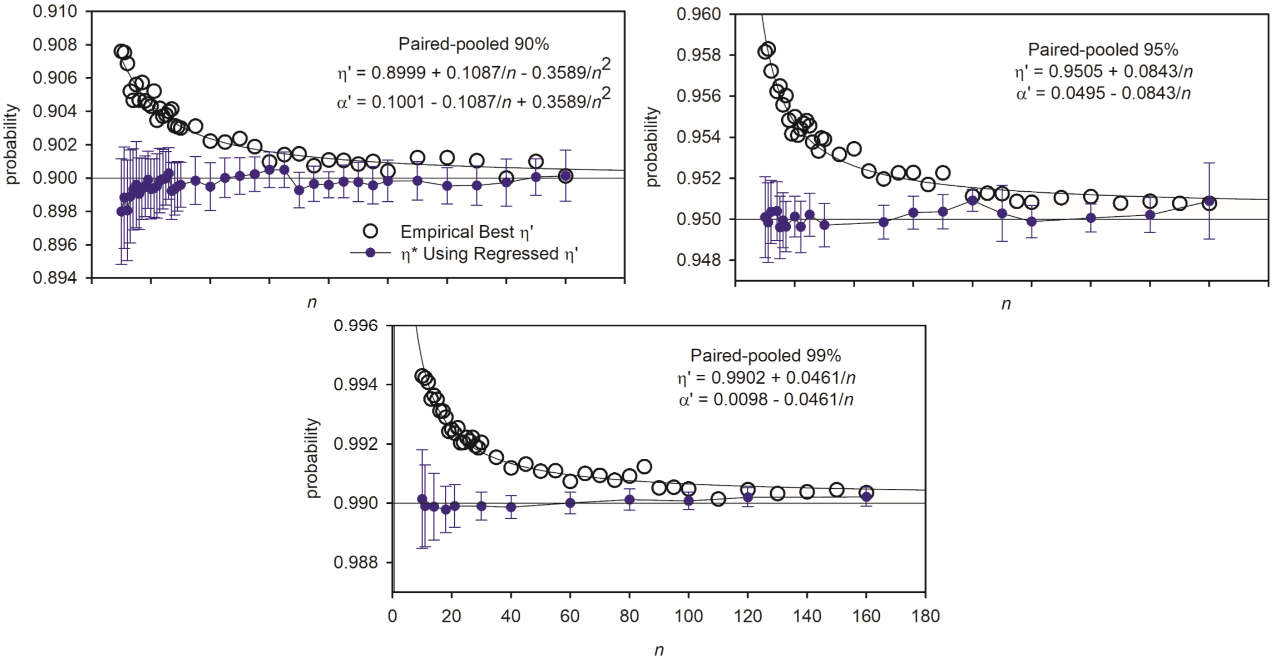
n	90%	95%	99%	n	90%	95%	99%
10	0.0928	0.0411	0.0052	35	0.0973	0.0471	0.0085
11	0.0932	0.0418	0.0056	40	0.0976	0.0474	0.0086
12	0.0935	0.0425	0.0060	45	0.0979	0.0476	0.0088
13	0.0939	0.0430	0.0063	50	0.0981	0.0478	0.0089
14	0.0942	0.0435	0.0065	55	0.0982	0.0480	0.0090
15	0.0944	0.0439	0.0067	60	0.0984	0.0481	0.0090
16	0.0947	0.0442	0.0069	65	0.0985	0.0482	0.0091
17	0.0949	0.0445	0.0071	70	0.0986	0.0483	0.0091
18	0.0952	0.0448	0.0072	75	0.0987	0.0484	0.0092
19	0.0954	0.0451	0.0074	80	0.0988	0.0484	0.0092
20	0.0956	0.0453	0.0075	85	0.0989	0.0485	0.0093
21	0.0957	0.0455	0.0076	90	0.0989	0.0486	0.0093
22	0.0959	0.0457	0.0077	95	0.0990	0.0486	0.0093
23	0.0961	0.0458	0.0078	100	0.0990	0.0487	0.0093
24	0.0962	0.0460	0.0079	110	0.0991	0.0487	0.0094
25	0.0963	0.0461	0.0080	120	0.0992	0.0488	0.0094
26	0.0965	0.0463	0.0080	130	0.0993	0.0489	0.0094
27	0.0966	0.0464	0.0081	140	0.0993	0.0489	0.0095
28	0.0967	0.0465	0.0082	150	0.0994	0.0489	0.0095
29	0.0968	0.0466	0.0082	160	0.0994	0.0490	0.0095
30	0.0969	0.0467	0.0083				
90%	$\alpha' = 0.1001 - 0.1087/n + 0.3589/n^2$						
95%	$\alpha' = 0.0495 - 0.0843/n$						
99%	$\alpha' = 0.0098 - 0.0461/n$						

ρ as in the previous paragraph. The α' for a planned n can be predicted from the equation for any δ or ρ. In blue are means and standard deviations of coverages for reliability simulations with 100,000 iterations each conducted using the best α' rounded to 3 decimal digits [η' regressed for each n; $\delta = 0.0$ to 1.0 by 0.2; $\rho = .0, .2, .4, .6$, and $.8$; iterations = 100,000] for η^0 = .90 or .95 and 4 decimal digits for η^0 = .99. The regressed best α' values for η^0 = .90, .95, and .99 between sample sizes of 10 and 160 are listed in Table 2 along with the regression equations. Closer inspection of these regression equations reveals that each begins with a number very close to the desired α^0 and then subtracts a small amount based on n to calculate α'.

This method demonstrates one way to compensate consistently for the low coverage induced in the observed CI by the use of r instead of ρ in a paired-pooled design with small samples. This is an important result for those who are more interested in obtaining an accurate 100η % CI with a small sample size than in achieving a fixed-width CI, and the single value of η' for each n does not require prior knowledge of σ or ρ. As seen in Figure 4 and explained in the Supplementary Material, the adjustment does not eliminate the possibility of differences in coverage across different values of ρ or δ, but it does adjust the overall mean coverage for all tested values of ρ or δ to the nominal η^0. If ρ and δ are known a priori, the same procedures can be employed to give a more focused adjustment for that combination of parameters using the supplied software and the instructions in the User's Notes (see Example 8).

The adjustment demonstrated in Figure 4 helps to generate paired-pooled CIs with an average coverage near the desired η^0, but it remains to be seen if the procedure with a practical coefficient from Table 2 can be paired with sample sizes from Tables S1 to S3 to generate CIs that have a known approximate width as well. If so, an estimate of the δ and ρ could be used to determine a sample size from Tables S1 to S3 that yields a CI with an average known width and a coverage of η^0 (compare to Tables 1, 2, and 3 in Kelley & Rausch, 2006). My simulation software supplies the average width w of the observed CI for each simulation as an output, so I created a scatterplot for the n from the blue symbols in Figure 4 versus the observed w of the observed CI in that experiment, and these are plotted as the symbols in Figure 5. Examples are shown in Figure 5 for observed

Figure 4 ■ Best η' values for 90%, 95% and 99% observed CIs using a paired samples design with a pooled S_p. Blue circles are the mean and S.D. coverages from simulations across 6 levels of δ and 5 levels of ρ from independent reliability runs of 100,000 iterations when using a 3-digit regressed α' for 90% and 95% and a 4-digit regressed α' for 99%. See Table 2. The η' corrects for the low coverage from using r instead of ρ seen in Figure 3.

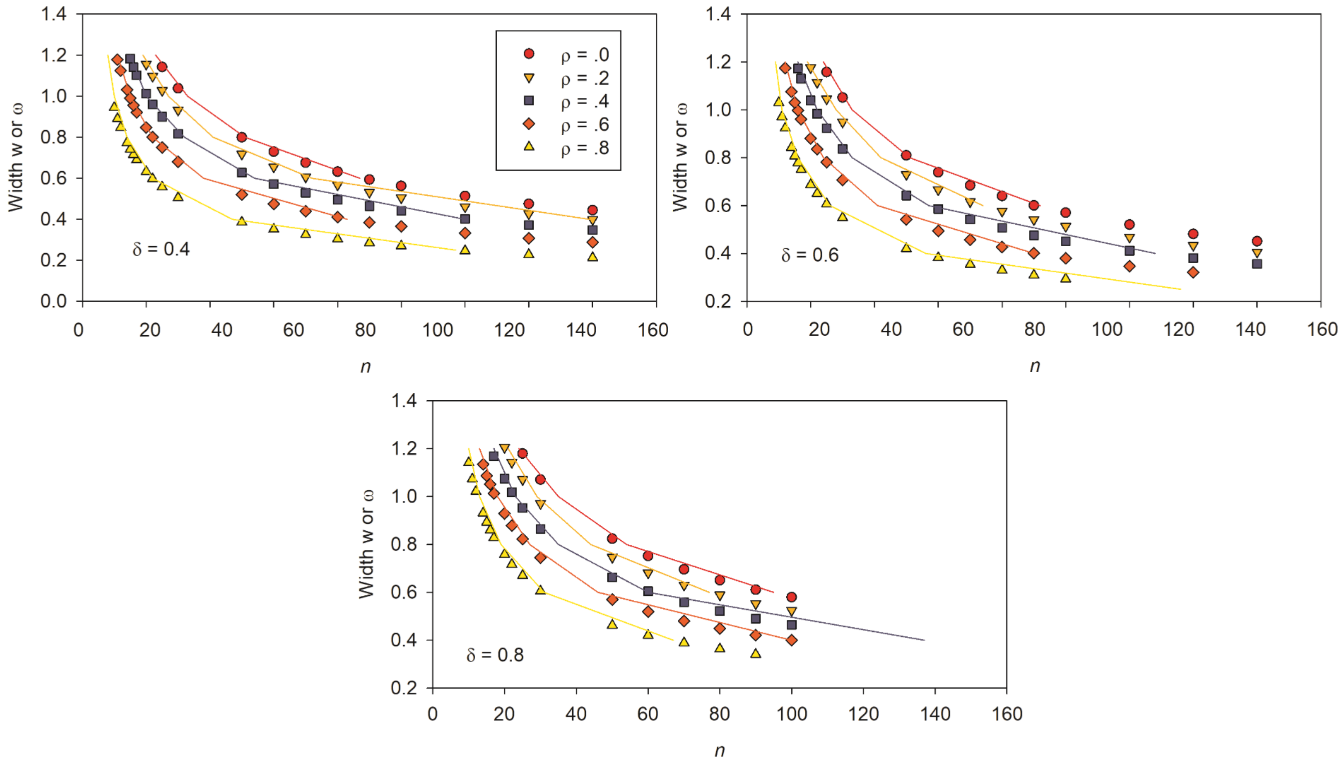


95% CIs with $\delta = 0.4, 0.6,$ and 0.8 and values of ρ from .0 to .8. In these simulations from Figure 4, n was set as an input to the simulation and the width w was a dependent variable in the output. Drawn on the same graph as solid lines with colors coordinated to the levels of ρ are data from Table S2 for 95% CIs where the best sample size is given to produce a 95% CI of a fixed width ω. In these simulations, ω was set as an input to the simulation and the best sample size n was determined from the output. One can see in Figure 5 that selecting a sample size for a given width ω, δ and ρ and then calculating the CI using the regressed α' from Table 2 produces an observed CI with width $w = \omega$ and coverage $\eta*$ = η^0 on average. This generalization at present can apply only to the range of values of ω, δ and ρ tested in these studies. In particular they apply only to positive values of ρ because positive correlations provide the advantage in a repeated measures design and no investigator would intentionally match subjects based on a negatively correlated matching variable. Values of the observed r were of course occasionally negative when randomly sampled from populations with low ρ. Cousineau and Goulet-Pelletier (2021) did include negative ρ values in their simulations.

Complex Example for Parameters Not in the Tables.

Tables 2 and S1, S2, and S3, are sufficient in many but not all situations. Figure 6 provides a graphical summary of the entire method from the sole perspective of the observed CI since that is the one we typically calculate. This large simulation used a basic CI protocol for which none of the parameters are in the current tables [$\omega = 0.7$, $\delta = 0.5$, $\rho = 0.3$, $\eta^0 = .995$, $\alpha^0 = .005$]. The coverage and standardized width were calculated for all test η' values between .995 and .998 by increments of .0001 (i.e., .9950, .9951, .9952, etc.) and for all sample sizes from 95 through 100 for 186 independent simulations of 500,000 experiments each. Figure 6A illustrates the coverage of the observed CI in each of the 186 simulations. The coverage using the nominal η^0 = .995 was too low. Clearly what mattered for coverage was the test η' rather than the sample size within this range, and $\eta' = .9954$ was the best for gaining a coverage of .995 at all sample sizes tested. Figure 6B demonstrates that a standardized width $\omega = 0.7$ can be achieved at any tested η' values between .995 and .996 depending on the sample size, but only a combination of $n = 97$ and $\eta' = .9954$ results in an observed CI with a

Figure 5 ■ Relation between n and the width of observed 95% CI at several levels of δ and ρ determined two ways: (1) Symbols represent the observed width w from the reliability simulations in Figure 4 where n was fixed and w was a dependent variable; (2) Solid lines represent values from Table S2 where ω was fixed and the best n for that width was determined from simulations. Using α' instead of α^0 produces the same average width expected from Table S2.



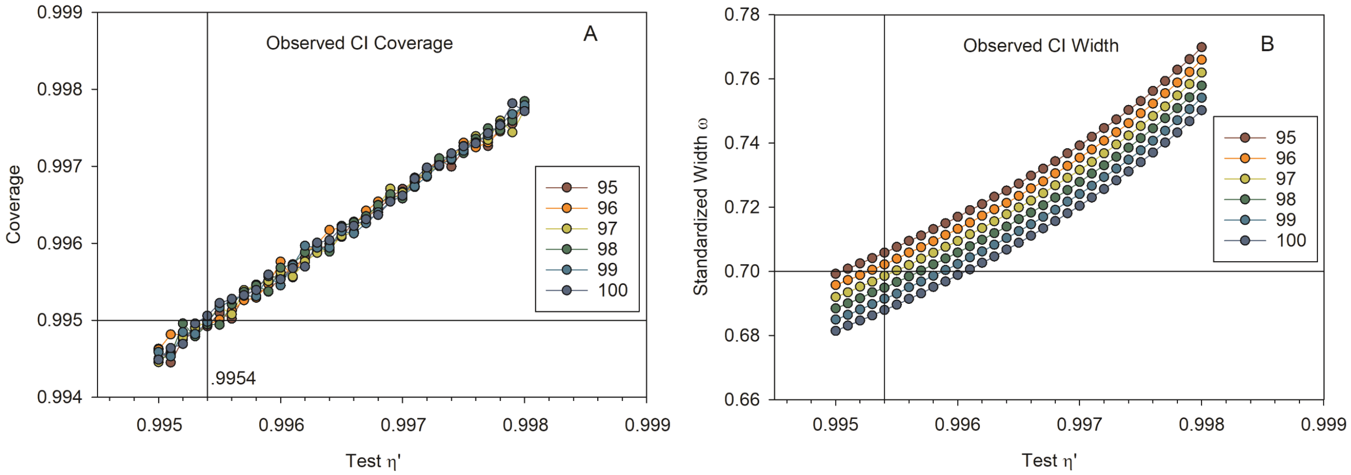
standardized width of $\omega = 0.7$ and also a coverage of .995. This entire simulation is instructive but not necessary in practice with sample sizes this large. Any of these sample sizes can be used to compute the best η' in one smaller simulation, and then that η' value can be tested at different sample sizes in another smaller simulation to determine the n that gives a coverage of the observed CI closest to the nominal η^0. See Example 8 in the User's Notes to repeat this simulation independent of the tables for any combination of ω, δ, ρ, and η^0. With small sample sizes, it is important to test both multiple n values and multiple η' values. The large number of iterations is important.

Summary of the Point Estimates

The biases for d_p and g_p from previous simulations are summarized in Figure 7 as a function of the sample size n, where the mean biases are calculated as (mean $d_p - \delta$) and (mean $g_p - \delta$). The range of δ was 0.0 to 1.0 and the range of ρ was 0.0 to 0.8 with each combination of δ and ρ weighted equally. In these simulated experiments, n was the calculated n for that protocol as listed in Table S2 and it was an independent variable. In all experiments reported here, g_p was an unbiased estimate of δ. The bias values for d_p at small sample sizes increased as a function of δ and ρ. The conclusion is that the use of the sample r to replace ρ in construction of g_p and a CI does not appear to affect the value of g_p itself as much as it does the variance of g_p (Figure 2) and the coverage of the CI (Figure 3).

The mean g_p values were determined from the experiment of Figure 4 where coverage was determined using a regressed value of α' at levels of δ between 0 and .8 and levels of δ between 0 and 1.0 across sample sizes ranging from 10 through 160, and the results are displayed in Figure 8. The data for the two graphs are identical but they are shown at different scales: the graph on the left shows detail and the graph on the right has the same ordinate scaling as Figure 7. Although there is a definite bias of g_p caused by the use of the regressed α', it is minuscule compared to the bias of d_p in Figure 7.

Figure 6 ■ Demonstrations (A) that the best η' to get nominal .995 coverage of the observed CI is .9954 with any n in the range, and (B) with $\eta' = .9954$ the best sample size to get a $\omega \leq 0.700$ with the observed CI is $n = 97$. Protocol [$\omega = 0.7$; $\delta = 0.5$; $\rho = 0.3$; η^0 = .995; iterations = 500,000 per symbol].



Software

Software that can reproduce the simulations for the figures and tables in this article is located at the Open Science Foundation site https://osf.io/q35g6/, including a 64-bit executable PC console application, complete C source code, and user's notes to explain how to conduct the simulations. Software is also supplied to compute a paired-pooled CI using the current methods from raw means, standard deviations, and correlations and to compute the required sample size from the noncentral t distribution with a known confidence coefficient if ρ is known as in Tables S1 to S3 using the method of Fitts (2021). These tables from the Online Supplement are reprinted in the user's notes. The notes include 8 detailed, fully worked examples of applications of the technique from the easiest to the most difficult. Other statistics packages may be useful for computing the approximate CI for paired-pooled designs if the functions use the method of Steiger and Fouladi (1997) with d and allow the user to supply both the noncentrality parameter and the degrees of freedom, $\hat{\lambda} = d_{\mathrm{P}}\sqrt{n/(2\,(1 - r_W)}$ and $\nu = 2(n{-}1)/(1{+}r^2_{OP})$. The R function `conf.limits.nct` in `MBESS` (Kelley, 2007) is one such function (see Cousineau & Goulet-Pelletier, 2021, their Listing 7). That listing uses r instead of r_{OP} in the calculation of ν, but r_{OP} can be calculated with the line:

```
rOP <- r * (1 + (1-r^2)/(2 * (n - 3)))
```

(from Equation 13).

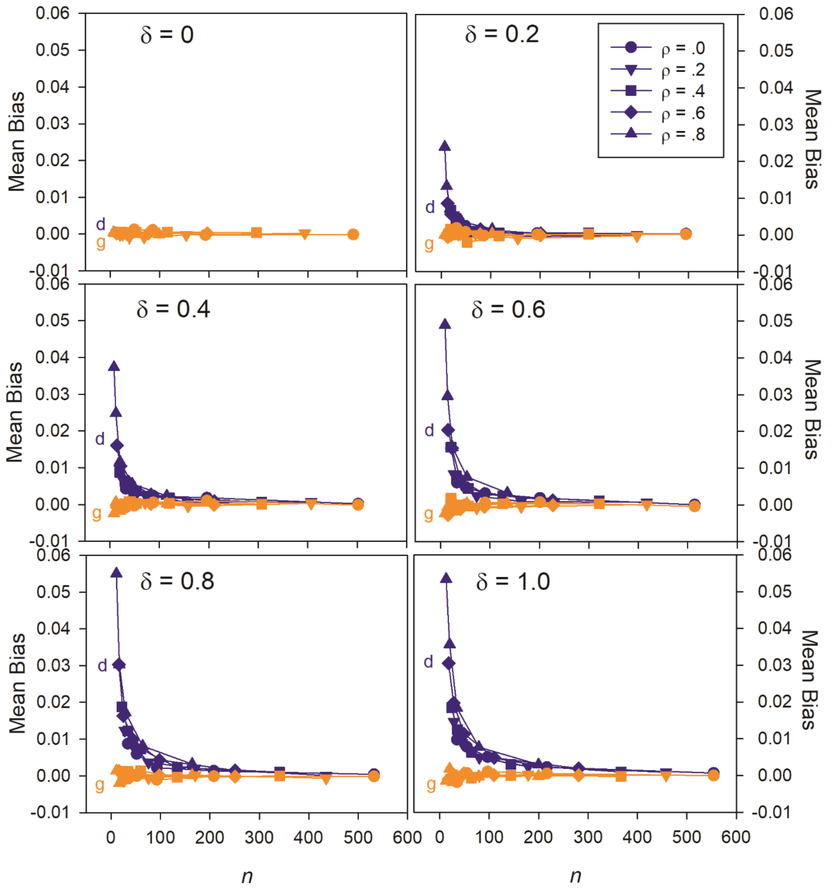
Discussion

The research presented here on the use of a standardized mean difference in a paired samples research design that employs a pooled standard deviation S_p for standardization of d_p (paired-pooled design) is experimental because it is based on simulations of unknown distributions rather than on analytical mathematics and a thorough understanding of the actual distribution of d_p when it is calculated using the r instead of ρ. The method is a useful way to approximate a CI for a standardized mean difference in a paired-pooled design with known coverage until the true distribution is discovered. In many research situations, a CI with nominal coverage can be constructed simply by looking up a practical coefficient in Table 2 and using that α' instead of α to construct the CI. If a certain width is desired, ρ and δ must be estimated, then the experiment can be conducted with augmented sample sizes determined from Tables S1 to S3 along with the practical coefficient from Table 2. More complicated procedures may be necessary if the needed δ, ρ, ω, or η^0 are not included in the tables, and all of these procedures are explained using detailed examples in the user's notes to the accompanying the software.

Summary of the Method

It is always valid to compare a standardized mean difference d_p from a paired-pooled design to a d_p in another study that uses two independent samples. The problem arises when trying to compare unbiased standardized

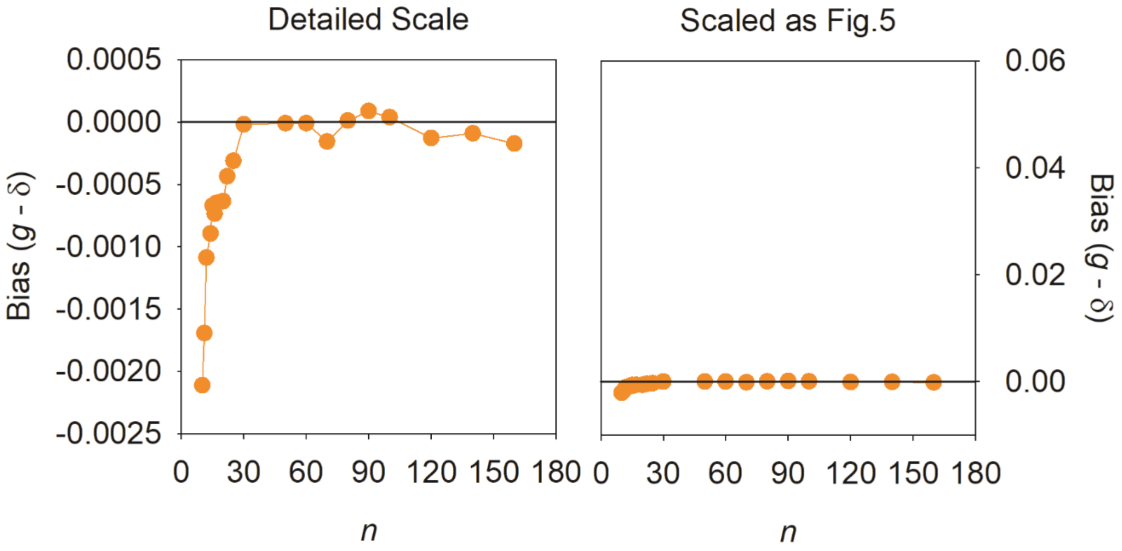
Figure 7 ■ Bias of d_p and g_p as a function of n in various conditions of ω, δ, and ρ in a one-sample paired pooled design using r to approximate ρ. Bias is computed as $(d_p - \delta)$ and $(g_p - \delta)$ so a positive number indicates a positive bias. The d_p became more biased with increasing δ and ρ, but g_p was always within .003 of δ. The sample size n was the n from Table 2 that yielded the width ω in a 95% noncentral t CI. The values of ω were 0.25, 0.4, 0.6, 0.8, and 1.0, and the paired-pooled procedures did not cause a bias in the estimate of δ from g_p. 50K iterations. See Table 1 for notation.



mean differences g_p or to compare CIs for d_p, because the paired-pooled design requires knowledge of ρ to calculate the degrees of freedom ν, and ν is necessary for the calculation of both g_p and the CI. Because ρ is rarely known, it must be approximated by the variables r_W and r_{OP}, and the use of the same formula for a CI that simply replaces ρ with r produces low coverage. In Figures 6 and 7, the calculation of g_p in a pooled-paired design that uses r_W to approximate ρ did not badly bias g_p itself, although the variance of g_p, $Var(g_p)$, was clearly biased (Figure 2). This implies that other methods of forming a CI for the unbiased standardized mean difference that employ the $Var(g_p)$ for the calculations (Cousineau & Goulet-Pelletier, 2021; Fitts, 2021; Goulet-Pelletier & Cousineau, 2018; Hedges & Olkin, 1985) will also have biased results. When using the noncentral t method of Steiger and Fouladi (1997) to calculate paired-pooled CIs with $\hat{\lambda} = d_{\text{P}}\sqrt{n/(2\,(1 - r_W)}$ and $\nu = 2(n-1)/(1 + r_{OP}^2)$, the coverages of the observed CIs are lower than the nominal η^0 at small sample sizes and even at $n = 100$ (Figure 3; Cousineau & Goulet-Pelletier, 2021).

Also shown in Figure 3 is the coverage of a fixed-width Steiger and Fouladi-compliant CI of exact standardized width ω. As sample size increases, the coverage of any fixed-width CI will increase, and at some sample size that coverage will equal the desired η^0. That sample size depends heavily on knowledge of δ and ρ before the experiment begins, because the sample size must be set at the beginning of the experiment and different values of δ and ρ require different sample sizes. This is the same require-

Figure 8 ■ "Bias" of g_p caused by the use of α' values from Figure 4 to generate 95% CIs at 5 levels of ρ and 6 levels of δ across various values of n. The same data are shown at two scales, a detailed scale on the left and the same scaling used in Figure 7 on the right. There is a definite bias at $n < 30$, but it is tiny compared with the bias of d_p. See Table 1 for notation.



ment as the need for ω and δ in Kelley and Rausch (2006) with the added requirement of an estimate of ρ. Tables S1, S2, and S3 give the sample size for a paired-pooled design when δ and ρ are known. Simulations that used the sample r to approximate ρ demonstrate that about 1, 2, or 3 additional subjects are needed in Tables S1, S2, and S3, respectively, when using r to generate a fixed-width CI. This does not mean that the use of the sample size from these tables will yield an observed CI with nominal coverage, however. It is the fixed-width CI that will have nominal coverage on average when using r. An example will help to explain this difference.

Suppose one desires a paired-pooled 95% CI of standardized width $\omega = 0.8$ and one knows a priori that the $\delta = 0.6$ and $\rho = 0.4$. The calculated sample size in Table S2 for $[\omega = 0.8, \delta = 0.6, \rho = 0.4]$ is $n = 32$. If one uses ρ to calculate the CI, then the interpretation of this sample size is exactly the same as that for the two-sample Table 2 in Kelley and Rausch (2006), i. e., that if one uses the selected sample size $n = 32$, the observed CI will have an average width of $\omega = 0.6$ and an average coverage of $\eta*$ = .95. At this n, the average width of the observed CI will equal that of the fixed-width CI and the average CI limits will be identical. The rationale for this in Kelley and Rausch (2006) is the fact that the coverage of the observed CI for a two-sample test in a simulation such as the left side of Figure 3 would have coverages of all observed CIs tracking exactly along the .95 guideline at all values of δ and n (ρ being irrelevant for independent samples). The sample size from the table of Kelley and Rausch does not affect coverage of the observed CI, it simply selects the n that produces an average width of ω. If we reproduced Figure 3 using ρ to calculate the observed CI, the coverages would also track exactly along the .95 guideline, and the selection of n from the table would simply assist in generating an observed CI of average width ω. If one does know that $\delta = 0.6$ and $\rho = 0.4$, both the observed CI (using ρ) and fixed-width CIs will have an average coverage of .95 when using $n = 32$. But this is clearly not what happens when we replace ρ with r.

When approximating ρ with r for $[\omega = 0.8; \delta = 0.6; \rho = 0.4]$, Table S2 advises that the sample size from simulations required about 2 more subjects than the calculation states, so we would expect to use $n = 34$. However, this does not mean that using $n = 34$ will produce an observed CI with $\eta*$ = .95 and $\omega = 0.8$. By inspecting Figure 3 for $\omega = 0.8$ and finding the coverage of the observed CI when $\delta = 0.6$, $\rho = 0.4$, and $n = 34$ we see that the coverage is less than the desired η^0 = .95. In fact, we may not have nominal coverage with the observed CI even if we used $n = 100$. Replacing ρ with r destroys that relationship, and the width (i.e., $w < \omega$) and coverage ($\eta* < \eta^0$) of the observed CI will be too low even with $n = 34$. The coverage of the fixed-width CI will be .95, but only if ρ is actually 0.4. If we do not really know ρ, we cannot select an exact sample size to use from Table S2. Thus, Tables S1 to S3 are not very helpful by themselves unless one is happy with the depressed coverages given for the observed CIs in Figure 3.

One could instead use the fixed-width CI with $n = 34$, and the coverage would average the nominal η^0 assuming that the a priori estimates of δ and ρ were accurate. As seen on the right side of Figure 3 with ordinate scaling on the right, a sample size of 34 can give very different coverages for a fixed-width CI if ρ has been poorly estimated. Many experiments do not have such accurate a priori estimates of δ and ρ. Furthermore, the fixed-width CI is ad hoc and not calculated from first principles based on a given η^0.

This is where the practical coefficients η' and α' are helpful (Table 2, Figure 4). Keeping with the same example, we can look up the regressed value for α' in Table 2 for $n = 34$ and η^0 = 95%, which is between .0467 at $n = 30$ and .0471 at $n = 35$. We can calculate an exact value using the regression equation or we can use .047 (the blue reliability symbols in Figure 4 used 3 digits for 95%). Note that the .047 does not depend importantly on any δ or ρ within the ranges of these parameters that were tested in this article. Figure 4 tells us that this practical coefficient will adjust the mean coverage from its biased value to the nominal η^0 = .95, and Figure 5 tells us that the average width of the interval will be the desired $\omega = 0.8$. Thus, the use of Tables S1 to S3 combined with the use of the adjusted practical coefficient from Table 2 and Figure 5 will yield an observed CI with nominal coverage and the desired width with a paired-pooled design. Tables S1, S2, S3, and 2 were all generated using the supplied software, so an investigator can use the software according to the instructions in the User's Notes to calculate and test most useful values for n, δ, ρ, and ω.

Assumptions of the Method

The calculations used in the Tables and Figures in this article were derived from populations with normal distributions, and the two arms of each repeated measures experiment had equal variances. Equal variances are necessary in order for a pooled error term S_p to make sense. The alternative procedure using S_D requires only a normal distribution of difference scores, but its d_D cannot be compared with d_p from a two-sample test. An alternative procedure not discussed here is the use of the error term from the control condition only with $\nu = n - 1$ (Becker, 1988). The calculation for r_W in Equation 14 requires knowledge not only of the Pearson r and S_p but also of the individual standard deviations S_1 and S_2. For a meta-analyst to try to calculate a corrected CI using the present methods from a published paper that used an incorrect ν or that used d_D instead of d_p, it means that the paper must have reported the correlation r and both S_1 and S_2. Using r or r_{OP} in place of r_W in the calculation of $\hat{\lambda} = d_\mathrm{P}\sqrt{n/(2\,(1-r_W)}$ can produce poor results (Fitts, 2020). In addition, a legitimate d_D may have been reported where the difference scores were normal in shape, but the distributions of the experimental and control conditions were not homogeneous as required for d_p. Violations of normality and homogeneity of variances are well known to adversely influence results, and such violations have not yet been studied with the current procedures. Bootstrap and other alternative procedures (Kelley, 2005) should be studied explicitly for paired-pooled tests.

Generalization to other CI protocols

I selected the Steiger and Fouladi (1997) protocol for computing CIs because it is the most widely discussed in the literature for noncentral t procedures. Cousineau and Goulet-Pelletier (2021) tested eight CI protocols to create approximate CIs for a paired-pooled design. They found that a new "Adjusted Λ'" approach produced approximate intervals for which the average coverage of the CI in simulations was sometimes greater than but never less than the nominal η^0. The "practical coefficient" technique does not depend on any one of these CI protocols, so it could be used with the Adjusted Λ' approach as well. The approximation using the Steiger and Fouladi method produces coverages below η^0, so the practical coefficient η' was larger than η^0 to bring the overall coverage up to η^0. Its use with the Adjusted Λ' could use a practical coefficient η' smaller than η^0 to bring the coverage down to η^0 in those circumstances where it was too high.

Authors' note

The author is a sponsored retiree at the University of Washington.

References

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

Bond, C. F., Jr., Wiitala, W. L., & Richard, F. D. (2003). Meta-analysis of raw mean differences. *Psychological Methods*, *8*, 406–418.

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

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

Cousineau, D. (2020). Approximating the distribution of Cohen's dp in within-subject designs. *The Quantitative*

Methods for Psychology, *16*, 418–421. doi:/10.20982/tqmp.16.4.p418

Cousineau, D., & Goulet-Pelletier, J.-C. (2021). A study of confidence intervals for Cohen's dp in within-subject designs with new proposals. *The Quantitative Methods for Psychology*, *17*, 51–75. doi:10.20982/tqmp.17.1.p051

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

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

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. doi:10.20982/tqmp.16.4.p281

Fitts, D. A. (2021). Expected and empirical coverages of different methods for generating noncentral t confidence intervals for a standardized mean difference. *Behavior Research Methods*, *53*, 2412–2429. doi:10.3758/s13428-021-01550-4

Fitts, D. A. (2022a). Absolute precision confidence intervals for unstandardized mean differences using sequential stopping rules. *Behavior Research Methods*, *Published online*, 1–34. doi:10.3758/s13428-022-01896-3

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

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., & Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego, CA: 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. doi:10.1177/0013164404264850

Kelley, K. (2007). Methods for the behavioral, educational, and social sciences: An r package. *Behavior Research Methods*, *39*, 979–984.

Kelley, K., Maxwell, S. E., & Rausch, J. R. (2003). Obtaining power or obtaining precision: Delineating methods of sample size planning. *Evaluation & the Health Professions*, *26*, 258–287. doi:1.1177/0163278703255242

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. doi:10.1037/1082-989X.11.4.363

Olkin, I., & Pratt, J. W. (1958). Unbiased estimation of certain correlation coefficients. *The annals of mathematical statistics*, 201–211.

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

Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594–604.

Open practices

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

Citation

Fitts, D. A. (2022b). Point and interval estimates for a standardized mean difference in paired-samples designs using a pooled standard deviation. *The Quantitative Methods for Psychology*, *18*(2), 207–223. doi:10.20982/tqmp.18.2.p207

Received: 15/03/2022 ~ Accepted: 15/07/2022