

# Bootstrap BCa confidence intervals for a standardized mean difference in a paired samples design using a pooled standard deviation

Douglas A. Fitts<sup>a</sup>

<sup>a</sup>University of Washington

**Abstract** ■ A standardized mean difference in a paired samples design using a pooled standard deviation,  $d_p$ , requires knowledge of the population correlation,  $\rho$ , in order to generate an accurate estimate of a noncentral  $t$  confidence interval (CI). Using the observed empirical correlation introduces a source of random error which causes bias in the variance of Hedges'  $g_p$  and in the coverage of the CI which is most notable at sample sizes below 50 pairs. By contrast, a bootstrap BCa (bias-corrected and accelerated) CI for  $d_p$  does not require an estimate of  $\rho$ , and with unpaired samples the BCa method produces a CI for  $d_p$  with nominal coverage when the underlying distribution is normal. New simulations demonstrate that the coverages of BCa CIs for  $d_p$  with paired samples and a pooled error term are also nominal across a wide range of values of  $\rho$  and the population effect sizes except at very small sample sizes ( $n = 10$  pairs) where the coverage is slightly depressed. Thus, the BCa CI is more accurate than a noncentral  $t$  CI when using  $d_p$  with a normal distribution. Because the estimation of  $\rho$  affects the variance of  $g_p$  rather than the central tendency, the value of  $g_p$  can be reported along with the BCa CI for  $d_p$ . However, simulations to generate a BCa CI for the unbiased  $g_p$  instead of  $d_p$  produced biased coverage results with 50 or fewer pairs, presumably because of the bias in the variance of  $g_p$  when estimating  $\rho$ . Software for generating the intervals is provided.

**Keywords** ■ Confidence Interval, Bootstrap, Cohen's  $d$ , Paired Samples, Point Estimation, Accuracy in Parameter Estimation.

**Acting Editor** ■ Denis Cousineau (Université d'Ottawa)

**Reviewers**  
■ One anonymous reviewer

[dfitts@uw.edu](mailto:dfitts@uw.edu)

[10.20982/tqmp.21.3.p125](https://doi.org/10.20982/tqmp.21.3.p125)

## Introduction

Psychology researchers often compare the standardized effect sizes from different studies, and increasingly these effect sizes are reported as confidence intervals (CIs) around a standardized effect size. When a study consists of two sets of measurements based on different treatments the standardized effect is often Cohen's  $d$  (Cohen, 1988), which is the difference between the means divided by the standard deviation of measurements. Because  $d$  has a positive bias with small sample sizes, it is often transformed to its unbiased equivalent Hedges'  $g$  (Hedges, 1981). The interpretation of  $d$  is most reliable when the distributions are normal and homoschedastic (Hedges, 2024). When this is true, the best estimate of the common standard deviation is a pooled

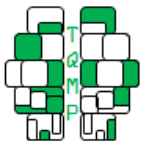
standard deviation,  $S_p$ , based on both sets of scores. For equal sample sizes this would be

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

and the  $d_p$  would be calculated as the raw difference between the means,  $D$ , divided by this pooled standard deviation, or

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

where the subscript  $p$  indicates that the  $d$  was formed using a pooled standard deviation instead of, for example, the standard deviation of the difference scores (known as  $d_z$  in tools such as GPower). The corresponding unbiased



Hedges'  $g$  is  $g_P$ :

$$g_P = (d_P)J(\nu);$$

$$J(\nu) = \frac{\Gamma\left(\frac{\nu}{2}\right)}{\sqrt{\frac{\nu}{2}} \Gamma\left(\frac{\nu-1}{2}\right)} \quad (3)$$

where  $\Gamma$  is the gamma function and  $\nu$  is the degrees of freedom. For independent samples instead of paired samples one can then calculate a CI for  $d_P$  using the noncentral  $t$  method of Steiger and Fouladi (1997) after which the investigator reports  $g_P$  as the point estimate and the CI for  $d_P$  as the interval estimate (Fitts, 2021). Unfortunately, one cannot use the same method to create a CI for  $d_p$  with a paired samples experiment (Fitts, 2020) because the degrees of freedom,  $\nu$ , for the paired samples experiment depends on the value of the correlation between the two sets of measures in the population (i.e.,  $\rho$ ), which is seldom known:

$$\text{Paired-pooled - population: } \nu = 2(n - 1)/(1 + \rho^2) \quad (4)$$

where  $n$  is the number of pairs of measurements (Cousineau, 2020). Substituting the sample  $r$  for  $\rho$  in this equation introduces a source of random error, and consequently the transformations of  $d_p$  are not distributed as a noncentral  $t$  (Cousineau, 2020; Cousineau & Goulet-Pelletier, 2021; Fitts, 2021). Fitts (2022) developed transformations that could approximately adjust for the bias in coverage when using the Steiger and Fouladi (1997) noncentral  $t$  method but different clumsy transformations were required for each chosen confidence coefficient (e.g., 90, 95 or 99%).

What is needed, then, is a simpler method for generating a CI for  $d_p$  in a paired samples design that does not depend either on the noncentral  $t$  distribution or the exact degrees of freedom. One possibility is a bootstrap BCa CI (Kelley, 2005). The bootstrap CI is generated by resampling data with replacement from the observed sample a large number of times, and then finding the desired percentiles from the new ad hoc distribution of resampled  $d_p$  values (e.g., 2.5 and 97.5 for a two-sided 95% CI). This CI is called a bootstrap percentile CI, which can be adjusted for bias and accelerated (BCa) using an algorithm provided by Efron to create the bootstrap BCa CI (Efron, 1985, 1987; Efron & Tibshirani, 1993). The bootstrap method does not presume that the results are distributed as a noncentral  $t$ , and it does not require the calculation or use of the degrees of freedom. When used with independent samples and normally distributed data the method generates CIs with nominal coverage for all but the smallest sample sizes ( $n = 10$  per group) (Kelley, 2005). Repeating this with dependent samples requires a coding of new software to calculate the bootstrap BCa CI, because existing programs work only with independent samples.

Thus, the goals of this project were to create new algorithms that allow computation of the BCa CI for  $d_p$  from a paired-samples design and to explore whether the observed coverage of such CIs is nominal when used with normally distributed, homoschedastic data.

## Method

### Bootstrap CI

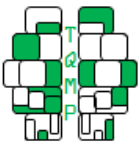
Kelley (2005) adapted the bootstrap BCa method of generating CIs for use with  $g_p$  in independent samples. His method for two-sided intervals was to collect data from  $n_1 + n_2 = N$  subjects, to collect 10,000 random samples of  $n_1$  and  $n_2$  subjects with replacement from that original data, to calculate 10,000  $d_p$  values, and then to convert each to  $g_p$ . Using the distribution of 10,000  $g_p$  resampled values he then calculated percentiles at  $\alpha/2$  and at  $1 - \alpha/2$ , where  $\alpha$  equals one minus the desired confidence level. For example, to create a .950 CI he found the percentiles at .025 and at .975 in the cumulative resampled distribution to generate the percentile CI with  $\alpha = .05$ . He then applied the BCa correction, described below, to the percentile CI to generate the bootstrap BCa confidence interval for  $g_p$  (Equation 3). The new algorithm presented here uses the same general method with the exceptions that (1) the research design was for paired samples instead of independent samples; and (2) the resampling consisted of 10,000  $d_p$  samples instead of converting to  $g_p$ . One rationale for this was because the bootstrap CI needed to avoid the calculation of degrees of freedom that would have been required for  $g_p$  (see Equation 3). In addition, Fitts (2021) found that the noncentral  $t$  CI for  $d$  had more accurate coverage than a CI for  $g$ . Both methods are tested in the following simulations.

The resampling for a paired samples design necessarily required samples of  $n$  pairs of scores rather than sampling values from each of the two independent groups as performed by Kelley (2005). A bootstrap percentile CI was then calculated using the desired  $\alpha$ , and we call each resampled  $d$  value  $d^*$  to distinguish it from the original  $d$ . For a bias-correction value,  $\hat{z}_0$ , one then calculates the proportion of  $d^*$  values that are less than the  $d$  calculated from the original sample, and computes:

$$\hat{z}_0 = \Phi^{-1} \left( \frac{\#(d^* < d)}{B} \right) \quad (5)$$

where  $\Phi^{-1}$  is the inverse of the standard normal cumulative distribution function and  $\#$  denotes "the number of".

The acceleration value,  $\hat{a}$ , depends on the mean of a jackknife procedure whereby  $d$  is calculated  $n$  times after the  $i$ th pair has been deleted ( $i = 1, \dots, n$ ), where  $n$  is the number of pairs of scores. The acceleration value is calcu-



lated as:

$$\hat{a} = \frac{\sum_{i=1}^N (\tilde{d} - d_{(-i)})^3}{6 \left( \left( \sum_{i=1}^N (\tilde{d} - d_{(-i)})^2 \right)^{3/2} \right)} \quad (6)$$

where  $\tilde{d}$  is the mean of the  $N$  means of the jackknifed samples with  $n - 1$  observations in each sample. The values  $\hat{z}_0$  and  $\hat{a}$  are then used to calculate lower and upper cumulative probability values:

$$P_{Low} = \Phi \left( \hat{z}_0 + \frac{(\hat{z}_0 + z_{\alpha/2})}{1 - \hat{a} (\hat{z}_0 + z_{\alpha/2})} \right) \quad (7)$$

and

$$P_{Up} = \Phi \left( \hat{z}_0 + \frac{(\hat{z}_0 + z_{(1-\alpha/2)})}{1 - \hat{a} (\hat{z}_0 + z_{(1-\alpha/2)})} \right). \quad (8)$$

The lower and upper bounds of the CI are determined by finding the quantiles of the ad hoc distribution of  $d^*$  corresponding to these two cumulative probability values. If  $\hat{z}_0$  and  $\hat{a}$  are both zero, these cumulative probability values revert to the original  $\alpha/2$  and  $1 - \alpha/2$ , and the bootstrap BCa CI limits are identical to those of the uncorrected percentile distribution.

**Degrees of freedom for  $g_p$**

For all experiments a  $g_p$  (Equation 3) was estimated from Equation 4 using the formula for the approximate degrees of freedom suggested by Cousineau and Goulet-Pelletier (2021):

$$\text{Paired-pooled - sample: } \nu = 2(n - 1)/(1 + r_{OP}^2) \quad (9)$$

where  $n$  is the number of pairs of scores and  $r_{OP}$  is the unbiased correlation coefficient. Their paper found little difference between using  $r_{OP}$  and the uncorrected  $r$ . This paper uses an approximation of  $r_{OP}$  suggested by Olkin and Pratt (1958), in their equation 2.7 that eliminates integration:

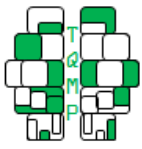
$$r_{OP} = r \left[ 1 + \frac{(1 - r^2)}{2(n - 3)} \right] \quad (10)$$

The following is a tiny example showing the calculations with 4 pairs of scores (measures 1 and 2). The simulation assumed  $\delta = 0.5$  and  $\rho = .5$  with a two-tailed  $\alpha = .05$ . Calculations are rounded to 3 decimal digits for brevity of presentation which is not recommended in practice.

The first step is to calculate the overall  $d_p$  from the two means and standard deviations. A pooled standard deviation (sdpool) is calculated using Equation 1, and then the

overall  $d_p$  is calculated using Equation 2. The best procedure uses this  $d_p$  instead of calculating  $g_p$  from Equations 3 and 4, so that procedure is omitted. The computer is used to resample 4 scores randomly with replacement to calculate a  $d_p^*$  from the 4 resampled scores. This procedure is repeated until  $B = 10,000$  resampled  $d_p^*$  values have been assembled into an array and the array is sorted from smallest to largest. The percentile 95% CI can be calculated directly from this array of  $d_p^*$  values by finding percentiles at 2.5% and 97.5%.

We now need to adjust the CI for bias and accelerate using the BCa formulas 5, 6, 7, and 8. For Equation 5, we count the number of these 10,000 resampled  $d_p^*$  values that are less than the overall  $d_p$  from our original sample (i.e.,  $<0.6$ ). In our simulation, this number was 4937, so  $4937/10000 = 0.4937$  is the proportion of  $d_p^*$  values less than the original  $d_p$  in this particular case. The inverse standard normal cumulative distribution function is used to calculate a quantile corresponding to this proportion (interpreted as a probability). This can be calculated in Excel with the command =NORM.INV(0.4937, 0, 1), and the answer is  $\hat{z}_0 = -0.01579$  to complete Equation 5. To calculate the bias coefficient we need to calculate jackknifed means and standard deviations and  $d_{p(-i)}$  values for 4 sets of means with  $n = 3$  each by successively dropping one pair of means 4 times in Equation 6. These are indicated in the worked example by the lines beginning “Drop”. For example, in line “Drop 1” the calculated mean was 0.509 for measure 1 from the scores for the three subjects 2, 3, and 4 (dropping subject 1), and the calculated mean was -0.440 for measure 2 from the second scores for the same 3 subjects. A pooled standard deviation is calculated from the two standard deviations for measures 1 and 2 with  $n = 3$ , and a  $d_{p(-i)}$  is calculated as usual using Equation 2. The  $\tilde{d}$  value is the mean of these four successively drawn  $d_{p(-i)}$  values, in this case, 0.587. Now we can calculate the bias for each  $d_{p(-i)}$  value by subtracting  $(\tilde{d} - d_{(-i)})$ . Equation 5 uses the sum of these squared and cubed values so we calculate those as bias squared and bias cubed and sum them. The sum of the cubed biases is the numerator for the calculation of  $\hat{a}$  in Equation 6. The denominator requires the sum of the squared biases taken to the 1.5 power, and multiplying the result by 6:  $\hat{a} = 0.040/0.590 = 0.068$ . Thus, we have our two coefficients  $\hat{z}_0 = -0.01579$  and  $\hat{a} = 0.068$  for Equations 7 and 8. If these were both 0.000, it would mean that there is no need for a BCa correction and the percentile CI values can be used intact. Here they are nonzero, so we proceed to Equations 7 and 8 to calculate the lower and upper probability values for our adjusted BCa CI. This requires a z value for our  $\alpha/2 = 0.025$  and  $1 - \alpha/2 = 0.975$ , which are determined from the inverse normal distribution in Excel as =NORM.INV(0.025, 0,



**Table 1** ■ Worked Example

Subject	Measurement 1	Measurement 2							
1	0.069	0.304							
2	0.172	0.541							
3	1.968	-0.038							
4	-0.613	-1.823							
M	0.399	-0.254	sdpool	Overall $d_p$					
SD	1.102	1.073	1.088	0.6					
Jackknife $n = 3$									
	$m_1$	$sd_1$	$m_2$	$sd_2$	sdpool	$d_{(-i)}$	bias $\tilde{d} - d_{(-i)}$	bias cubed	bias squared
Drop 1	0.509	1.323	-0.44	1.232	1.278	0.743	-0.156	-0.004	0.024
Drop 2	0.475	1.337	-0.519	1.142	1.243	0.799	-0.212	-0.01	0.045
Drop 3	-0.124	0.427	-0.326	1.302	0.969	0.208	0.379	0.054	0.144
Drop 4	0.736	1.068	0.269	0.291	0.783	0.597	-0.01	0	0
					sum	2.347		0.04	0.213
					$\tilde{d}$	0.587			
From Equation 6:									
$\hat{a}$ num	0.040								
$\hat{a}$ den	0.590								
$\hat{a}$	0.068								

1) and =NORM.INV(0.975, 0, 1), or the usual  $\pm 1.96$  for  $\alpha = .05$ . With  $\hat{z}_0 = -0.01579$  and  $\hat{a} = 0.068$  we can complete Equations 7 and 8 to find intermediate z values of  $z_1 = -1.759$  and  $z_2 = 2.223$ , which yield the probability values for our PLow and PUp, i.e., =NORM.DIST(-1.759, 0, 1, 1) which is .039 and =NORM.DIST(2.223, 0, 1, 1) which is .987. The computer then finds the percentiles in the ad hoc resampled array of 10,000  $d_p^*$  values corresponding to the probabilities 3.9% and 98.7% by linear interpolation, and the final BCa CI is [-2.867, 2.120]. See Software section for additional help.

**Simulations**

Simulations consisted of 100,000 replications of experiments having normally distributed scores with equal variances in both conditions and varying levels of  $\rho$  (.0, .2, .4, .6, and .8), population standardized effect size  $\delta$  (0.0, 0.2, 0.4, 0.6, 0.8, and 1.0), sample size  $n$  (10, 20, 30, 40, 50, 75, 150, and 250 pairs) and confidence coefficient (90, 95, and 99%). For each experiment, records were made of the  $d_p, g_p, \nu, r, r_{OP}, D, S_p$  (see Equations 1, 2, 3, 9, 10), both the percentile and BCa CIs and their respective full widths, and a notation of whether or not each CI included  $\delta$ . The average of each variable over the 100,000 replications were calculated along with the coverage, defined as the number of experiments where the CI included  $\delta$  divided by 100,000.

Bias was calculated for the averages of  $d_p$  as  $(d_p - \delta)$ ,

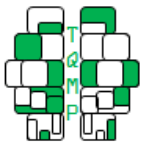
for  $g_p$  as  $(g_p - \delta)$ ,  $r$  as  $(r - \rho)$ , and  $r_{OP}$  as  $(r_{OP} - \rho)$ . Thus a positive or negative result indicated a positive or negative bias of the estimator.

**Software**

Simulations were conducted using programs written in C. Executable 64-bit code for a PC, source code, and a help file are all available at OSF site <https://osf.io/txumj/>.

Two programs are available, one for calculating bootstrap BCa CIs from raw data that conforms to a paired-samples design with a pooled error term. The second is a program that can be used to replicate the simulation studies reported here. A script in R to calculate a bootstrap BCa CI from raw data is also provided at the OSF site for the convenience of R programmers. An Excel file showing the calculations from the above worked example is also included.

The simulation studies use original code based on an R program published by Kelley (2005) for computing bootstrap BCa nonparametric CIs for a standardized mean difference with independent data that has been translated to C and further adapted here for the paired-pooled design using a B array of resampled  $d_p$  values instead of  $g_p$  values. The number of resampled  $d_p$  values to use, B, was always 10,000 in these simulations, but the user should be aware that this number should be increased if used with  $\alpha$  values more stringent than .01. The program detects percentiles in this ad hoc array of resampled values at probability val-



ues corresponding to  $\alpha/2$  and  $1 - \alpha/2$ . With  $\alpha = .001$  these percentiles would represent scores  $(10,000)(.0005) = 5$  and  $(10,000)(.9995) = 9995$  in the sorted array which are very close to the boundaries, so  $B \geq 20,000$  would be preferred. This increase in random resampling obviously increases computing time for simulations with 100,000 iterations of each experiment, but it is tolerable when computing a CI using raw data from a single experiment.

The programs to compute a bootstrap BCa CI for a standardized mean difference (present study; Kelley, 2005) have a peculiarity that distinguishes them from similar efforts with unstandardized data, and that has to do with equation 6. For a single sample, or even for an unstandardized mean difference between two samples, the mean of the  $N$  jackknifed means is equal to the overall mean of the original sample. That equality between the mean of the jackknifed means and the overall mean of the full sample is destroyed by the division by the standard deviation to form the  $d_p$ . Here I distinguish between the mean of the jackknifed  $d$  values (i.e.,  $\tilde{d}$ ) and the overall  $d$  for the original full dataset, which are not equal for the  $d_p$ . In the calculation of variable  $\hat{a}$  (Equation 6), where the deviations of the individual jackknifed  $d_p$ s from the mean of the  $N$  jackknifed  $d_p$ s are summed, the mean of the jackknifed  $d_p$ s is used instead of the overall  $d_p$  of the original sample. The difference is small (in the above example,  $d_p = 0.6$  and  $\tilde{d} = 0.587$ ) and the method appears to work well in practice. For unstandardized values there is no difference between  $D_p$  and  $\tilde{D}$ . The  $d$  in equation 5 remains the overall  $d_p$  of the original sample both in the program of Kelley (2005) and here. This difference between standardized and unstandardized analyses is illustrated in the Excel file of the worked example at the OSF site.

## Results

### Simulations using resampled $d_p$

A complete set of simulations was conducted with 10,000 resampled  $d_p$  values to generate the percentile CI. A limited set of simulations conducted with 10,000 resampled  $g_p$  values as a demonstration of the bias is presented later.

Before examining coverage, it is useful to examine the bias of the estimators obtained during the simulations. The bias of  $d_p$  was greatest at small sample size as expected, and the conversion to  $g_p$  eliminated the bias as expected. As a reminder, the previous study with paired samples (Fitts, 2022) observed a bias in the variance of  $g_p$ , not in the central tendency, and here we consider the mean  $g_p$ . The bias of  $r$  was greatest at small sample sizes as expected, and the conversion to the approximate  $r_{OP}$  virtually eliminated the bias.

In general, the coverage of the percentile interval was

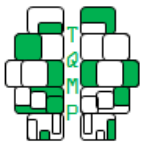
farther from the nominal value for the confidence coefficient than the BCa interval, with coverage of the percentile interval being lower than for the BCa interval, so data will be presented for the BCa interval.

The mean coverage data are summarized in Table 2 for each value of  $\delta$  averaged across all values of  $\rho$  for the 90, 95, and 99% confidence coefficients. In general, larger values of  $\delta$  produced very slightly lower coverages than smaller values of  $\delta$  at small sample sizes, but overall the coverages were very nearly nominal for all confidence coefficients. The smallest sample size of  $n = 10$  pairs fared worst with low coverages. With sample sizes of 20 pairs or more, all averages rounded to two decimal places were .89-.90 at 90%, .94-.95 at 95%, and .98-.99 at 99%.

The mean coverage data are summarized in Table 3 for each value of  $\rho$  averaged across all values of  $\delta$  for the 90, 95, and 99% confidence coefficients. Again, for all sample sizes of 20 pairs or more, all averages rounded to two decimal places were .89-.90 at 90%, .94-.95 at 95%, and .98-.99 at 99%. Larger effect sizes had slightly lower coverages than smaller effect sizes, and the coverage was most below nominal with  $n = 10$  pairs.

The residual very small bias tended to be in the direction of lower than nominal coverage (e.g., .89 instead of .90). Circumstances that call for the most caution in interpreting the coverage as nominal include combinations of very small sample sizes, very large effects, and very high correlations, situations which rarely apply in cognitive, social, or educational psychology. Areas such as neuroscience and physiological psychology will rarely encounter this problem because the dependent variables tend to be more directly comparable in unstandardized form (e.g., number of double-labeled neurons or differences in mean arterial pressure in mmHg).

The average widths of the BCa CIs at all values of  $\delta$ ,  $\rho$ , and  $n$  are presented in Table 4 for 90% CIs, Table 5 for 95% CIs, and Table 6 for 99% CIs. As expected, larger sample sizes generated narrower widths. Although the investigator is unlikely to know  $\delta$  and  $\rho$  when planning an experiment, the tables are useful for providing a rough idea of the sample size required to generate a CI of a given width. For example, if the investigator desires to generate a 95% CI with a width of at most 0.5 standard deviations, one can see in Table 5 that the average widths of all CIs were below 0.5 with a sample size of 150 pairs regardless of the final value of  $\delta$  or  $\rho$  (width range 0.20 to 0.48). The ranges can be reduced considerably if the investigator has reliable a priori information about  $\delta$  or  $\rho$ . In general, the width increased with increasing effect size and decreased with increasing correlation. Negative correlations were not considered here because the point of matching subjects is to produce a positive correlation that increases power, and



**Table 2** ■ Mean coverages of BCa CIs for each value of  $\delta$  averaged across all values of  $\rho$  for the 90, 95, and 99% confidence coefficients at different sample sizes using the  $d_p$  ad hoc resampling array. For sample sizes 20 and above, the coverage was either at or only very slightly below the nominal value.

$\delta$	$n$	90%	95%	99%	$n$	90%	95%	99%
0	10	0.892	0.937	0.976	50	0.902	0.950	0.989
0.2		0.890	0.936	0.975		0.901	0.950	0.989
0.4		0.887	0.933	0.974		0.900	0.949	0.989
0.6		0.881	0.928	0.971		0.899	0.948	0.989
0.8		0.874	0.923	0.969		0.897	0.947	0.988
1		0.866	0.917	0.964		0.895	0.945	0.987
0	20	0.900	0.947	0.986	75	0.900	0.950	0.990
0.2		0.899	0.947	0.986		0.901	0.950	0.989
0.4		0.897	0.945	0.985		0.901	0.949	0.989
0.6		0.894	0.944	0.984		0.900	0.948	0.989
0.8		0.890	0.940	0.983		0.898	0.947	0.989
1		0.887	0.937	0.982		0.896	0.947	0.988
0	30	0.901	0.949	0.988	150	0.900	0.950	0.990
0.2		0.901	0.948	0.988		0.901	0.950	0.990
0.4		0.900	0.948	0.988		0.900	0.950	0.990
0.6		0.896	0.947	0.987		0.900	0.950	0.990
0.8		0.895	0.944	0.986		0.899	0.949	0.989
1		0.891	0.942	0.985		0.898	0.949	0.989
0	40	0.902	0.950	0.989	250	0.900	0.950	0.990
0.2		0.901	0.949	0.989		0.900	0.950	0.990
0.4		0.900	0.948	0.988		0.900	0.949	0.990
0.6		0.898	0.948	0.988		0.900	0.949	0.990
0.8		0.896	0.946	0.987		0.900	0.950	0.990
1		0.894	0.945	0.987		0.899	0.949	0.990

a negative correlation could make power even worse than an independent samples test. In the context of a CI, an increase in positive correlation greatly reduces the width of the CI, thus increasing the accuracy of the estimation of the parameter. A negative correlation would do the opposite.

**Simulations using resampled  $g_p$**

The previous simulations were repeated exactly for sample sizes of 10, 20, 30, 40, 50, and 75 except that the resampled distribution was 10,000 estimated values of  $g_p$  instead of  $d_p$ . For each simulated experiment and each resampled or jackknifed experiment, the  $d_p$  was immediately converted to  $g_p$  using Equations 3, 4, 9, and 10, and used instead of  $d_p$  in the calculations of Equations 5, 6, 7, and 8. The coverage results for each effect size averaged over all correlations and for each correlation averaged over all effect sizes for each confidence coefficient of 90, 95, and 99% are presented in Table 7. The data for sample sizes of 75 and above are identical to the  $d_p$  data from the previous tables and are omitted for brevity. At sample sizes of 10, 20, and 30 the coverages are lower than for the simulations using  $d_p$ . At sam-

ple sizes of 40 and above the data begin to converge on the  $d_p$  results from previous tables. The poorer performance from resampled  $g_p$  distributions at small sample sizes is consistent with previous research demonstrating a bias in the variance of  $g_p$  rather than the mean of  $g_p$ , and the distortion of the variance of  $g_p$  would be worst at small sample sizes where the estimate of the variance is less stable. The distortion of the variance is attributable to the random error introduced in the calculation of the degrees of freedom used to calculate  $g_p$  by using the sample  $r_{OP}$  (Equation 9) instead of  $\rho$  (Equation 4), which would also be worst with small sample sizes. At large sample sizes  $d_p$  and  $g_p$  converge (Hedges, 1981).

**Variation of Width**

Coverage is a principal measure of the quality of a CI. For CIs with equal coverage, another measure of quality is its width, with narrower CIs providing better estimates of the location of the parameter  $\delta$  than wider CIs. For a given confidence level and sample size, the width will vary as a random variable with different random samples. With

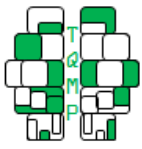


Table 3 ■ Mean coverages of BCa CIs for each value of ρ averaged across all values of δ for the 90, 95, and 99% confidence coefficients at different sample sizes using the d\_p ad hoc resampling array. For sample sizes 20 and above, the coverage was either at or only very slightly below the nominal value.

Table with 9 columns: rho, n, 90%, 95%, 99%, n, 90%, 95%, 99%. Rows represent different sample sizes (10, 20, 30, 40) and rho values (0, 0.2, 0.4, 0.6, 0.8).

forms of CIs using a noncentral t such as those suggested by Steiger & Fouladi (1997) or Hedges & Olkin (1985) the randomness of independent samples is the only source of variability of the width because each random sample has one unique solution to the location of the CI's bounds (Fitts, 2021). The same is not true of bootstrap CIs for which the width can vary from run-to-run of a single set of data because of the requirement for random resampling to form the ad hoc sampling distribution that is the basis of the bootstrap. It is not unreasonable to wonder how much extra variability this feature of the method adds. One way to measure this is to compare the variability of repeated calculations of the width from a single fixed sample with the variability of calculations of the width from many different random samples based on identical parameters.

The simulations used the same methods as previous d\_p experiments, drawing 50 different random samples from populations with varying delta (0.0, 0.5, or 1.0), varying rho (.0, .2, .4, .6, or .8), and varying n (10, 20, 30, 40, 50, 75, 150, or 250 pairs) for a total of 50 \* 3 \* 5 \* 8 = 6,000 different random samples. Then, for each individual random sample, the width of the CI was calculated by repeating the BCa CI algorithm on the same fixed data 50 times, and the average standard deviation of the widths was calculated for estimates of the CI from the same data (FIXED dataset cal-

culations). For each random sample, the mean width of 50 FIXED calculations was determined as a stable estimate of the mean width for a given sample and the mean and standard deviation of these means was calculated for the different random samples (RANDOM dataset calculations) based on identical parameters. The standard deviations for the RANDOM calculations should have little contribution from the variation caused by repeated sampling of the same dataset. The mean width was calculated as the mean of 50 different random samples based on the means of the 50 FIXED sample calculations (i.e., the mean of means). The standard deviation of the width was calculated twice, once for the FIXED dataset and once for the RANDOM dataset.

The data are presented as summary means for each value of delta collapsed across all values of rho (Figure 1) and again for each value of rho collapsed across all values of delta (Figure 2). In each figure, the top panel is the mean width for each condition and the bottom panel is the standard deviations for both the FIXED (dashed lines) and RANDOM datasets (solid lines). From Figure 1 it is apparent that differences in delta contribute very little to differences in the standardized width, and from Figure 2 it is apparent that differences in rho cause an obvious reduction in width with increasing correlations. Naturally, the mean widths of these 50 stabilized means of 50 samples are similar to but less ac-

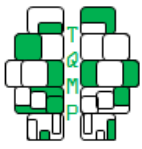


Table 4 ■ Average 90% BCa CI width for each value of  $\delta$ ,  $\rho$  and  $n$ .

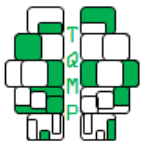
$\delta$	$\rho$	npairs							
		10	20	30	40	50	75	150	250
0	0	1.58	1.07	0.87	0.75	0.66	0.54	0.38	0.29
0.2	0	1.58	1.07	0.87	0.75	0.67	0.54	0.38	0.30
0.4	0	1.59	1.08	0.87	0.75	0.67	0.55	0.38	0.30
0.6	0	1.60	1.09	0.88	0.76	0.68	0.55	0.39	0.30
0.8	0	1.61	1.10	0.89	0.77	0.69	0.56	0.40	0.31
1	0	1.63	1.12	0.91	0.79	0.70	0.57	0.40	0.31
0	0.2	1.42	0.96	0.78	0.67	0.60	0.48	0.34	0.26
0.2	0.2	1.42	0.96	0.78	0.67	0.60	0.49	0.34	0.26
0.4	0.2	1.43	0.97	0.78	0.68	0.60	0.49	0.35	0.27
0.6	0.2	1.44	0.98	0.79	0.68	0.61	0.50	0.35	0.27
0.8	0.2	1.46	1.00	0.81	0.70	0.62	0.51	0.36	0.28
1	0.2	1.48	1.02	0.83	0.71	0.64	0.52	0.37	0.28
0	0.4	1.25	0.84	0.68	0.58	0.52	0.42	0.30	0.23
0.2	0.4	1.25	0.84	0.68	0.58	0.52	0.42	0.30	0.23
0.4	0.4	1.26	0.85	0.69	0.59	0.53	0.43	0.30	0.23
0.6	0.4	1.28	0.87	0.70	0.60	0.54	0.44	0.31	0.24
0.8	0.4	1.30	0.89	0.72	0.62	0.55	0.45	0.32	0.25
1	0.4	1.32	0.91	0.74	0.64	0.57	0.47	0.33	0.25
0	0.6	1.04	0.69	0.56	0.48	0.42	0.34	0.24	0.19
0.2	0.6	1.05	0.70	0.56	0.48	0.43	0.35	0.24	0.19
0.4	0.6	1.06	0.71	0.57	0.49	0.44	0.35	0.25	0.19
0.6	0.6	1.08	0.73	0.59	0.51	0.45	0.37	0.26	0.20
0.8	0.6	1.11	0.76	0.61	0.53	0.47	0.38	0.27	0.21
1	0.6	1.15	0.79	0.64	0.56	0.50	0.41	0.29	0.22
0	0.8	0.76	0.50	0.40	0.34	0.30	0.24	0.17	0.13
0.2	0.8	0.77	0.51	0.40	0.35	0.31	0.25	0.17	0.13
0.4	0.8	0.79	0.53	0.42	0.36	0.32	0.26	0.18	0.14
0.6	0.8	0.83	0.56	0.45	0.39	0.35	0.28	0.20	0.15
0.8	0.8	0.88	0.61	0.49	0.42	0.38	0.31	0.22	0.17
1	0.8	0.95	0.66	0.54	0.47	0.42	0.34	0.24	0.19

curate than the mean widths of the 100,000 samples in Tables 3 to 5, so the tabled values are preferred for planning purposes. In both figures it is clear that the standard deviation of the FIXED calculations is far less than the standard deviation of the RANDOM calculations.

### Discussion

A task force on statistical inference in psychology encourages CIs on effect sizes, and standardized effect sizes should be presented when the measurements are not meaningful or comparable on a practical level (Wilkinson & the Task Force on Statistical Inference, 1999). Cohen's  $d$  is one example of a standardized effect size based on two means (Cohen, 1988) and the means can be derived from two independent groups, from two repeated measures on a single group, or from two groups of subjects that have been

matched on a positively correlated variable that is relevant to the dependent variable of interest. For independent samples from normally distributed measures with equal variances and sample sizes (Hedges, 2024) the best standard deviation to use is  $S_p$  (Equation 1). For repeated or matched scores, the difference between the means can be standardized by the standard deviation of the difference scores (called  $d_D$  or  $d_z$ ) as is commonly done in a  $t$  test (where  $t = d_D\sqrt{n}$ ). However,  $d_D$  cannot be directly compared with the result for  $d_p$  (Equation 2) from another study with either independent or dependent samples. Computations to attempt a conversion of  $d_D$  to  $d_p$  exist, but are not generally recommended because the assumption of homoschedasticity that is required for  $d_p$  is not required for a valid  $d_D$  (Goulet-Pelletier & Cousineau, 2018, 2019; Lakens, 2013). For an accurate comparison of the standard-



**Table 5** ■ Average 95% BCa CI width for each value of  $\delta$ ,  $\rho$  and  $n$ .

$\delta$	$\rho$	npairs							
		10	20	30	40	50	75	150	250
0	0	1.94	1.29	1.04	0.89	0.79	0.65	0.45	0.35
0.2	0	1.94	1.29	1.04	0.89	0.80	0.65	0.46	0.35
0.4	0	1.94	1.30	1.05	0.90	0.80	0.65	0.46	0.35
0.6	0	1.95	1.31	1.06	0.91	0.81	0.66	0.46	0.36
0.8	0	1.97	1.33	1.07	0.92	0.82	0.67	0.47	0.36
1	0	1.99	1.35	1.09	0.94	0.84	0.68	0.48	0.37
0	0.2	1.74	1.16	0.93	0.80	0.71	0.58	0.41	0.31
0.2	0.2	1.74	1.16	0.93	0.80	0.71	0.58	0.41	0.32
0.4	0.2	1.75	1.17	0.94	0.81	0.72	0.59	0.41	0.32
0.6	0.2	1.77	1.18	0.95	0.82	0.73	0.59	0.42	0.32
0.8	0.2	1.78	1.20	0.97	0.84	0.74	0.61	0.43	0.33
1	0.2	1.81	1.23	0.99	0.85	0.76	0.62	0.44	0.34
0	0.4	1.53	1.01	0.81	0.69	0.62	0.50	0.35	0.27
0.2	0.4	1.53	1.01	0.81	0.70	0.62	0.50	0.35	0.27
0.4	0.4	1.54	1.02	0.82	0.71	0.63	0.51	0.36	0.28
0.6	0.4	1.56	1.04	0.84	0.72	0.64	0.52	0.37	0.28
0.8	0.4	1.59	1.07	0.86	0.74	0.66	0.54	0.38	0.29
1	0.4	1.62	1.10	0.89	0.76	0.68	0.56	0.39	0.30
0	0.6	1.28	0.84	0.67	0.57	0.51	0.41	0.29	0.22
0.2	0.6	1.29	0.84	0.67	0.58	0.51	0.41	0.29	0.22
0.4	0.6	1.30	0.86	0.68	0.59	0.52	0.42	0.30	0.23
0.6	0.6	1.32	0.88	0.71	0.61	0.54	0.44	0.31	0.24
0.8	0.6	1.36	0.91	0.74	0.63	0.56	0.46	0.32	0.25
1	0.6	1.41	0.95	0.77	0.67	0.59	0.48	0.34	0.26
0	0.8	0.94	0.60	0.48	0.41	0.36	0.29	0.20	0.16
0.2	0.8	0.95	0.61	0.49	0.41	0.37	0.30	0.21	0.16
0.4	0.8	0.98	0.64	0.51	0.43	0.39	0.31	0.22	0.17
0.6	0.8	1.02	0.68	0.54	0.47	0.42	0.34	0.24	0.18
0.8	0.8	1.08	0.73	0.59	0.51	0.45	0.37	0.26	0.20
1	0.8	1.16	0.79	0.64	0.56	0.50	0.41	0.29	0.22

ized effect size of a study with independent subjects to that from a study with paired measures (repeated measures or matched subjects) one needs to compute  $d_p$  from the paired measures (reviewed in Fitts, 2022).

A CI for  $d_p$  from normally distributed and homoschedastic independent scores is most reliably created using the noncentral  $t$  method of Steiger and Fouladi (1997) as reported by Fitts (2021). However, the calculation of this interval from dependent measures is problematic because of the requirement to calculate the degrees of freedom,  $\nu$ . Various authors suggested using  $\nu = n - 1$ , as in the  $t$  test (Borenstein et al., 2009; Cumming, 2012; Cumming & Finch, 2001; Lakens, 2013), although Cumming (2012) noted that  $g_p$  was not completely unbiased. Goulet-Pelletier and Cousineau (2018) suggested  $\nu = 2(n - 1)$  as in an independent measures experiment with equal sample sizes. Fitts

(2020) demonstrated that the degrees of freedom actually vary according to the value of  $\rho$  from  $\nu = (n - 1)$  with  $\rho = 1.0$  to  $\nu = 2(n - 1)$  with  $\rho = 0$ . Cousineau (2020) rapidly supplied a calculation formula (Equation 4). Replacing  $\rho$  with an unbiased estimate of the sample  $r$  as in Equation 9 introduces random error in the calculation of either  $g_p$  or the CI for  $d_p$  and the distribution of the noncentrality parameter corresponding to  $d_p$  is no longer distributed exactly as a noncentral  $t$ . Fitts (2022) suggested an empirical adjustment to correct for bias from using  $r_{OP}$  instead of  $\rho$  in the calculation of Steiger & Fouladi intervals, and numerous approximations have also been suggested (Cousineau, 2020; Cousineau & Goulet-Pelletier, 2021; Viechtbauer, 2007).

The present simulations demonstrate that a CI with accurate coverage can be calculated using a BCa bootstrap

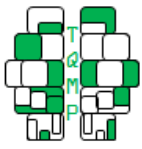


Table 6 ■ Average 99% BCa CI width for each value of  $\delta$ ,  $\rho$  and  $n$ .

$\delta$	$\rho$	npairs							
		10	20	30	40	50	75	150	250
0	0	2.74	1.74	1.38	1.18	1.05	0.85	0.60	0.46
0.2	0	2.74	1.74	1.39	1.19	1.06	0.85	0.60	0.46
0.4	0	2.75	1.75	1.40	1.19	1.06	0.86	0.60	0.47
0.6	0	2.76	1.77	1.41	1.21	1.07	0.87	0.61	0.47
0.8	0	2.78	1.79	1.43	1.22	1.09	0.88	0.62	0.48
1	0	2.80	1.82	1.45	1.25	1.11	0.90	0.63	0.49
0	0.2	2.47	1.56	1.24	1.06	0.94	0.76	0.54	0.41
0.2	0.2	2.47	1.56	1.24	1.06	0.95	0.77	0.54	0.42
0.4	0.2	2.48	1.58	1.25	1.07	0.95	0.77	0.54	0.42
0.6	0.2	2.49	1.59	1.27	1.09	0.97	0.78	0.55	0.43
0.8	0.2	2.52	1.62	1.29	1.11	0.99	0.80	0.56	0.43
1	0.2	2.54	1.65	1.32	1.13	1.01	0.82	0.58	0.45
0	0.4	2.17	1.36	1.08	0.92	0.82	0.66	0.46	0.36
0.2	0.4	2.18	1.37	1.08	0.93	0.82	0.67	0.47	0.36
0.4	0.4	2.19	1.38	1.10	0.94	0.83	0.68	0.47	0.37
0.6	0.4	2.21	1.40	1.12	0.96	0.85	0.69	0.48	0.37
0.8	0.4	2.23	1.44	1.15	0.98	0.87	0.71	0.50	0.38
1	0.4	2.27	1.47	1.18	1.01	0.90	0.73	0.52	0.40
0	0.6	1.83	1.13	0.89	0.76	0.67	0.54	0.38	0.29
0.2	0.6	1.84	1.14	0.90	0.77	0.68	0.55	0.38	0.30
0.4	0.6	1.85	1.16	0.92	0.78	0.69	0.56	0.39	0.30
0.6	0.6	1.88	1.19	0.94	0.81	0.72	0.58	0.41	0.31
0.8	0.6	1.92	1.23	0.98	0.84	0.75	0.61	0.43	0.33
1	0.6	1.97	1.28	1.03	0.89	0.79	0.64	0.45	0.35
0	0.8	1.37	0.82	0.64	0.54	0.48	0.39	0.27	0.21
0.2	0.8	1.38	0.83	0.65	0.55	0.49	0.39	0.28	0.21
0.4	0.8	1.40	0.86	0.68	0.58	0.51	0.41	0.29	0.22
0.6	0.8	1.46	0.92	0.73	0.62	0.55	0.45	0.31	0.24
0.8	0.8	1.53	0.98	0.79	0.67	0.60	0.49	0.34	0.27
1	0.8	1.63	1.07	0.86	0.74	0.66	0.54	0.38	0.29

CI, although sample sizes below 20 pairs may experience a slight depression of coverage (the worst among those tested was a coverage of .92 instead of .95 with  $n = 10$  and  $\delta = 1.0$ , Table 1). Few experiments striving for accuracy in parameter estimation (Kelley & Rausch, 2006; Maxwell et al., 2008) would plan experiments with fewer than 20 pairs.

The width of a CI indicates the precision of the estimate, with narrow CIs providing greater precision in the location of  $\delta$ . Kelley and Rausch (2006) included tables that gave sample sizes for experiments with two independent groups in order to achieve a predetermined standardized width of CI on average if one is able to estimate  $\delta$  with precision, because larger values of  $\delta$  produce wider CIs for a given sample size. For example, to achieve a 95% CI with an average width of .80 one needs 49 subjects per group if the  $\delta$  is 0.05 and 55 per group if the  $\delta$  is 1.0 (see Kelley & Rausch, 2006,

Table 2, top panel).

We can compare these sample sizes for independent groups to the present data in Table 4 for a 95% BCa CI in a paired-pooled design for several effect sizes. Because an independent groups design is uncorrelated, we first compare the sample size for  $\rho = .00$  from the present simulations with the results of Kelley and Rausch (2006). By scanning the top row of Table 4 for  $\delta = 0$  and  $\rho = 0$  we find that the sample size closest to an average width of .8 is 50 pairs of scores. By the sixth row of the table for  $\delta = 1.0$  we see that the average width is 0.84, i.e., wider than our desired 0.80, and this indicates that we would require a few additional pairs with  $\delta = 1.0$  to achieve the desired width of 0.80. Thus, the required sample size for a matched pairs experiment would require about the same sample size per group as an independent groups experiment if the correlation be-

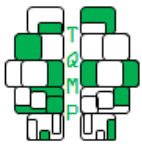


Table 7 ■ Worse coverage for each value of  $\delta$  and  $\rho$  for small sample  $g_p$  ad hoc resampling arrays with confidence coefficients of 90, 95, or 99%. Data for each  $\delta$  are averaged across all levels of  $\rho$  and vice versa. Above  $n = 50$  coverage converged with the findings of Tables 2 and 3.

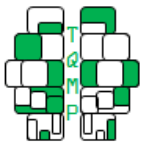
Table with 10 columns: delta, n, 90%, 95%, 99%, rho, n, 90%, 95%, 99%. Rows represent combinations of delta (0, 0.2, 0.4, 0.6, 0.8, 1) and n (10, 20, 30, 40, 50).

tween the scores is 0.0 (i.e., 49-55 from Kelley & Rausch), and the sample size for a repeated measures experiment would be half the total number of subjects of the independent groups experiment. This is the worst that could happen unless the unfortunate investigator happened to match subjects using a negatively correlated measure.

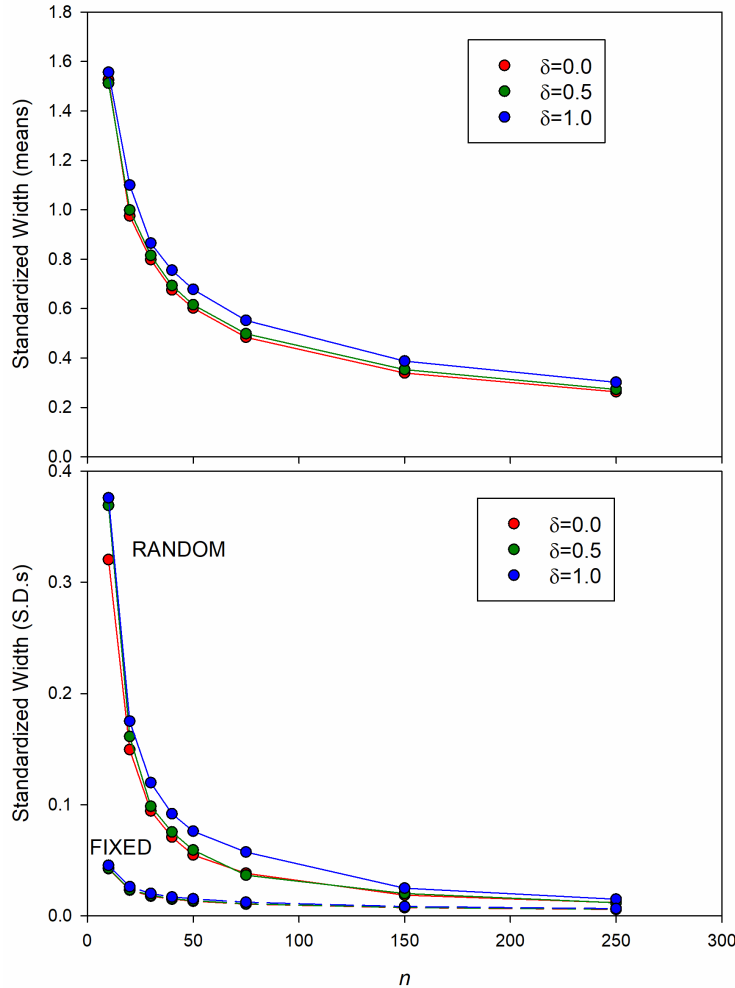
For  $\rho = .20$ , the range of expected widths of our BCa experiment with 50 pairs would be 0.71 to 0.76 for effect sizes of 0 and 1.0. For  $\rho = .40$ , the expected widths have been cut to 0.62 to 0.68 for the same sample sizes, indicating considerable improvement in efficiency by using a correlated design. With the tables of Kelley and Rausch (2006) for the independent groups design and a large effect size one can estimate the minimum width of the interval in a BCa CI if the observed  $\rho$  is only .00. If the scores are indeed positively

correlated in the paired design, which is the point of matching scores, the average width will be narrower than that of the independent groups design with either the same sample size (for matched subjects) or half the total sample size (for repeated measures). This effect is clearly illustrated in Figure 2.

Because the bootstrap method itself does not require a normal distribution of scores, one might wonder how the method works with distributions that are not normally distributed and homoschedastic. When used to generate a CI for a standardized mean difference, the problem lies partly in the interpretation of the statistic itself rather than the method. Hedges (2024) concludes, "If the data are not approximately normally distributed or if they have substantially unequal standard deviations, the relation between



**Figure 1** ■ Means and standard deviations of standardized widths of BCa CIs for 3 levels of  $\delta$  collapsed across 5 levels of  $\rho$  for different sample sizes. S.D.s were averaged for 50 reps of the FIXED data or for 50 different (RANDOM) random samples.



*d* and overlap between distributions can be very different, and interpretations of *d* that apply when the data are normal with equal variances are unreliable.” The BCa adjustment to the percentile bootstrap CI corrects for bias in asymptotic samples (Efron, 1985, 1987; Efron and Tibshirani, 1993), but the correction has limitations in terms of distribution shape and sample size (Algina et al., 2006; Chen & Peng, 2015; DasPeddada & Patwardhan, 1992; Efron & Tibshirani, 1993; Schenker, 1985; Zhou & Dinh, 2005). The present study did not include simulations with nonnormal or heteroschedastic distributions.

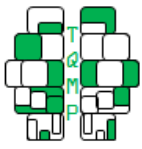
**Authors’ note**

The author is a sponsored retiree of the University of Washington, has no funding, and has no conflicts of interest.

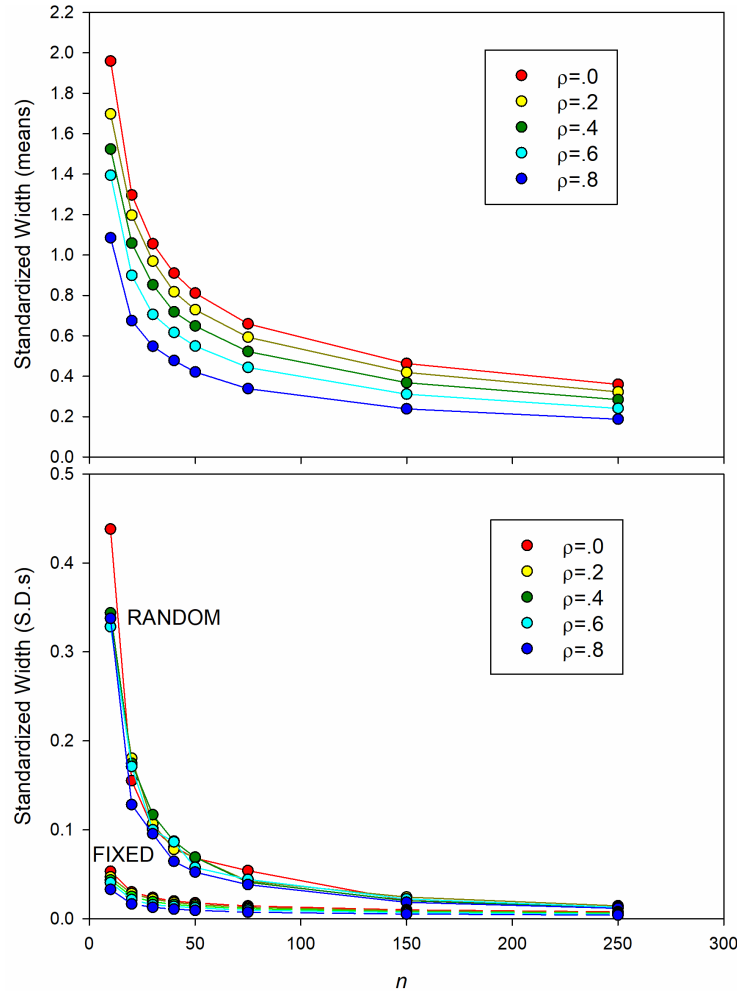
**References**

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

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. John Wiley & Sons. doi: 10.1002/9780470743386.



**Figure 2** ■ Means and standard deviations of standardized widths of BCa CIs for 5 levels of  $\rho$  collapsed across 3 levels of  $\delta$  for different sample sizes. S.D.s were averaged for 50 reps of the FIXED data or for 50 different (RANDOM) random samples.



Chen, L.-T., & Peng, C.-Y. J. (2015). The sensitivity of three methods to nonnormality and unequal variances in interval estimation of effect sizes. *Behavior Research Methods*, 47, 107–126. doi: [10.3758/s13428-014-0461-3](https://doi.org/10.3758/s13428-014-0461-3).

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum. doi: [10.4324/9780203771587](https://doi.org/10.4324/9780203771587).

Cousineau, D. (2020). Approximating the distribution of cohen's  $d_p$  in within-subject designs. *The Quantitative Methods for Psychology*, 16, 418–421. doi: [10.20982/tqmp.16.4.p418](https://doi.org/10.20982/tqmp.16.4.p418).

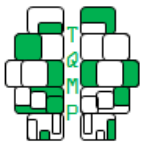
Cousineau, D., & Goulet-Pelletier, J.-C. (2021). A study of confidence intervals for cohen's  $d_p$  in within-subject designs with new proposals. *The Quantitative Methods*

*for Psychology*, 17, 51–75. doi: [10.20982/tqmp.17.1.p051](https://doi.org/10.20982/tqmp.17.1.p051).

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge. doi: [10.4324/9780203807002](https://doi.org/10.4324/9780203807002).

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](https://doi.org/10.1177/0013164401614002).

DasPeddada, S., & Patwardhan, G. (1992). Qualms about bca bootstrap confidence intervals. *Statistics and Probability Letters*, 15, 77–83. doi: [10.1016/0167-7152\(92\)90288-G](https://doi.org/10.1016/0167-7152(92)90288-G).



- Efron, B. (1985). Bootstrap confidence intervals for a class of parametric problems. *Biometrika*, 72, 45–58. doi: [10.1093/biomet/72.1.45](https://doi.org/10.1093/biomet/72.1.45).
- Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, 82, 171–185. doi: [10.2307/2289144](https://doi.org/10.2307/2289144).
- Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Chapman & Hall. doi: [10.1201/9780429246593](https://doi.org/10.1201/9780429246593).
- 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](https://doi.org/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](https://doi.org/10.3758/s13428-021-01550-4).
- Fitts, D. A. (2022). Point and interval estimates for a standardized mean difference in paired-samples designs using a pooled standard deviation. *Quantitative Methods for Psychology*, 18, 207–223. doi: [10.20982/tqmp.18.2.p207](https://doi.org/10.20982/tqmp.18.2.p207).
- 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](https://doi.org/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](https://doi.org/10.20982/tqmp.15.1.p054).
- Hedges, L. V. (1981). Distribution theory for glass’s estimator of effect size and related estimators. *Journal of Educational Statistics*, 6, 107–128. doi: [10.2307/1164588](https://doi.org/10.2307/1164588).
- Hedges, L. V. (2024). *Interpretation of the standardized mean difference effect size when distributions are not normal or homoschedastic*. Published online. doi: [10.1177/00131644241278928](https://doi.org/10.1177/00131644241278928).
- 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](https://doi.org/10.1177/0013164404264850).
- 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](https://doi.org/10.1037/1082-989X.11.4.363).
- 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](https://doi.org/10.3389/fpsyg.2013.00863).
- Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology*, 59, 537–563. doi: [10.1146/annurev.psych.59.103006.093735](https://doi.org/10.1146/annurev.psych.59.103006.093735).
- Olkin, I., & Pratt, J. W. (1958). Unbiased estimation of certain correlation coefficients. *The Annals of Mathematical Statistics*, 29, 201–211. doi: [10.1214/aoms/1177706717](https://doi.org/10.1214/aoms/1177706717).
- Schenker, N. (1985). Qualms about bootstrap confidence intervals. *Journal of the American Statistical Association*, 80, 360–361. doi: [10.2307/2287897](https://doi.org/10.2307/2287897).
- Steiger, J. H., & Fouladi, R. T. (1997). Noncentrality interval estimation and the evaluation of statistical methods. In L. L. H. S. A. Mulaik & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 221–257). Lawrence Erlbaum Associates. doi: [10.4324/9781315629049](https://doi.org/10.4324/9781315629049).
- Viechtbauer, W. (2007). Approximate confidence intervals for standardized effect sizes in the two-independent and two-dependent samples design. *Journal of Educational and Behavioral Statistics*, 32, 39–60. doi: [10.3102/1076998606298034](https://doi.org/10.3102/1076998606298034).
- Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54, 594–604. doi: [10.1037/0003-066X.54.8.594](https://doi.org/10.1037/0003-066X.54.8.594).
- Zhou, X. H., & Dinh, P. (2005). Nonparametric confidence intervals for the one- and two-sample problems. *Biostatistics*, 6, 187–200. doi: [10.1093/biostatistics/kxi002](https://doi.org/10.1093/biostatistics/kxi002).

## Citation

Fitts, D. A. (2025). Bootstrap BCa confidence intervals for a standardized mean difference in a paired samples design using a pooled standard deviation. *The Quantitative Methods for Psychology*, 21(3), 125–138. doi: [10.20982/tqmp.21.3.p125](https://doi.org/10.20982/tqmp.21.3.p125).

Copyright © 2025, Fitts. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Received: 23/05/2025 ~ Accepted: 19/10/2025