Journal of Educational and Behavioral Statistics
March 2007, Vol. 32, No. 1, pp. 39–60
DOI: 10.3102/1076998606298034

Approximate Confidence Intervals for Standardized Effect Sizes in the Two-Independent and Two-Dependent Samples Design

Wolfgang Viechtbauer
University of Maastricht

Standardized effect sizes and confidence intervals thereof are extremely useful devices for comparing results across different studies using scales with incommensurable units. However, exact confidence intervals for standardized effect sizes can usually be obtained only via iterative estimation procedures. The present article summarizes several closed-form approximations to the exact confidence interval bounds in the two-independent and two-dependent samples design. Monte Carlo simulations were conducted to determine the accuracy of the various approximations under a wide variety of conditions. All methods except one provided accurate results for moderately large sample sizes and converged to the exact confidence interval bounds as sample size increased.

Keywords: *effect size; standardized mean difference; confidence intervals; two-independent samples design; two-dependent samples design*

Introduction

There is a growing consensus that all null hypothesis significance tests should be supplemented with effect size estimates and confidence intervals (e.g., American Psychological Association, 2001; Cohen, 1994; Cumming & Finch, 2001; Hyde, 2001; Kirk, 1996; Schmidt, 1996; Thompson, 2002; Wilkinson & APA Task Force on Statistical Inference, 1999). Procedures for obtaining confidence intervals (CIs) in the raw (unstandardized) units for the two-independent and two-dependent samples design are covered in most introductory textbooks on statistics, commonly known to researchers, and implemented in statistical software packages. However, the units of the measurement scales used by researchers are often chosen arbitrarily. Reporting effect sizes and corresponding CIs in standardized units allows comparisons between measurements on scales that use incommensurable units.

For example, Marcus, Marquis, and Sakai (1997) conducted a study to investigate the effectiveness of eye movement desensitization and reprocessing (EMDR), a controversial treatment for a variety of psychological disorders

I would like to thank David Thissen and three anonymous reviewers for their valuable comments on an earlier draft of this article.

including posttraumatic stress disorder (PTSD). Patients in the EMDR group scored on average 19.76 points below patients in a standard care (SC) group following treatment, as measured by the modified PTSD Symptom Scale. Another study investigating EMDR treatment for PTSD was conducted by Carlson, Chemtob, Rusnak, Hedlund, and Muraoka (1998). Here, the mean difference between the EMDR and a control condition amounted to 3.2 points on a self-report measure devised by the authors of this study.

Those two outcomes are not directly comparable because the scales are based on a different number of items and scoring criteria. A common solution is to standardize the mean difference by the pooled standard deviation of the two groups. Doing so yields standardized mean differences of 0.76 and 1.33 points in the first and second study, respectively. In other words, the EMDR group scored 0.76 standard deviations below the SC group in the study by Marcus et al. (1997), whereas Carlson et al. (1998) found a difference of 1.33 standard deviations between the two groups. The differences in treatment efficacy between the two studies are now more apparent. Moreover, CIs can be used to indicate the precision of these effect size estimates. The corresponding 95% CIs are given by (0.26, 1.25) and (0.39, 2.25), respectively.

Obtaining exact CIs in standardized units usually requires the use of noncentral distributions and iterative estimation procedures (Cumming & Finch, 2001; Hedges & Olkin, 1985; Smithson, 2003a; Steiger & Fouladi, 1997). At the time of this writing, the methods required to find exact CIs are not covered in most textbooks, are not commonly known to researchers, and have not been implemented in most statistical analysis software. In an effort to address this problem, six articles published in the August 2001 issue of *Educational and Psychological Measurement* (Thompson, 2001) provided researchers with the necessary information to calculate effect sizes and CIs for a wide variety of experimental designs. In addition, a monograph dealing with CIs based on central and noncentral distributions was published recently (Smithson, 2003a). The article by Steiger and Fouladi (1997) also provides an excellent introduction to this topic. Finally, specialized software and scripts to be used in conjunction with standard statistical software packages are available (Cumming, 2003; Smithson, 2003b).

Researchers can either familiarize themselves with the specialized tools or rely on various approximate methods based on central distributions and closed-form expressions to calculate CIs for standardized effect sizes. Numerous such approximations have been suggested in the literature. The purpose of the present article is to examine the accuracy of such approximations in the context of the two-independent and the two-dependent samples design.

The Two-Independent Samples Design

In the two-independent samples design, participants are randomly assigned to an experimental (E) or a control (C) group. Assume that the scores within each group are sampled from normal distributions with expectations μ_E and μ_C and

common variance σ^2. The null hypothesis H_0: $\mu_E - \mu_C = 0$ (i.e., the absence of a difference in the population means) can be tested with the familiar two-independent samples *t* test.

This null hypothesis significance test provides us with a simple dichotomous decision rule, namely, whether to reject H_0 or not, but neither informs us about the direction, magnitude, or precision of the measured effect. Clearly, $\bar{X}_E - \bar{X}_C$ provides an unbiased estimate of $\mu_E - \mu_C$, where $\bar{X}_E$ and $\bar{X}_C$ denote the sample means of the n_E and n_C scores in the two groups. The precision of this estimate can be indicated by a $(1 - \alpha) \times 100\%$ CI for $\mu_E - \mu_C$, given by

$$\bar{X}_E - \bar{X}_C \pm t_{(n_E+n_C-2),1-\alpha/2} s_p \sqrt{\left(\frac{1}{n_E} + \frac{1}{n_C}\right)}, \tag{1}$$

where $t_{m,1-\alpha/2}$ denotes the $100 \times (1 - \alpha/2)^{th}$ quantile of a central *t* distribution with *m* degrees of freedom and s_p^2 the pooled variance of the two groups.

In the ideal case where one is investigating a particular outcome variable whose raw units can be compared across related experiments, the unstandardized mean difference $\mu_E - \mu_C$ represents a reasonable choice for the population effect size. However, as discussed earlier, the measurement units are often chosen arbitrarily. Therefore, working with standardized units can be more informative as this allows comparisons of parameter estimates across scales using different units. The population effect size is then defined as

$$\delta_2 = \frac{\mu_E - \mu_C}{\sigma}, \tag{2}$$

which reflects the difference in the population means in standard deviation units. An estimate of δ_2 is given by

$$d_2 = \frac{\bar{X}_E - \bar{X}_C}{s_p}. \tag{3}$$

However, d_2 is a positively biased estimator of δ_2. The bias of d_2 was first demonstrated by Hedges (1981), who also derived the unbiased estimator

$$g_2 = c(m)\left(\frac{\bar{X}_E - \bar{X}_C}{s_p}\right), \tag{4}$$

where

$$c(m) = \frac{\Gamma\left(\frac{m}{2}\right)}{\sqrt{\frac{m}{2}}\,\Gamma\left(\frac{m-1}{2}\right)} \approx 1 - \frac{3}{4m-1}. \tag{5}$$

Based on Hedges (1981, 1982, 1983), we note the following set of results. The exact variances of d_2 and g_2 are given by Equations 21 and 22 in Table 1. However, $\sigma^2_{d_2}$ and $\sigma^2_{g_2}$ depend on the unknown value δ_2, which in practice is replaced by either d_2 or g_2, leading to estimates $\hat{\sigma}^{2\,(B)}_{d_2}$ and $\hat{\sigma}^{2\,(B)}_{g_2}$ (Equations 23 and 24). This introduces a certain amount of bias into the estimated sampling variances because $\mathrm{E}[d_2^2] \neq \mathrm{E}[g_2^2] \neq \delta_2^2$. Unbiased estimates of $\sigma^2_{d_2}$ and $\sigma^2_{g_2}$, denoted by $\hat{\sigma}^{2(U)}_{d_2}$ and $\hat{\sigma}^{2\,(U)}_{g_2}$, are given by Equations 25 and 26. Finally, d_2 and g_2 are asymptotically normal with mean δ_2 and variance $1/\tilde{n} + \delta_2^2/(2m)$, where $\tilde{n} = n_E n_C/(n_E + n_C)$. Replacing the unknown value of δ_2 by either sample estimate leads to $\hat{\sigma}^{2\,(L1)}_{d_2}$ and $\hat{\sigma}^{2\,(L1)}_{g_2}$, the large sample variance estimators (Equations 28 and 29). However, in the literature, one usually finds m replaced with the total sample size $N = n_E + n_C$ in Equation 27. This leads to the large sample estimators $\hat{\sigma}^{2\,(L2)}_{d_2}$ and $\hat{\sigma}^{2\,(L2)}_{g_2}$ (Equations 30 and 31).

The Two-Dependent Samples Design

Researchers often choose to measure the same set of n participants on two different occasions, such as before and after receiving some treatment. Because the same group of participants is measured twice, the two sets of scores are no longer independent. Assume that the scores X_1 and X_2 obtained at Time 1 and Time 2 are sampled from normal distributions with expected values μ_1 and μ_2 and common variance σ^2. Now define the random variable $D = X_2 - X_1$. It follows that D is normally distributed with expected value $\mu_D = \mu_2 - \mu_1$ and variance $\sigma^2_D = 2\sigma^2(1 - \rho)$, where ρ is the correlation between the scores at Time 1 and Time 2.

The null hypothesis H_0: $\mu_2 - \mu_1 = 0$ (i.e., H_0: $\mu_D = 0$) can be tested by carrying out a one-sample t test on the D scores. The value of $\mu_D = \mu_2 - \mu_1$ is easily estimated with $\bar{D} = \bar{X}_2 - \bar{X}_1$, with a $(1 - \alpha) \times 100\%$ CI given by

$$\bar{D} \pm t_{(n-1),1-\alpha/2} \frac{s_D}{\sqrt{n}}, \tag{6}$$

where s^2_D is the observed variance in the D scores.

Again, we would like to obtain a standardized point estimate. Two different standardized parameters have been suggested in the literature (Becker, 1988; Gibbons, Hedeker, & Davis, 1993; Morris, 2000; Morris & DeShon, 2002), namely,

$$\delta_D = \frac{\mu_D}{\sigma_D} = \frac{\mu_2 - \mu_1}{\sigma\sqrt{2(1 - \rho)}}, \tag{7}$$

based on the standard deviation in the D scores, and

$$\delta_{D2} = \frac{\mu_D}{\sigma} = \frac{\mu_2 - \mu_1}{\sigma}, \tag{8}$$

TABLE 1
Variances and Variance Estimators in the Two-Independent Samples Case (δ_2) and the Two-Dependent Samples Case (δ_D)

	Note	Equation
$\sigma_d^2 = \frac{m[1+\tilde{n}\delta^2]}{(m-2)\tilde{n}} - \frac{\delta^2}{[c(m)]^2}$	Exact variance of d	(21)
$\sigma_g^2 = \frac{[c(m)]^2 m[1+\tilde{n}\delta^2]}{(m-2)\tilde{n}} - \delta^2$	Exact variance of g	(22)
$\hat{\sigma}_d^{2\,(B)} = \frac{m[1+\tilde{n}d^2]}{(m-2)\tilde{n}} - \frac{d^2}{[c(m)]^2}$	Biased Estimate of σ_d^2	(23)
$\hat{\sigma}_g^{2\,(B)} = \frac{[c(m)]^2 m[1+\tilde{n}g^2]}{(m-2)\tilde{n}} - g^2$	Biased estimate of σ_g^2	(24)
$\hat{\sigma}_d^{2\,(U)} = \frac{1}{\tilde{n}[c(m)]^2} + \left(1 - \frac{(m-2)}{m[c(m)]^2}\right) d^2$	Unbiased estimate of σ_d^2	(25)
$\hat{\sigma}_g^{2\,(U)} = \frac{1}{\tilde{n}} + \left(1 - \frac{(m-2)}{m[c(m)]^2}\right) g^2$	Unbiased estimate of σ_g^2	(26)
$\sigma_{d/g}^{2\,(\infty)} = \frac{1}{\tilde{n}} + \frac{\delta^2}{2m}$	Large sample variance	(27)
$\hat{\sigma}_d^{2\,(L1)} = \frac{1}{\tilde{n}} + \frac{d^2}{2m}$	Estimate of $\sigma_{d/g}^{2\,(\infty)}$	(28)
$\hat{\sigma}_g^{2\,(L1)} = \frac{1}{\tilde{n}} + \frac{g^2}{2m}$	Estimate of $\sigma_{d/g}^{2\,(\infty)}$	(29)
$\hat{\sigma}_d^{2\,(L2)} = \frac{1}{\tilde{n}} + \frac{d^2}{2N}$	Estimate of $\sigma_{d/g}^{2\,(\infty)}$	(30)
$\hat{\sigma}_g^{2\,(L2)} = \frac{1}{\tilde{n}} + \frac{g^2}{2N}$	Estimate of $\sigma_{d/g}^{2\,(\infty)}$	(31)

Note: $c(m)$ is defined in Equation 5. In the two-independent samples case, $d = d_2$, $g = g_2$, $\delta = \delta_2$, $\tilde{n} = n_E n_C/(n_E + n_C)$, $m = n_E + n_C - 2$, and $N = n_E + n_C$. In the two-dependent samples case with parameter δ_D, $d = d_D$, $g = g_D$, $\delta = \delta_D$, $\tilde{n} = n$, $m = n - 1$, and $N = n$.

which is of the same form as δ_2 defined in Equation 2 for the two-independent samples design. When $\rho = .5$, then $\delta_D = \delta_{D2}$. However, in many cases, one would expect the correlation between the scores at the two occasions to be greater than .5, which implies $\delta_D > \delta_{D2}$ (see Ray & Shadish, 1996, for some empirical evidence relevant to this issue). Simply assuming that $\rho = .5$ is not recommended, and estimates of δ_D are therefore not directly comparable to estimates of δ_{D2}.

If δ_D is chosen as the effect size parameter of interest, then biased and unbiased estimates are given by

$$d_D = \frac{\bar{D}}{s_D} \tag{9}$$

and

$$g_D = c(m)\left(\frac{\bar{D}}{s_D}\right), \tag{10}$$

respectively, where $m = n - 1$. The exact variances and variance estimates for these effect sizes are given in Table 1 with $\tilde{n} = n$ and $N = n$ (Becker, 1988; Gibbons et al., 1993; Morris, 2000; Morris & DeShon, 2002).

Working with δ_{D2} is usually preferable because it is directly comparable to δ_2 from two-independent samples designs. However, estimating δ_{D2} from dependent samples data poses some additional difficulties. Naturally, one might consider estimators of the form given by d_2 and g_2 (Equations 3 and 4) to be appropriate here. The problem with this approach is that their exact distributions, expected values, and variances are unknown. Instead, the estimators

$$d_{D2} = \frac{\bar{D}}{s_1} \tag{11}$$

and

$$g_{D2} = c(m)\left(\frac{\bar{D}}{s_1}\right) \tag{12}$$

have been suggested in the literature and are biased and unbiased estimators of δ_{D2}, respectively, where $m = n - 1$ and s_1 is the pretreatment standard deviation (Becker, 1988; Morris, 2000; Morris & DeShon, 2002). The exact variance of d_{D2} and g_{D2} and the large sample approximations are given in Table 2 (Becker, 1988; Morris & DeShon, 2002). In addition, biased estimates of the exact sampling variance are obtained by replacing the parameter δ_{D2} by either d_{D2} or g_{D2} and ρ by the sample correlation r. Unbiased estimates of the exact sampling variance can also be derived and are given in Table 2 (Equations 36 and 37).

Note that one must use an unbiased estimate of ρ in Equations 36 and 37 to obtain unbiased estimates of the sampling variances of d_{D2} and g_{D2}. An unbiased estimate of ρ, derived by Olkin and Pratt (1958), is given by

$$r^{u} = r\, F\left(\frac{1}{2}, \frac{1}{2}, \frac{n-2}{2}, 1 - r^2\right), \tag{13}$$

where

$$F(\alpha, \beta, \gamma, x) = \sum_{j=0}^{\infty} \frac{\Gamma(\alpha + j)\Gamma(\beta + j)\Gamma(\gamma)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\gamma + j)} \frac{x^j}{j!} \tag{14}$$

denotes the hypergeometric function. Olkin and Pratt also suggested

$$r^{u} \approx r + \frac{r(1 - r^2)}{2(n - 4)} \tag{15}$$

as a simple yet accurate approximation to the unbiased estimator of ρ.

TABLE 2
Variances and Variance Estimators in the Two-Dependent Samples Case (δ_{D2})

	Note	Equation
$\sigma_d^2 = \frac{m[2(1-\rho)+n\delta^2]}{(m-2)n} - \frac{\delta^2}{[c(m)]^2}$	Exact variance of d	(32)
$\sigma_g^2 = \frac{[c(m)]^2 m[2(1-\rho)+n\delta^2]}{(m-2)n} - \delta^2$	Exact variance of g	(33)
$\hat{\sigma}_d^{2\,(B)} = \frac{m[2(1-r)+nd^2]}{(m-2)n} - \frac{d^2}{[c(m)]^2}$	Biased estimate of σ_d^2	(34)
$\hat{\sigma}_g^{2\,(B)} = \frac{[c(m)]^2 m[2(1-r)+ng^2]}{(m-2)n} - g^2$	Biased estimate of σ_g^2	(35)
$\hat{\sigma}_d^{2\,(U)} = \frac{2(1-r^\mu)}{n[c(m)]^2} + \left(1 - \frac{(m-2)}{m[c(m)]^2}\right)d^2$	Unbiased estimate of σ_d^2	(36)
$\hat{\sigma}_g^{2\,(U)} = \frac{2(1-r^\mu)}{n} + \left(1 - \frac{(m-2)}{m[c(m)]^2}\right)g^2$	Unbiased estimate of σ_g^2	(37)
$\sigma_{d/g}^{2\,(\infty)} = \frac{2(1-\rho)}{n} + \frac{\delta^2}{2m}$	Large sample variance	(38)
$\hat{\sigma}_d^{2\,(L1)} = \frac{2(1-r)}{n} + \frac{d^2}{2m}$	Estimate of $\sigma_{d/g}^{2\,(\infty)}$	(39)
$\hat{\sigma}_g^{2\,(L1)} = \frac{2(1-r)}{n} + \frac{g^2}{2m}$	Estimate of $\sigma_{d/g}^{2\,(\infty)}$	(40)
$\hat{\sigma}_d^{2\,(L2)} = \frac{2(1-r)}{n} + \frac{d^2}{2n}$	Estimate of $\sigma_{d/g}^{2\,(\infty)}$	(41)
$\hat{\sigma}_g^{2\,(L2)} = \frac{2(1-r)}{n} + \frac{g^2}{2n}$	Estimate of $\sigma_{d/g}^{2\,(\infty)}$	(42)

Note: $c(m)$ is defined in Equation 5. In the two-dependent samples case with parameter δ_{D2}, $d = d_{D2}$, $g = g_{D2}$, $\delta = \delta_{D2}$, and $m = n - 1$. r is the sample correlation coefficient, and r^μ is the unbiased estimate of the population correlation ρ as defined in Equation 13.

Confidence Intervals for δ

In the following sections, δ refers to one of the three different effect size parameters discussed previously, namely, δ_2, δ_D, or δ_{D2}. We will now consider methods for finding a CI for δ. Finding the exact CI bounds for δ is problematic because the shape of the distribution of d and g depends directly on the parameter for which the interval is being constructed. Therefore, the CI cannot be given as a closed-form expression.

Iterative methods to find the exact CI for the two-independent and two-dependent samples design with parameter δ_D have been discussed in the literature (Cumming & Finch, 2001; Hedges & Olkin, 1985; Smithson, 2003a; Steiger & Fouladi, 1997). Exact CI bounds for the two-dependent samples design with parameter δ_{D2} can also be obtained when ρ is known by multiplying the bounds for δ_D by $\sqrt{2(1-\rho)}$. However, in practice, ρ must be estimated from the data. The

additional variability in estimates of ρ must be considered when using an iteration procedure to find the exact CI for δ_{D2}. No such method has been developed yet.

We will now consider several approximations to the exact CI bounds. Let q be equal to the $100 \times (1 - \alpha/2)^{th}$ quantile of either the standard normal or the central t distribution with m degrees of freedom. Approximate $(1 - \alpha) \times 100\%$ CIs for δ are given by methods B, U, L1, and L2 as shown in Table 3. Two different sets of approximations are given, depending on whether one uses the biased or the unbiased estimate of the corresponding population parameter. Also, one can use either the normal distribution or the t distribution to construct such approximate CIs. Use of the normal distribution for obtaining CIs can be justified based on the fact that d and g are asymptotically normal with expectation δ and variances given by Equations 27 or 38. On the other hand, when $\delta = 0$, then $\sqrt{\tilde{n}}d$ and $c(m)^{-1}\sqrt{\tilde{n}}g$ are distributed central t with m degrees of freedom, which suggests use of the t distribution for small δ.

Another approach to obtain approximate CI bounds for δ is to first use a variance stabilizing transformation. The variance stabilizing transformation for the standardized mean difference, as suggested by Hedges and Olkin (1985), is here generalized to include the two-dependent samples design. Based on the delta method, the random variable

$$z_g = h(g) = \sqrt{2}\sinh^{-1}\frac{g}{a} = \sqrt{2}\log\left(\frac{g}{a} + \sqrt{\frac{g^2}{a^2} + 1}\right) \tag{16}$$

can be shown to be asymptotically normal with expectation $h(\delta)$ and variance $1/N$, where $a = \sqrt{4 + 2(n_E/n_C) + 2(n_C/n_E)}$ in the two-independent samples case, $a = \sqrt{2}$ in the two-dependent samples case with parameter δ_D, and $a = \sqrt{4(1 - r)}$ in the two-dependent samples case with parameter δ_{D2}. Therefore, one calculates the lower and upper bounds of a CI using the distribution of z_g with

$$z_g \pm q \times \sqrt{1/N} \tag{17}$$

and then transforms these bounds back to the original metric with

$$g = h^{-1}(z_g) = a\sinh\left(\frac{z_g}{\sqrt{2}}\right) = a\left(\frac{\exp[z_g/\sqrt{2}] - \exp[z_g/\sqrt{2}]}{2}\right), \tag{18}$$

the inverse function of $h(g)$. This method (denoted by the letter H) could also be applied using the biased estimate d in place of g.

Finally, Fidler and Thompson (2001) suggested that a CI for δ_2 in the two-independent samples design could be approximated by dividing each X_E and X_C score by s_p and obtaining a CI for the raw mean difference as described by Equation 1 using the transformed data. It is easy to show that this method is

TABLE 3
Methods to Obtain Approximate Confidence Intervals for δ

Method	
gB	$g \pm q \times \hat{\sigma}_g^{(B)}$
dB	$d \pm q \times \hat{\sigma}_d^{(B)}$
gU	$g \pm q \times \hat{\sigma}_g^{(U)}$
dU	$d \pm q \times \hat{\sigma}_d^{(U)}$
gL1	$g \pm q \times \hat{\sigma}_g^{(L1)}$
dL1	$d \pm q \times \hat{\sigma}_d^{(L1)}$
gL2	$g \pm q \times \hat{\sigma}_g^{(L2)}$
dL2	$d \pm q \times \hat{\sigma}_d^{(L2)}$
gH	$h^{-1}(z_g \pm q \times \sqrt{1/N})$
dH	$h^{-1}(z_d \pm q \times \sqrt{1/N})$
	Equation 1 divided by s_p (two-independent samples case)
F	Equation 6 divided by s_D (two-dependent samples case with δ_D)
	Equation 6 divided by $s_D/\sqrt{2(1-r)}$ (two-dependent samples case with δ_{D2})

Note: See text for details. q is the $100 \times (1-\alpha/2)^{th}$ quantile of either the standard normal distribution or the central t distribution with m degrees of freedom.

identical to dividing the CI bounds for $\mu_E - \mu_C$ obtained from Equation 1 by s_p. A similar approach was also discussed by Bird (2002). However, after dividing Equation 1 by s_p, we obtain $d_2 \pm t_{(n_E+n_C-2),1-\alpha/2}\sqrt{1/\tilde{n}}$. This in turn reveals that the Fidler and Thompson approach is essentially the same as using the biased estimate d_2 and constructing a CI based on the t distribution and the large sample variance $\sigma^{2\,(\infty)}_{d/g}$ (Equation 27) where δ_2 is assumed to be zero. The same principle could be extended to the two-dependent samples design where a CI for δ_D is sought by dividing the CI bounds for μ_D obtained from Equation 6 by s_D. Finally, dividing the CI bounds for μ_D obtained from Equation 6 by $s_D/\sqrt{2(1-r)}$ would provide an approximate CI for δ_{D2}. This method of finding approximate CIs for δ will be denoted by the letter F.

In total, this yields 21 different approximations (see again Table 3). Specifically, methods B, U, L1, L2, and H can either be based on d or g and can either employ critical values from the normal or the t distribution, whereas method F is by default always based on d and the t distribution. The letter z or t will be appended to the name of the approximation to indicate which distribution was used. For example, method B based on g and the normal distribution will be denoted by gBz.

Examples

Three examples will illustrate the various approximations discussed in the previous section. First, consider the two-independent samples case. Assume that (25, 19, 21, 14, 16, 23, 24, 24, 22, 22) and (28, 26, 27, 19, 23, 29, 25, 31, 32, 30) represent test scores from students randomly assigned to a control and an experimental group, respectively, in an experiment investigating the impact of a new teaching technique on students' performance. The mean difference in test scores between the two groups is 6.0 points with a 95% CI given by (2.44, 9.56). To obtain the standardized estimate g_2, we use Equation 4 with $m = 18$ and $c(18) = .96$. In standardized units, it turns out that the students in the experimental group scored $g_2 = 1.52$ standard deviations above the control group. Using iterative estimation procedures, we find that an exact 95% CI for δ_2 is given by (.55, 2.58).

We now compare the exact CI with the approximate CI bounds obtained by the methods discussed earlier (for conciseness, the examples in this section will be based on the normal distribution and g only). For method gBz, we first calculate $\hat{\sigma}_g^{2\,(B)}$ (Equation 24), which is equal to .28. Next, we find that $1.52 \pm 1.96 \times \sqrt{.28}$ provides approximate CI bounds (.48, 2.55). For method gUz, we find that $\hat{\sigma}_g^{2(U)}$ is equal to .27 and $1.52 \pm 1.96 \times \sqrt{.27}$ then provides the approximate bounds (.50, 2.54). Continuing with methods gL1z and gL2z, we find the CI bounds to be equal to (.51, 2.52) and (.52, 2.51), respectively. For the variance stabilizing transformation (method gHz), we use Equation 17 with $a = \sqrt{8}$ and obtain $z_g = .73$. We then calculate the CI in the transformed units with Equation 17, yielding $.73 \pm 1.96 \times \sqrt{1/20} = (.29, 1.16)$. Transforming these bounds back into the original units using Equation 18 results in (.58, 2.60). Finally, for method F, we divide the bounds of the CI based on the raw units by $s_p = 3.79$, yielding (.65, 2.52).

Next, consider the dependent samples case where interest is focused on δ_D. Assume that the scores given earlier were obtained from 10 participants tested before and after receiving some treatment. The change scores $D = X_2 - X_1$ are equal to (3, 7, 6, 5, 7, 6, 1, 7, 10, 8). The unstandardized mean difference is estimated by $\bar{D}$, which is 6.0 as in the two-independent sample case. A 95% CI for μ_D, obtained with Equation 6, yields the bounds (4.18, 7.82). Because the scores are highly correlated ($r = .78$), this CI is narrower than the one obtained from the same data when treating the two sets of scores as coming from two independent samples.

Using Equation 10 with $m = 9$ and $c(9) = 0.91$, we find $g_D = 2.16$. An exact CI for δ_D is given by (1.11, 3.59). Approximate 95% CIs are obtained in the same manner as before except that $\tilde{n} = n = 10$, $N = n = 10$, and $a = \sqrt{2}$. The approximate bounds are equal to (.84, 3.48) using method gBz, (.89, 3.43) using method gUz, (.99, 3.33) using method gL1z, (1.03, 3.29) using method gL2z, (1.20, 3.54) using method gHz, and (1.65, 3.08) using method F.

As a final example, consider the case where an estimate and CI for δ_{D2} is sought. Using Equation 12 with $m = 9$ and $c(9) = 0.91$, we find $g_{D2} = 1.51$. An exact CI for δ_{D2} is obtained by multiplying the bounds for δ_D, namely,

$(1.11, 3.59)$, by $\sqrt{2(1-\rho)}$. The scores given earlier were drawn from a population where $\rho = .80$, and therefore, the exact CI for δ_{D2} is given by $(.70, 2.27)$.

Method gBz in this case is based on Equation 35, which yields an approximate CI of $(.60, 2.43)$. Continuing through the list of variance estimates in Table 2 and approximation methods in Table 3, we obtain the bounds $(.64, 2.38)$ with method gUz, $(.70, 2.33)$ with method gL1z, and $(.73, 2.30)$ with method gL2z. Method gHz, based on the variance stabilizing transformation, yields the approximate bounds $(.86, 2.47)$. And finally, an approximate CI is given by $(1.11, 2.04)$ when using method F.

The examples illustrate that the methods can differ widely in terms of how well they approximate the exact CI bounds. Hedges (1982) studied method gL2z in the two-independent samples design (using sample sizes in the range $10 \leq n_E = n_C \leq 100$ and values of δ_2 between 0.25 and 1.50) and found the approximation to be quite accurate. Morris (2000) studied method dUz in the two-dependent samples design with parameter δ_{D2}. However, δ_{D2} and ρ were treated as known in the simulations, which bears little relevance to practice, where only sample estimates of δ_{D2} and ρ are available. It is still unknown how well CIs based on the remaining methods approximate the exact CI bounds and in particular, whether one method should be preferred over the others. Results for unequal and very small sample sizes and values of $|\delta|$ above 1.5 are also still warranted. Finally, the two-dependent samples design with parameter δ_D has not been studied at all so far. Therefore, Monte Carlo simulations were conducted to compare the accuracy of the various approximations.

Method

Three sets of simulations were conducted. The first set of simulations corresponds to the two-independent samples case. Values of δ_2 between -2 and *2* in steps of .25, seven different CI widths ($1 - \alpha = .50$, .60, .70, .80, .90, .95, and .99.), and various sample size configurations were used: (a) equal sample sizes of $n_E = n_C = 4$, 8, 16, 32, and 64 participants per group; (b) unequal sample sizes with $(n_E, n_C) = (2, 6)$, (4, 12), (8, 24), (16, 48), and (32, 96), each corresponding to a 25/75% split of participants; and (c) unequal sample sizes with $(n_E, n_C) = (2,$ 14), (4, 28), (8, 56), and (16, 112), each corresponding to a 12.5/87.5% split of participants. Therefore, there were a total of 17 $\delta_2 \times 14$ sample size $\times 7$ CI width $= 1,666$ conditions. On each of the 100,000 iterations for a particular condition, a d_2 value was directly simulated from $Z/\sqrt{X/m}$, where Z is a random normal variable with distribution $N(\delta_2, 1/n_E + 1/n_C)$ and X is a random chi-square variable with $m = n_E + n_C - 2$ degrees of freedom. The exact $(1 - \alpha) \times 100\%$ CI bounds for δ_2 were then determined using iterative methods. Next, approximate CI bounds were obtained with each of the 21 methods discussed earlier.

The second set of simulations corresponds to the two-dependent samples case with parameter δ_D. Because values of δ_D tend to be larger than δ_2 values, values

of δ_D between -4 and 4 in steps of .50 were used. Five sample size conditions ($n = 8$, 16, 32, 64, and 128) were simulated. In total, this yields a total of $17\delta_D \times 5$ sample size $\times 7$ CI width $= 595$ conditions. Again, 100,000 values of d_D were directly simulated for each condition from $Z/\sqrt{X/m}$, where Z is a random normal variable with distribution $N(\delta_D, 1/n)$ and X is a random chi-square variable with $m = n - 1$ degrees of freedom. Exact and approximate CI bounds were then obtained for δ_D.

The third set of simulations corresponds to the two-dependent samples case with parameter δ_{D2}. For each trial, two vectors of random standard normal data were generated for various values of n with population correlation coefficient ρ using the Cholesky decomposition. Next, a constant was added to one of the two sets such that all values of δ_{D2} between -2 and *2* in steps of .25 were represented in the simulations. Sample sizes of n equal to 8, 16, 32, 64, and 128 and values of ρ equal to 0, .1, .3, .5, .7, and .9 were included in the simulations. This yields a total of $17\delta_{D2} \times 5$ sample size $\times 6\rho \times 7$ CI width $= 3,570$ conditions. After generating the data, d_{D2} and g_{D2} were calculated and the exact CI bounds for δ_{D2} were determined using iterative methods (assuming known ρ). Next, approximate CI bounds for δ_{D2} were obtained with each of the approximation methods. The third set of simulations was based on 10,000 trials per condition because simulating the raw data required substantially more computing time.

The accuracy of the various approximations was assessed with two measures: the empirical coverage probability and the ratio of the length of the approximate CI compared to the exact interval. Specifically, the empirical coverage probability was estimated with

$$\hat{P} = \frac{1}{R}\sum_{i=1}^{R} I_{(\hat{C}_{L_i}, \hat{C}_{U_i})}[\delta], \tag{19}$$

where $\hat{C}_{L_i}$ and $\hat{C}_{U_i}$ denote the lower and upper CI bounds on the i^{th} iteration, $I_{(\hat{C}_{L_i}, \hat{C}_{U_i})}[\delta] = 1$ if $\delta \in (\hat{C}_{L_i}, \hat{C}_{U_i})$ and 0 otherwise, and $R = 100,000$ for the first two sets of simulations, and $R = 10,000$ for the third set. The maximum standard error of the $\hat{P}$ values is .002 for $R = 100,000$ and .005 for $R = 10,000$.

The empirical coverage probability indicates whether the approximation captures the parameter as often as it should. On the other hand, the ratio of the length of the approximate CI to the true CI indicates to what degree the width of the true interval was over- or underestimated. The average width ratio for a particular method was estimated with

$$\hat{W} = \frac{1}{R}\sum_{i=1}^{R} \frac{(\hat{C}_{U_i} - \hat{C}_{L_i})}{(C_{U_i} - C_{L_i})}, \tag{20}$$

where C_{L_i} and C_{U_i} are the lower and upper bounds of the exact CI.

The error in the empirical coverage probability of an approximation is given by $\hat{P} - (1 - \alpha)$. The error in the width ratio is given by $\hat{W} - 1$. Because of the large number of approximations considered in the present article, the methods were first examined in terms of their maximum error in coverage probability and width ratio. These values represent the worst-case results and allow us to rule out methods that can yield grossly inaccurate approximations (the full set of results can be obtained by contacting the author).

Results

Two-Independent Samples Design

The accuracy of methods B, U, L1, L2, and H depended only on the total sample size ($N = n_E + n_C$) and not on the proportion of scores falling into each group. Also, all these methods converged to the nominal coverage probabilities and interval widths as N increased, regardless of whether the method was based on d_2, g_2, the normal, or the t distribution critical values. In fact, the maximum value of $(\hat{P} - (1 - \alpha)) \times 100\%$ over all conditions and methods (excluding method F) was never larger than 6.5% for $N = 16$, 3.0% for $N = 32$, 1.5% for $N = 64$, and 1.0% for $N = 128$. In terms of width ratios, convergence was somewhat slower. The maximum value of $(\hat{W} - 1) \times 100\%$ was 28.6%, 11.9%, 5.5%, and 2.6% for a total sample size of 16, 32, 64, and 128 participants, respectively.

The convergence of the empirical coverage probabilities of methods B, U, L1, L2, and H to the nominal $(1 - \alpha)$ values and CI widths was to be expected due to three reasons. First, the distributions of d and g are asymptotically normal with expectation δ. Second, these methods are based on consistent variance estimates. And finally, for large sample sizes, the normal and t distributions converge, yielding very similar quantiles.

However, for small sample sizes, some approximations were noticeably more accurate than others. Table 4 indicates the maximum value of $(\hat{P} - (1 - \alpha)) \times 100\%$ and $(\hat{W} - 1) \times 100\%$ for each method and each CI width over all values of δ_2 for $N = 8$ in the equal sample size condition. Negative signs indicate underestimation of coverage probabilities and widths, whereas positive signs indicate overestimation. The table shows that the approximate bounds based on the normal distribution generally yielded more accurate results than the corresponding bounds based on the t distribution, especially with respect to width ratios and 90% or wider CIs.

Table 4 also indicates that methods gUz, gL1z, dL1z, dL2z, gHz, and dHz provided the most accurate results when considering empirical coverage probabilities and width ratios simultaneously. To determine whether one of these methods is the most accurate, one should examine individual values of $\hat{P}$ and $\hat{W}$ instead of focusing on the maximum errors as done in Table 4. Accordingly, the empirical coverage probabilities and width ratios of these methods are plotted by δ_2 values for the 95% CI condition in Figure 1 when $N = 8$. In addition, the

TABLE 4
Maximum Error (in %) in Empirical Coverage Probabilities and Width Ratios Over All Values of δ_2 for N = *8 in the Two-Independent Samples Case (Equal Group Sizes)*

Confidence Interval	Coverage Probability					Width Ratio				
	50%	70%	90%	95%	99%	50%	70%	90%	95%	99%
gBz	7.3	7.9	5.5	3.6	1.0	8.6	8.6	8.7	8.8	8.9
gBt	10.2	12.1	8.5	4.9	1.0	15.5	18.8	28.4	35.8	56.7
dBz	11.6	12.8	7.6	4.4	1.0	33.6	33.6	33.7	33.8	34.0
dBt	14.7	16.8	9.3	5.0	1.0	42.1	46.2	58.0	67.1	92.9
gUz	4.5	4.9	3.8	2.7	0.9	2.0	2.1	2.2	2.2	2.4
gUt	7.3	9.2	7.6	4.8	1.0	8.6	11.7	20.7	27.6	47.3
dUz	5.5	6.5	5.0	3.2	0.9	17.5	17.5	17.6	17.7	17.8
dUt	8.4	10.8	8.2	4.8	1.0	25.0	28.6	38.9	46.9	69.6
gL1z	4.1	4.0	2.6	1.6	0.6	−4.6	−4.6	−4.5	−4.4	−4.3
gL1t	6.9	8.2	6.5	4.3	1.0	5.4	8.4	17.1	23.7	42.7
dL1z	−1.9	−2.1	−1.2	−0.5	0.4	0.9	0.9	1.0	1.1	1.2
dL1t	1.4	3.5	5.3	4.0	1.0	7.3	10.4	19.3	26.2	45.6
gL2z	3.9	3.5	1.7	0.9	0.2	−9.1	−9.1	−9.0	−9.0	−8.9
gL2t	6.7	7.6	5.7	3.9	1.0	4.5	7.5	16.0	22.6	41.4
dL2z	−3.8	−3.9	−2.5	−1.7	−0.5	−4.7	−4.7	−4.6	−4.6	−4.4
dL2t	−1.2	1.8	3.4	3.1	1.0	5.4	8.4	17.1	23.7	42.6
gHz	3.5	2.5	−1.7	−1.4	−1.1	−8.7	−8.1	−6.4	−5.3	5.2
gHt	6.2	6.4	4.2	2.6	0.8	5.1	8.9	20.7	30.4	62.5
dHz	−4.0	−5.0	−4.9	−4.5	−2.9	−4.3	−3.6	−1.9	3.1	6.1
dHt	−1.5	−0.7	1.1	0.9	0.5	6.0	9.9	21.7	31.6	63.9
F	−9.4	−10.2	−7.5	−6.2	−3.4	−20.4	−18.0	12.8	19.2	37.4
F[a]	−8.4	−10.1	−8.2	−6.0	−2.6	−18.5	−18.4	−18.1	−17.9	−17.5
F[b]	−6.6	−8.1	−6.0	−4.5	−1.9	−14.8	−14.7	−14.4	−14.2	−13.7
F[c]	−4.2	−4.7	−3.7	−2.7	−1.0	−9.4	−9.3	−9.1	−8.8	−8.3

a. $(n_E, n_C) = (64, 64)$.
b. $(n_E, n_C) = (32, 96)$.
c. $(n_E, n_C) = (16, 112)$.

results for method F are shown to illustrate its distinctive performance. Clearly, method dL1z provided the most accurate approximation to the exact CI bounds for δ_2. Its width ratios were slightly above 1 but with no substantial impact on coverage probabilities.

Method F showed a very different pattern of results when compared to the other methods. At $\delta_2 = 0$, the empirical coverage probabilities were equal to the nominal $(1 - \alpha)$ values for all sample sizes (even for $N = 8$) and all values of $(1 - \alpha)$, although its width ratios were quite inaccurate. For larger values of $|\delta_2|$, $\hat{P}$ fell below the nominal $(1 - \alpha)$ value, with no improvement in accuracy

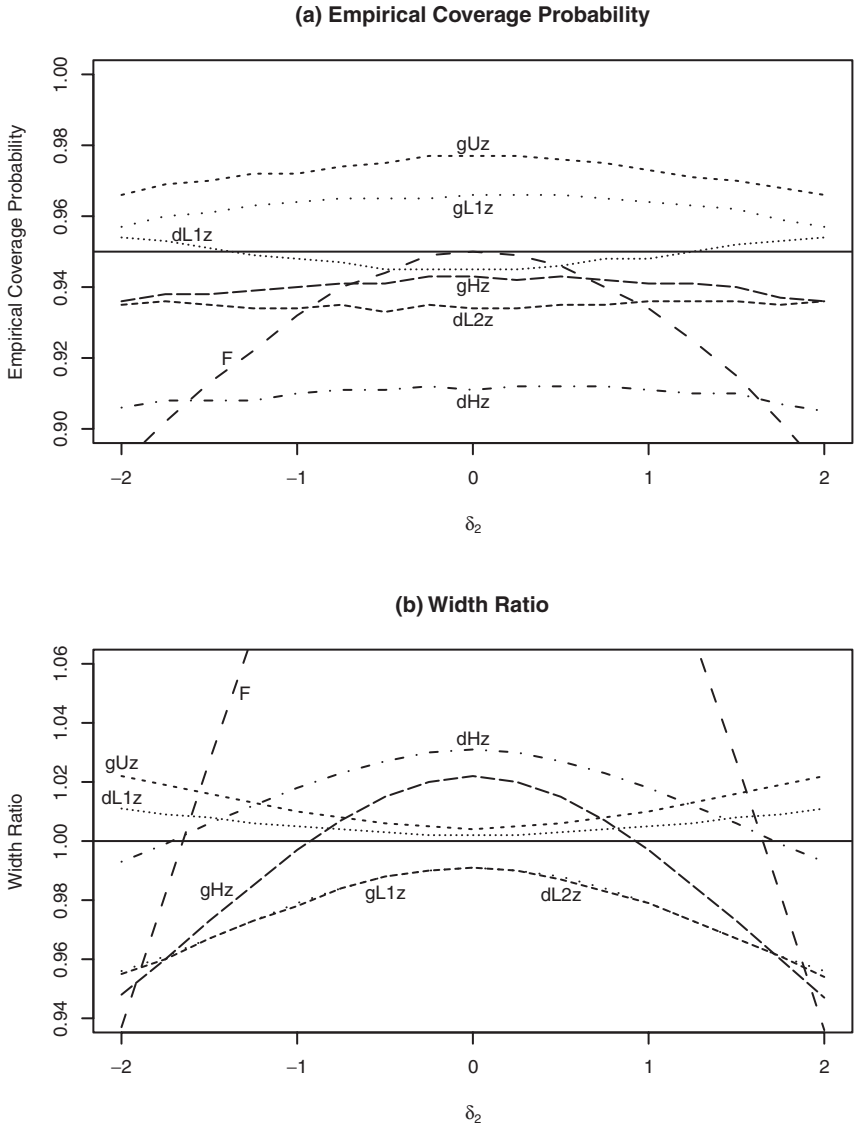

FIGURE 1. *Empirical coverage probability and width ratio in the two-independent sample case for* $\mathrm{N} = 8$ *and 95% confidence intervals (CIs) (Equal Group Sizes).*

for larger sample sizes (see last rows in Table 4). This finding is not surprising as method F essentially assumes that δ_2 is equal to zero. Therefore, no matter how large N becomes, method F will not provide accurate results when $\delta_2 \neq 0$. Finally, method F actually showed an increase in accuracy in the unequal sample

TABLE 5
Maximum Error (in %) in Empirical Coverage Probabilities and Width Ratios Over All Values of δ_D for n = *8 in the Two-Dependent Samples Case With Parameter* δ_D

Confidence	Coverage Probability					Width Ratio				
Interval	50%	70%	90%	95%	99%	50%	70%	90%	95%	99%
gBz	5.9	6.8	4.5	3.1	0.9	8.4	8.6	9.2	9.6	10.6
gBt	8.3	10.3	7.5	4.7	1.0	14.3	17.3	25.8	32.3	50.3
dBz	13.0	14.1	7.9	4.4	1.0	36.1	36.3	37.1	37.6	38.8
dBt	15.7	17.3	9.2	4.9	1.0	43.4	47.2	57.9	66.0	88.6
gUz	3.5	4.4	3.1	2.3	0.8	3.2	3.4	3.9	4.3	5.2
gUt	6.0	8.0	6.6	4.4	1.0	8.8	11.6	19.7	25.8	43.0
dUz	5.5	6.7	4.8	3.2	0.8	16.1	16.4	17.0	17.4	18.5
dUt	8.0	10.4	7.7	4.6	1.0	22.5	25.7	34.8	41.7	61.0
gL1z	−4.3	−4.8	−3.8	−2.9	−1.2	−9.0	−8.8	−8.3	−8.0	−7.2
gL1t	5.8	7.4	5.8	4.0	1.0	4.7	7.3	14.4	19.9	35.0
dL1z	−1.6	−1.5	−0.9	0.8	0.4	1.1	1.4	1.9	2.3	3.2
dL1t	1.6	3.6	4.9	3.8	1.0	6.6	9.4	17.4	23.4	40.2
gL2z	−6.7	−7.7	−6.1	−4.6	−2.0	−14.1	−14.0	−13.5	−13.2	−12.4
gL2t	5.7	7.1	5.4	3.7	1.0	−9.5	−7.1	14.0	19.4	34.5
dL2z	−3.5	−3.4	−1.7	−0.9	−0.2	−4.7	−4.5	−4.0	−3.7	−2.9
dL2t	−1.3	2.1	3.6	3.2	1.0	5.1	7.6	14.8	20.3	35.4
gHz	−6.3	−6.6	−4.2	−2.5	−1.2	−13.7	−13.0	−11.1	−9.7	−6.2
gHt	5.2	6.0	3.8	2.5	0.7	−9.0	8.3	18.3	26.5	52.3
dHz	−3.9	−4.5	−4.4	−3.9	−2.5	−4.3	−3.5	2.5	3.7	6.7
dHt	−1.6	−1.1	1.1	1.0	0.5	5.6	9.0	19.1	27.4	53.4
F	−32.7	−43.3	−46.1	−41.8	−28.0	−67.7	−66.9	−64.5	−62.7	−57.5
F[a]	−32.4	−43.2	−48.0	−46.7	−38.5	−66.8	−66.7	−66.6	−66.5	−66.2

a. $n = 128$.

size conditions, with more extreme splits yielding more accurate $\hat{P}$ values (again, see last rows in Table 4).

Two-Dependent Samples Design With Parameter δ_D

All methods except F again converged to the correct coverage probabilities and interval widths as sample size increased. Specifically, $(\hat{P} - (1 - \alpha)) \times 100\%$ never exceeded 7.6% when $n = 16$, 3.9% when $n = 32$, 2.1% when $n = 64$, and 1.1% when $n = 128$ over all values of δ_D and $(1 - \alpha)$ for methods B, U, L1, L2, and H. For $n = 16$, 32, 64, and 128, the highest width ratio error was 31.7%, 13.9%, 7.5%, and 3.9%, respectively. Table 5 provides the maximum value of $(\hat{P} - (1 - \alpha)) \times 100\%$ and $(\hat{W} - 1) \times 100\%$ for each method and

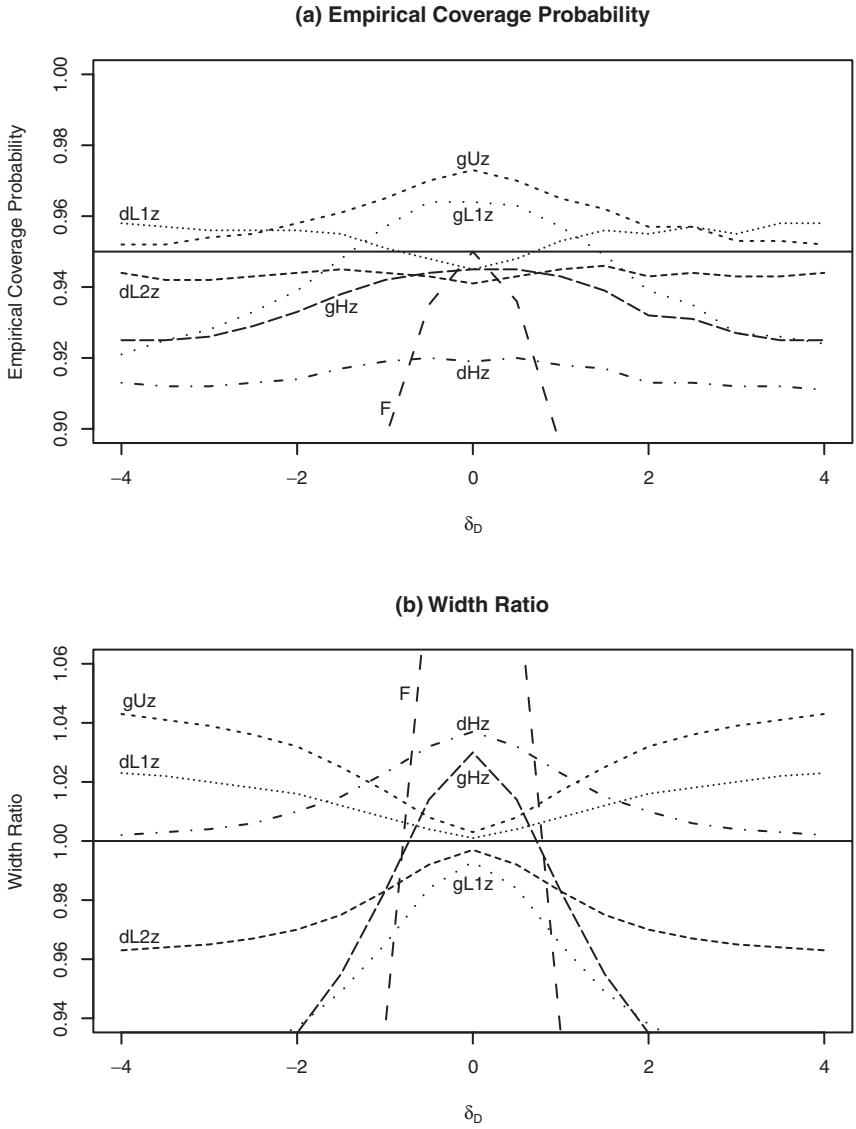

FIGURE 2. *Empirical coverage probability and width ratio in the two-dependent sample case with parameter* δ_D *for* $n = 8$ *and 95% confidence intervals (CIs).*

each CI width over all values of δ_D for $n = 8$. The results are similar to those from the two-independent samples case.

Figure 2 shows the empirical coverage probabilities and width ratios for methods gUz, gL1z, dL1z, dL2z, gHz, dHz, and F for the 95% CI condition

when $n = 8$, which again indicates that method dL1z performed favorably. Because of the wider range of δ_D values considered in this set of simulations, method F performed even worse when compared to the two-independent samples case and again showed no sign of improvement with larger sample sizes.

Two-Dependent Samples Design With Parameter δ_{D2}

Once again, all of the approximations except method F converged to the exact CI bounds in terms of empirical coverage probabilities and widths. The largest error in the coverage probabilities amounted to 8.3% for $n = 16$, 3.9% for $n = 32$, 2.7% for $n = 64$, and 1.7% for $n = 128$. For the width ratios, the largest errors were 27.1%, 16.4%, 8.5%, and 4.4%, respectively. Table 6 provides the maximum value of $(\hat{P} - (1 - \alpha)) \times 100\%$ and $(\hat{W} - 1) \times 100\%$ for each method over all values of δ_{D2} and ρ in the two-dependent samples case when $n = 8$. The approximations were slightly less accurate than in the previous two sets of simulations, which can be attributed to the additional error introduced by having to estimate ρ. Methods gL1z, gL2z, dL2z, and gHz provided the most accurate coverage probabilities and width ratios for small sample sizes.

Empirical coverage probabilities and width ratios for these four methods are plotted in Figure 3 across individual values of δ_{D2} for the 95% CI condition when $\rho = .7$ and $n = 8$. Method dL1z, which provided the most accurate results in the previous two sets of simulations, is also plotted for comparison purposes, whereas method gHz was omitted, as it made the graphs difficult to read (it did not provide more accurate results than the two methods discussed in the following). Here, none of the methods could be considered generally superior in all aspects. In fact, all of these approximations tended to capture the parameter not often enough despite the fact that intervals provided by methods dL1z and dL2z were usually too wide on average. Overall, methods gL1z and dL2z appear to be most accurate in terms of interval widths while still providing quite accurate coverage probabilities.

Discussion

Finding the exact CI for δ requires iterative estimation procedures. However, the present article shows that various approximate methods can be used without concern as long as sample sizes are at least moderately large. On the other hand, when sample sizes are small, then researchers can consult Tables 4, 5, and 6 when choosing an approximation. These tables provide the maximum error over conditions that are not under the control of the experimenter, namely, the true value of δ and the true value of ρ in the two-dependent samples case. In other words, because these parameters are unknown in practice, one cannot pick a method that would be optimal for particular δ and ρ values. A reasonable alternative is a minimax approach, choosing a method that minimizes the maximum possible error.

Ideally, one method would be most accurate for all the designs considered. Method dL1z was extremely accurate in the two-independent and two-dependent

TABLE 6
Maximum Error (in %) in Empirical Coverage Probabilities and Width Ratios Over All Values of δ_{D2} and ρ for n = *8 in the Two-Dependent Samples Case With Parameter δ_{D2}.*

Confidence Interval	Coverage Probability					Width Ratio				
	50%	70%	90%	95%	99%	50%	70%	90%	95%	99%
gBz	5.7	5.4	3.1	1.7	−1.8	17.4	17.1	17.9	18.5	20.1
gBt	8.3	8.9	6.1	4.3	1.0	23.8	26.4	35.8	43.0	63.1
dBz	13.2	14.1	7.9	4.3	1.0	47.4	47.2	48.3	49.1	51.0
dBt	15.8	17.4	9.3	4.8	1.0	55.4	59.0	70.8	79.8	105.1
gUz	3.3	3.1	−3.2	−3.4	−2.8	11.0	10.8	11.6	12.2	13.7
gUt	5.9	6.6	5.1	3.8	1.0	17.8	20.3	29.2	36.1	55.2
dUz	5.1	6.8	4.6	2.9	−1.4	25.0	24.8	25.7	26.3	28.0
dUt	8.1	10.5	7.6	4.4	1.0	32.7	35.4	45.5	53.2	74.8
gL1z	−4.3	−5.2	−−5.6	−4.1	−2.9	−3.6	−3.4	−3.2	−3.0	−3.3
gL1t	4.8	5.6	4.3	3.0	1.0	5.1	7.5	14.7	20.1	36.5
dL1z	−4.1	−5.0	−4.4	−3.2	−1.8	9.6	9.2	10.0	10.5	12.0
dL1t	2.0	4.1	4.7	3.4	1.0	15.5	17.9	26.7	33.4	52.1
gL2z	−6.5	−8.1	−7.7	−5.9	−3.5	−8.2	−8.0	−7.3	−7.2	−6.7
gL2t	4.5	5.4	3.9	2.7	1.0	4.7	7.0	13.6	19.1	35.1
dL2z	−4.3	−5.5	−4.8	−3.8	−2.0	3.2	2.8	3.6	4.0	5.3
dL2t	−2.1	−2.2	3.3	3.0	0.9	8.8	11.0	19.3	25.4	43.1
gHz	−6.3	−6.9	−5.2	−3.5	−2.4	−7.7	−7.0	−4.6	−3.4	6.4
gHt	4.1	4.4	2.5	2.0	0.8	5.2	8.4	17.9	26.1	53.0
dHz	−4.6	−6.5	−7.2	−6.3	−3.8	3.7	3.9	6.5	8.2	12.8
dHt	−2.6	−3.2	−2.3	1.4	0.6	9.4	12.4	23.8	32.9	62.1
F	−35.1	−47.5	−52.9	−49.9	−37.9	−69.1	−68.3	−66.0	−64.2	−59.1
F[a]	−33.6	−44.3	−51.2	−49.1	−40.9	−69.8	−69.8	−69.7	−69.6	−69.5

a. $n = 128$.

samples case with parameter δ_D. However, it did not perform quite as well in the two-dependent samples case with parameter δ_{D2}, where methods gL1z and dL2z provided the most accurate results. The results in the two-dependent samples case were of particular interest because methods to find the exact CI of δ_{D2} without knowledge of ρ have not been developed yet. Therefore, researchers must rely on approximations in practice and methods gL1z and dL2z are to be recommended here. The fact that methods based on the large-sample variances provided some of the most accurate approximations to the exact CI bounds is certainly a welcomed finding because these methods are also the easiest to compute among all the suggested approximations.

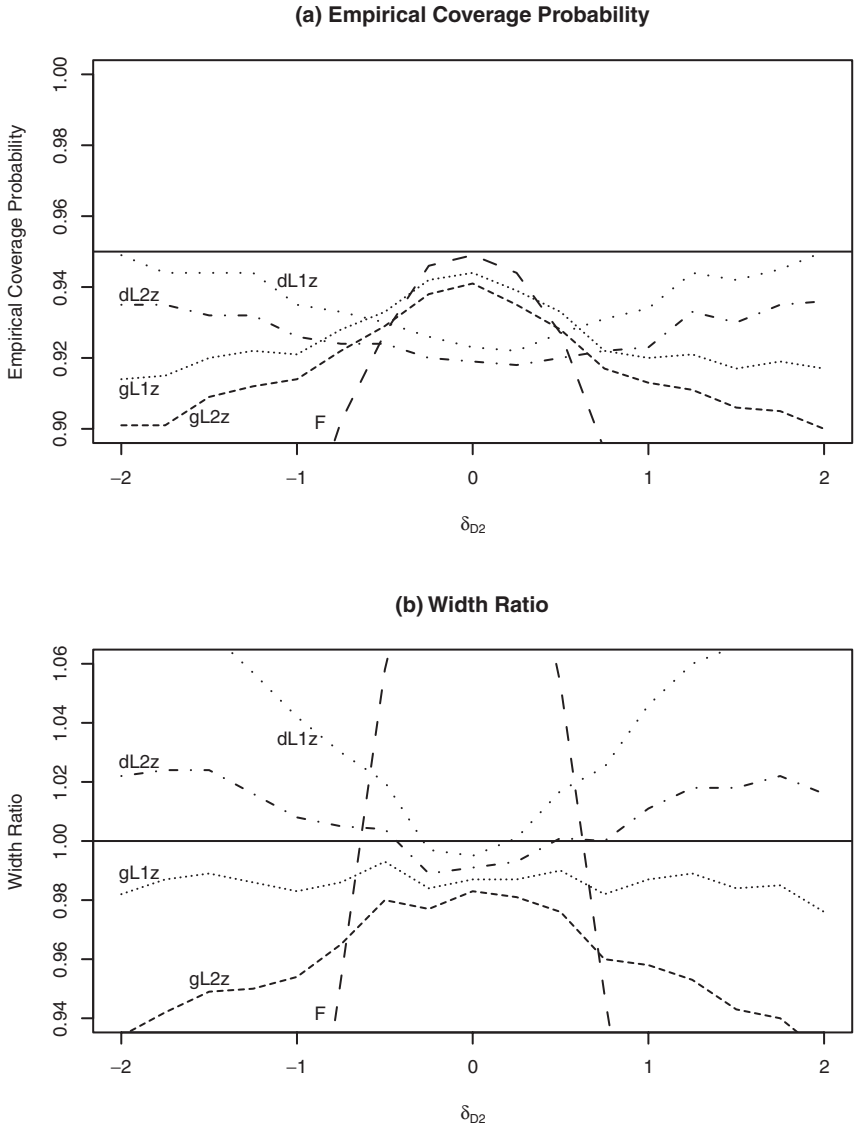

FIGURE 3. *Empirical coverage probability and width ratio in the two-dependent sample case with parameter* δ_{D2} *for* $n = 8$, $\rho = .7$, *and 95% confidence intervals (CIs).*

A final word of caution is warranted when discussing standardized effect sizes. The standardized effect size δ is useful when comparing results from multiple studies using measurement instruments whose raw units are not directly comparable. If the different instruments provide scores that are linear transformations of

each other, then standardizing the raw effect sizes allows comparisons across different instruments. The problem with standardized effect sizes is their dependence on the amount of variability in the population. The problem is twofold. First of all, δ assumes homoscedasticity of the scores in the groups or across repeated measures. When σ is not homogeneous, use of δ might be problematic. However, using standardized units creates an even more notable problem because δ depends on the particular characteristics of the population being studied, specifically, its variance (Cohen, 1994). In other words, two *d* or *g* values for the same outcome measure obtained from two experiments could be incommensurable if the samples were drawn from populations with unequal variances.

Cohen (1994) emphasized that researchers must begin to "respect the units they work with" (p. 1001). In the ideal case where the raw units of measurement have a natural interpretation and are consistent across multiple studies, it is not necessary to standardize the effect size. Using raw units eliminates the dependency of the effect size on the population variance. CIs for effect sizes in raw units are easily obtained and are exact. However, in the social sciences, the multitude of scales and measurement instruments will necessitate the use of standardized effect sizes in the future.

References

American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author.

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

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

Carlson, J. G., Chemtob, C. M., Rusnak, K., Hedlund, N. L., & Muraoka, M. Y. (1998). Eye movement desensitization and reprocessing (EMDR) treatment for combat-related posttraumtic stress disorder. *Journal of Traumatic Stress*, *11*, 3-24.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997-1003.

Cumming, G. (2003). Exploratory software for confidence intervals [Computer software]. Retrieved January 25, 2007, from http://www.latrobe.edu.au/psy/esci/index.html

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.

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

Gibbons, R. D., Hedeker, D. R., & Davis, J. M. (1993). Estimation of effect size from a series of experiments involving paired comparisons. *Journal of Educational Statistics*, *18*, 271-279.

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

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

Hedges, L. V. (1983). A random effects model for effect sizes. *Psychological Bulletin*, *93*, 388-395.

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

Hyde, J. S. (2001). Reporting effect sizes: The roles of editors, textbook authors, and publication manuals. *Educational and Psychological Measurement*, *61*, 225-228.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, *56*, 746-759.

Marcus, S. V., Marquis, P., & Sakai, C. (1997). Controlled study of treatment of PTSD using EMDR in an HMO setting. *Psychotherapy*, *34*, 307-315.

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

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods*, *7*, 105-125.

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

Ray, J. W., & Shadish, W. R. (1996). How interchangeable are different estimators of effect size? *Journal of Consulting and Clinical Psychology*, *64*, 1316-1325.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, *1*, 115-129.

Smithson, M. J. (2003a). *Confidence intervals*. Thousand Oaks, CA: Sage.

Smithson, M. J. (2003b). Scripts and software for noncentral confidence interval and power calculations [Computer software]. Retrieved January 25, 2007, from http://www.anu.edu.au/psychology/people/smithson/details/CIstuff/CI.html

Steiger, J. H., & Fouladi, R. T. (1997). Noncentrality interval estimation and the evaluation of statistical models. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 221-257). Mahwah, NJ: Lawrence Erlbaum.

Thompson, B. (Ed.). (2001). Confidence intervals for effect sizes [Special section]. *Educational and Psychological Measurement*, *61*, 517-667.

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher*, *31*, 25-32.

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

Author

WOLFGANG VIECHTBAUER is Assistant Professor, Department of Methodology and Statistics, at the University of Maastricht; P.O. Box, 6200 MD Maastricht, The Netherlands; e-mail: wolfgang.viechtbauer@stat.unimaas.nl. His research interests include mixed-effects models, meta-analysis, multilevel modeling, longitudinal analysis, and effect size measures.

Manuscript received June 28, 2004
Accepted January May 4, 2005