THE EFFECTS OF NONNORMAL DISTRIBUTIONS ON CONFIDENCE INTERVALS AROUND THE STANDARDIZED MEAN DIFFERENCE: BOOTSTRAP AND PARAMETRIC CONFIDENCE INTERVALS

KEN KELLEY
University of Notre Dame

The standardized group mean difference, Cohen's *d*, is among the most commonly used and intuitively appealing effect sizes for group comparisons. However, reporting this point estimate alone does not reflect the extent to which sampling error may have led to an obtained value. A confidence interval expresses the uncertainty that exists between *d* and the population value, δ, it represents. A set of Monte Carlo simulations was conducted to examine the integrity of a noncentral approach analogous to that given by Steiger and Fouladi, as well as two bootstrap approaches in situations in which the normality assumption is violated. Because *d* is positively biased, a procedure given by Hedges and Olkin is outlined, such that an unbiased estimate of δ can be obtained. The bias-corrected and accelerated bootstrap confidence interval using the unbiased estimate of δ is proposed and recommended for general use, especially in cases in which the assumption of normality may be violated.

Keywords: *effect size; standardized effect size; confidence intervals; bootstrap methods; nonnormal data*

Methodological recommendations within the behavioral sciences have increasingly emphasized the importance and utility of confidence intervals (Cumming & Finch, 2001; Smithson, 2001), effect sizes (Olejnik & Algina,

I would like to thank Ke-Hai Yuan and Joseph R. Rausch, both of the Department of Psychology at the University of Notre Dame, for helpful comments and suggestions during the preparation of this article. Correspondence concerning this article should be addressed to Ken Kelley, 118 Haggar Hall, Department of Psychology, University of Notre Dame, Notre Dame, IN 46556; e-mail: kkelley@nd.edu.

Educational and Psychological Measurement, Vol. 65 No. 1, February 2005 51-69
DOI: 10.1177/0013164404264850

2000; Roberts & Henson, 2002), and confidence intervals around effect sizes (Steiger & Fouladi, 1997; Thompson, 2002). Along these lines, the American Psychological Association's Task Force on Statistical Inference (Wilkinson & Task Force on Statistical Inference, 1999, p. 599) recommended presenting effect sizes for primary outcomes, as well as forming confidence intervals for effect sizes involving such primary outcomes. Thus, the future of quantitative behavioral science research may indeed be based in large part on confidence intervals around effect sizes, while being less reliant on null hypothesis significance testing (Thompson, 2002).

In the context of group comparisons, the most commonly used and perhaps the most intuitively appealing effect size is the standardized difference between two group means, typically termed Cohen's *d* (Cohen, 1988, chap. 3) or sometimes Hedges's *g* (Hedges, 1981). The population-standardized difference between groups means is defined by

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

where μ_j is the population mean for the j^{th} group ($j = 1, 2$); and σ is the population standard deviation, assumed equal for the two groups. The most commonly used estimate of δ is given by

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

where $\overline{X}_j$ is the sample mean from the j^{th} group; and *s* is the square root of the usual estimate of the mean square within, that is, the square root of the unbiased estimate of the pooled variance. Even though *d* is typically used as an estimate of δ, *d* is known to be positively biased (Hedges & Olkin, 1985, chap. 5), a complication that will be dealt with momentarily (note that Glass's *g′* is a variation on δ, where the group mean difference is divided by the control group standard deviation rather than the pooled within-group standard deviation; Glass, 1976).

The use of equation 2 provides a measure of the effect group membership has on the mean difference, assuming a common standard deviation, that is scale free (standardized). As with any point estimate, reporting *d* in the absence of a confidence interval arguably does a disservice to those otherwise interested in the phenomenon under study. A point estimate alone fails to convey the uncertainty associated with an estimate as it relates to the corresponding population parameter. It is the population value that is of interest, not the observed point estimate. Thus, whenever there is an interest in *d*, there should also be an interest in the limits of the confidence interval that probabilistically bound the value of δ.

It is well known that data are often not normally distributed in behavioral research (Micceri, 1989). It is also well known that most inferential statistical methods assume that data are independently sampled from some normally distributed population. Thus, analyzing nonnormal data by way of procedures that assume normality can have serious implications for the conclusions reached from (mis)using such inferential techniques.

The purpose of the present article is to explore the appropriateness of three contending methods for forming confidence intervals around the population standardized mean difference when the assumption of normality is violated. Specifically, the article examines an exact parametric method based on a noncentral *t* distribution, the bootstrap percentile method, and the bootstrap bias-corrected and accelerated method. These three methods are evaluated in terms of the accuracy of the confidence interval coverage, the precision of the estimate, and statistical power.

Methods of Formulating Confidence Intervals Around δ

Two classes of confidence intervals are explored in the present article. The first method is a parametric procedure based on the standard general linear model assumptions that (a) data are randomly sampled from a normally distributed parent population conditional on group membership, (b) homogeneity of variance exists for all groups of interest (only two groups are considered in the present article, yet the ideas extend to *j* groups), and (c) the units of analysis are independent of one another. This method of confidence interval formulation is analogous to that described in Steiger and Fouladi (1997) and is based on a noncentral *t* distribution (also see Cumming & Finch, 2001, for another procedure). The second method of confidence interval formulation is based on two different nonparametric bootstrap resampling procedures. In the following two subsections, an overview of the framework and details of the application is given for each of the two bootstrap methods.

Parametric Confidence Interval

When the assumptions are met and the null hypothesis is true, the difference between two group means is normally distributed about zero. When this difference is divided by its standard error, it follows a central *t* distribution with $\nu = n_1 + n_2 - 2$ degrees of freedom, where n_j is the sample size for the j^{th} group. However, when the null hypothesis is false, the difference between the means divided by its standard error does not follow a central *t* distribution; rather, it follows a nonsymmetric distribution that is known as a noncentral *t* distribution with ν degrees of freedom and noncentrality parameter λ. The noncentrality parameter is a function of δ and the within-group sample sizes:

$$\lambda = \delta\sqrt{\frac{n_1 n_2}{n_1 + n_2}}. \tag{3}$$

The observed t value, $t_{\text{obs.}}$, is used to estimate the noncentrality parameter λ.

By the confidence interval transformation principle (Steiger & Fouladi, 1997, p. 234), finding the confidence limits for λ leads to the confidence limits for δ. (Although a one-sided confidence interval around δ may be of interest in certain circumstances, the discussion in the present article is restricted to two-sided confidence intervals. However, the calculation of a one-sided confidence interval for δ is straightforward given the ensuing discussion.) The lower confidence limit for λ is obtained by finding the noncentral parameter whose $1 - \alpha / 2$ quantile is $t_{\text{obs.}}$. Likewise, the upper confidence limit for λ is obtained by finding the noncentral parameter whose $\alpha / 2$ quantile is $t_{\text{obs.}}$. These upper and lower limits bracket λ with 100(1 – α)% confidence. The noncentral confidence limits for λ can be obtained in a straightforward fashion with the following SAS syntax:

$$\text{LowNC\_CV} = \text{TNONCT}(t_{\text{obs.}}, \nu, 1 - \alpha / 2),$$

and

$$\text{UpNC\_CV} = \text{TNONCT}(t_{\text{obs.}}, \nu, \alpha / 2),$$

where LowNC_CV and UpNC_CV are the lower $t'_{(\alpha/2,\nu,\lambda)}$ and upper $t'_{(1-\alpha/2,\nu,\lambda)}$ critical values from the particular noncentral t distribution. The critical values can also be obtained using R or S-Plus. (The critical values themselves are not directly available in the computer programs R and S-Plus, but special scripts were developed to obtain the critical values in these programs and are available on request.)

Once the confidence limits for λ have been obtained, they can be transformed into confidence limits for δ. This transformation holds because δ is a function of λ and the within-group sample size. The confidence interval around δ is computed in the following manner:

$$\text{Prob.}\left[t'_{(\alpha/2,\nu,\lambda)}\sqrt{\frac{n_1 + n_2}{n_1 n_2}} \le \delta \le t'_{(1-\alpha/2,\nu,\lambda)}\sqrt{\frac{n_1 + n_2}{n_1 n_2}}\right] = 1 - \alpha. \tag{4}$$

Thus, given that the statistical assumptions are met, equation 4 provides the 100(1 – α)% confidence limits around δ. However, it is important to realize that to the extent that the assumptions are not met, equation 4 can potentially yield misleading confidence interval limits such that the empirical coverage is greater or less than the nominal coverage specified. Momentarily, the statistical validity of this procedure will be examined when the assumption of normality is violated via a set of Monte Carlo simulations.

Bootstrap Confidence Intervals

The general bootstrap technique is a resampling procedure whereby random samples are repeatedly drawn from the set of observed data a large number of times (say 10,000) to study the distribution of the statistic(s) of interest given the obtained data. In the present context, interest lies in examining the distribution of the B bootstrapped d values, where $\mathbf{d}^*$ represents the vector of length B of the bootstrap results such that nonparametric confidence limits can be formed. The bootstrap procedure makes no assumption about the parent population from which the data were drawn other than that the data are randomly sampled and thus representative of the parent population.

Within the bootstrap framework, two methods of confidence interval formulation are delineated. The first type is the percentile method, whereby the values representing the $\alpha / 2$ and $1 - \alpha / 2$ quantiles of the empirical bootstrap distribution are taken as the confidence limits. That is, the confidence limits from the percentile method are obtained simply by finding the values from the bootstrap distribution, $\mathbf{d}^*$, that correspond to the α / 2 and $1 - \alpha$ / 2 cumulative probabilities.

The percentile method is first-order accurate. The order of accuracy in the sense of confidence intervals is the rate at which the errors of over- or undercoverage of the $100(1 - \alpha)\%$ confidence interval limits approach zero. First-order accuracy means that the error of the percentage of confidence interval coverage approaches zero at a rate related to $1/\sqrt{\min(n_1, n_2)}$ (Efron & Tibshirani, 1993, p. 187). The second type of bootstrap confidence interval of interest, and the one to be generally recommended, is the bias-corrected and accelerated confidence interval (BC*a*). The BC*a* is second-order accurate, meaning that the over- or undercoverage of the $100(1 - \alpha)\%$ confidence interval approaches zero at a rate related to $1 / \min(n_1, n_2)$ (Efron & Tibshirani, 1993, p. 187).

The computation of the BC*a* proceeds in three steps. First, a bootstrap sample of size B is collected, as was the case with the percentile method. Rather than just stopping there, however, a bias correction value is obtained, as is an acceleration value. The bias correction value, $\hat{z}_0$, is obtained by calculating the proportion of the $\mathbf{d}^*$ values that are less than the sample d and then finding the quantile from the normal distribution with that cumulative probability:

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

where Φ is the standard normal cumulative distribution function and Φ^{-1} its inverse (e.g., $\Phi[1.645] = 0.95$ and $\Phi^{-1}[0.975] = 1.96$), and # is read as "the number of." The acceleration value, $\hat{a}$, is obtained by first performing a jackknife procedure, whereby d is calculated N times, once after the i^{th} case ($i =$

1, . . . , N) has been deleted. Let $d_{(-i)}$ be the value of d when the i^{th} data point has been deleted and $\tilde{d}$ be the mean of the N jackknifed $d_{(-i)}$ values. The acceleration parameter is then computed as follows:

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

Details of the rationale for the use of the BC*a* are given in Efron and Tibshirani (1993, chap. 14).

Once $\hat{z}_0$ and $\hat{a}$ have been calculated, the limits of the confidence interval are calculated by finding the values from the bootstrap sample that correspond to the CI_{Low} and CI_{Up} quantiles of the observed bootstrap distribution. The CI_{Low} and CI_{Up} values are found from the following formulations:

$$\text{CI}_{\text{Low}} = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(\alpha/2)}}{1 - \hat{a}(\hat{z}_0 + z^{(\alpha/2)}}\right), \tag{7}$$

and

$$\text{CI}_{\text{Up}} = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(1-\alpha/2)}}{1 - \hat{a}(\hat{z}_0 + z^{(1-\alpha/2)}}\right), \tag{8}$$

such that CI_{Low} and CI_{Up} represent the quantiles from the distribution of $\mathbf{d}^*$. That is, the confidence limits from the bias-corrected and accelerated approach are obtained by finding the values from the bootstrap distribution, $\mathbf{d}^*$, that correspond to the CI_{Low} and CI_{Up} cumulative probabilities. It should be pointed out that when $\hat{a}$ and $\hat{z}_0$ equal zero, $\text{CI}_{\text{Low}} = \alpha / 2$ and $\text{CI}_{\text{Up}} = 1 - \alpha / 2$, which corresponds to the values obtained from the percentile method. Appendix A provides syntax to calculate confidence intervals on the basis of the percentile method and the generally recommended bias-corrected and accelerated method in R or S-Plus. (It should be noted that much of the discussion regarding the noncentral approach and the bootstrap techniques is also applicable to g'. The noncentral approach can be modified for a confidence interval around the population g', as was done for δ, by adjusting the degrees of freedom to $n_c - 1$, where n_c is the control group sample size (Hedges & Olkin, 1985, chap. 5). Furthermore, minor modifications can be made to the bootstrap procedures so that they can be applied to a confidence interval for the population g'.)

Estimating δ in an Unbiased Fashion

The commonly used estimate of δ is given by equation 2. However, $E[d] = \delta / G(\nu)$, where

$$\mathrm{G}(v) = \frac{\Gamma(v/2)}{\sqrt{v/2}\,\Gamma((v-1)/2)}, \tag{9}$$

and $\Gamma(k)$ is the γ function evaluated at k (Hedges & Olkin, 1985, p. 104). An unbiased estimate of δ, d_u, can thus be obtained in the following manner:

$$d_u = dG(\nu). \tag{10}$$

The bias in d is particularly problematic for a small ν. Hedges and Olkin (1985) provided a table (Table 2, p. 80) from which the values of $G(\nu)$ can be obtained for ν from 2 to 50. Appendix B provides a program with a function for calculating d and d_u using R or S-Plus. (Note that Hedges & Olkin, 1985, p. 79, gave an approximation to d_u that does not require the use of the Γ function. The approximation of d_u is given as $d\{1 - 3 / [4\nu - 1]\}$. However, because d_u is an unbiased estimate, it is to be preferred.)

Methods of the Monte Carlo Simulations

The nonnormal data were generated using Fleishman's (1978) power method. This method yields a random variate distributed with mean zero and variance one, for which the skew and kurtosis are specified for the particular nonnormal distribution. The nonnormal data are generated by first generating random variates from a standard normal distribution and then transforming the variates by a polynomial equation of order three with coefficients specific to a particular case of nonnormality. The necessary coefficients are given in Table 1 of Fleishman (pp. 524-525).

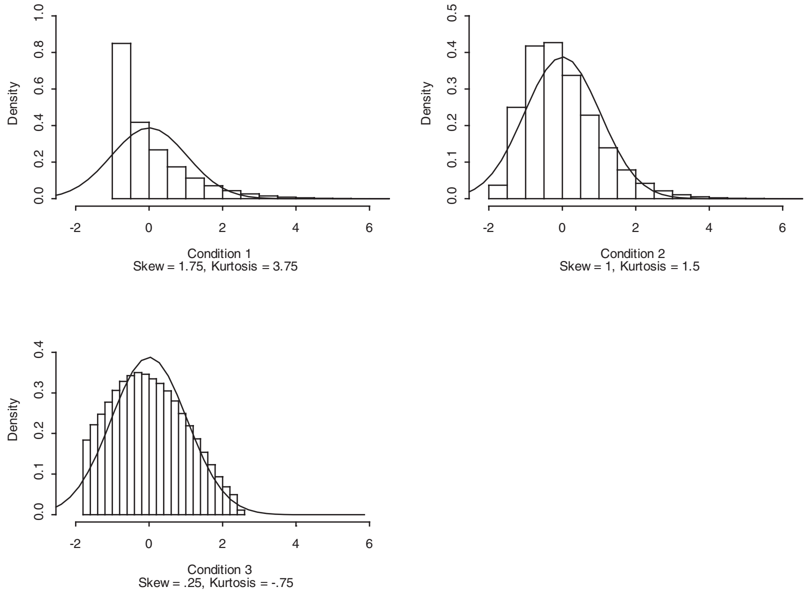
Skew and kurtosis are defined respectively as

$$\gamma_1 = \frac{\mu_3}{\mu_2^{3/2}} = \frac{\kappa_3}{\kappa_2^{3/2}}, \tag{11}$$

and

$$\gamma_2 = \frac{\mu_4}{\mu_2^2} - 3 = \frac{\kappa_4}{\kappa_2^2}, \tag{12}$$

where μ_r is the r^{th} central moment, and κ_r is the r^{th} cumulant of the particular distribution (Stuart & Ord, 1994, chap. 3). It should be noted that γ_1 and γ_2 are zero for the normal distribution.



Condition 1
Skew = 1.75, Kurtosis = 3.75

Condition 2
Skew = 1, Kurtosis = 1.5

Condition 3
Skew = .25, Kurtosis = -.75

Figure 1. Illustration of the three nonnormal distributions, with a normal density line included for comparative purposes.

Four different distributional forms were examined. To gauge the effectiveness of the three confidence interval procedures, the normal distribution was examined along with three nonnormal distributions. The nonnormal distributions were chosen because they were thought to be realistic representations of distributions encountered in the behavioral sciences. The skew and kurtosis of each condition are given, along with their graphical depiction accompanied by a normal density line, in Figure 1.

Results

The results for each of the 72 scenarios (Table 1: three sample sizes by three methods for d and two methods for d_{u}; Table 2: three sample sizes by three distributions by two methods for d and one method for d_{u}; Table 3: five effect sizes by three distributions by two methods for d) are based on 10,000 replications in the program R. For the bootstrap conditions, there were B = 10,000 bootstrap samples drawn within each of the 10,000 replications of the Monte Carlo procedure. Table 1 provides a comparison of the noncentral method, the bootstrap percentile method, and the bootstrap bias-corrected and accelerated method using both d and d_u as estimates of δ. Table 1 shows

Table 1
*Results for the Noncentral (NC), Bootstrap (BS) Percentile, and BS Bias-Corrected and Accelerated (BC*a*) Confidence Intervals When All Parametric Assumptions Were Met*

		d Estimate		d_u Estimate	
Statistic	NC	BS Percentile	BS BC*a*	BS Percentile	BS BC*a*
Skew = 0, kurtosis = 0 (normal distribution), $n_1 = n_2 = 5$, $\delta = 0$					
% of coverage	95.0100	87.9900	95.5100	87.6200	95.2200
M width	2.5731	3.5093	3.2819	3.5154	3.2822
Median width	2.5165	3.2458	3.1387	3.2459	3.1417
Mean low bound	–1.2802	–1.7406	–1.6378	–1.7673	–1.6475
Median low bound	–1.2346	–1.5600	–1.6071	–1.5812	–1.6129
SD low bound	0.7192	1.4194	0.8504	1.4614	0.8481
M up bound	1.2929	1.7687	1.6442	1.7481	1.6347
Median up bound	1.2446	1.5806	1.6085	1.5533	1.6005
SD up bound	0.7206	1.3926	0.8214	1.4185	0.8516
Skew = 0, kurtosis = 0 (normal distribution), $n_1 = n_2 = 15$, $\delta = 0$					
% of coverage	94.8400	92.9900	95.9600	93.2800	96.3600
M width	1.4447	1.5232	1.5157	1.5247	1.5168
Median width	1.4370	1.5111	1.5086	1.5120	1.5091
M low bound	–0.7224	–0.7606	–0.7566	–0.7618	–0.7578
Median low bound	–0.7175	–0.7526	–0.7537	–0.7534	–0.7554
SD low bound	0.3741	0.4284	0.8504	0.4274	0.3653
M up bound	0.7223	0.7627	1.6442	0.7629	0.7590
Median up bound	0.7138	0.7499	1.6085	0.7501	0.7590
SD up bound	0.3745	0.4272	0.8214	0.4271	0.3654
Skew = 0, kurtosis = 0 (normal distribution), $n_1 = n_2 = 50$, $\delta = 0$					
% of coverage	94.9500	94.3900	95.3900	94.0400	94.9500
M width	0.7860	0.7964	0.7964	0.7964	0.7963
Median width	0.7849	0.7955	0.7957	0.7956	0.7956
M low bound	–0.3936	–0.3964	–0.3966	–0.3972	–0.3972
Median low bound	–0.3932	–0.3966	–0.3972	–0.3955	–0.3949
SD low bound	0.2022	0.2100	0.2007	0.2111	0.2018
M up bound	0.3925	0.4000	0.3998	0.3992	0.3991
Median up bound	0.3907	0.3979	0.3978	0.3982	0.3992
SD up bound	0.2023	0.2100	0.2008	0.2111	0.2018

the results in the special case in which the null hypothesis was true and all of the statistical assumptions were met. The sample sizes examined were chosen to represent very small (5), small (15), and medium (50) per-group sample sizes that may arise in applied research settings. Perhaps the single most important value given in Table 1 is the percentage of confidence intervals whose bounds correctly bracketed the population value (% of coverage),

Table 2
*Results for the Noncentral (NC), Bootstrap (BS) Percentile, and BS Bias-Corrected and Accelerated (BC*a*) Confidence Intervals When the Null Hypothesis Was True and Both Sample Distributions Followed the Same Nonnormal Distribution*

Statistic	NC	BS BC*a* (d)	BS BC*a* (d_u)
Condition 1 (skew = 1.75, kurtosis = 3.75), δ = 0			
$n_1 = n_2 = 5$			
% of coverage	96.1400	96.0700	96.5200
M width	2.5683	3.2278	3.2326
Median width	2.5195	3.0792	3.0869
$n_1 = n_2 = 15$			
% of coverage	95.5600	95.2100	95.2500
M width	1.4445	1.4803	1.4819
Median width	1.4376	1.4864	1.4877
$n_1 = n_2 = 50$			
% of coverage	95.2000	95.8500	95.3700
M width	0.7860	0.7872	0.7868
Median width	0.7849	0.7898	0.7893
Condition 2 (skew = 1, kurtosis = 1.5), δ = 0			
$n_1 = n_2 = 5$			
% of coverage	95.4000	96.1900	95.9200
M width	2.5709	3.2649	3.2679
Median width	2.5171	3.1250	3.1288
$n_1 = n_2 = 15$			
% of coverage	95.2600	95.9800	95.8800
M width	1.4443	1.5043	1.5056
Median width	1.4371	1.5044	1.5014
$n_1 = n_2 = 50$			
% of coverage	95.4600	95.3400	95.3400
M width	0.7860	0.7929	0.7929
Median width	0.7850	0.7933	0.7933
Condition 3 (skew = 1, kurtosis = 1.5), δ = 0			
$n_1 = n_2 = 5$			
% of coverage	94.9600	95.9100	95.7900
M width	2.5731	3.3243	3.3320
Median width	2.5142	3.1639	3.1617
$n_1 = n_2 = 15$			
% of coverage	94.9300	96.5000	96.3300
M width	1.4449	1.5282	1.5290
Median width	1.4371	1.5161	1.5163
$n_1 = n_2 = 50$			
% of coverage	95.5600	95.3600	94.8500
M width	0.7860	0.7990	0.7990
Median width	0.7849	0.7978	0.7977

which was specified to be 95% for all scenarios. The means, medians, and standard deviations of the confidence bounds are also included. These descriptive values illustrate the relative precision of each of the methods. Notice that in all cases, the noncentral method outperformed the bootstrap

Table 3

Results for the Noncentral (NC) and Bootstrap (BS) Bias-Corrected and Accelerated (BCa) Confidence Intervals When the Null Hypothesis Was False and Statistical Power Was 0.80 in Each of Five Different Effect-Size Scenarios (under the incorrect assumption of normality conditional on group), for Which One Sample Distribution Was Normal and the Other Nonnormal to the Degree Specified

	δ = 0.20, $n_1 = n_2 = 394$		δ = 0.50, $n_1 = n_2 = 64$		δ = 0.80, $n_1 = n_2 = 26$		δ = 1.00, $n_1 = n_2 = 17$		δ = 1.60, $n_1 = n_2 = 8$	
Statistic	NC	BS BC*a*	NC	BS BC*a*	NC	BS BC*a*	NC	BS BC*a*	NC	BS BC*a*
Condition 1										
(skew = 1.75, kurtosis = 3.75)										
% of coverage	93.8700	95.4400	92.2900	95.0400	90.3200	94.4700	89.2400	94.1800	88.0500	90.0900
M width	0.2801	0.2875	0.7060	0.7517	1.1440	1.2477	1.4597	1.6060	2.4255	2.6378
Median width	0.2800	0.2873	0.7040	0.7460	1.1342	1.2275	1.4369	1.5667	2.3282	2.5223
Empirical power × 100	79.5700	79.1100	78.6700	75.7800	78.7600	72.5400	78.7000	70.6300	81.5800	72.2000
Condition 2										
(skew = 1.00, kurtosis = 1.50)										
% of coverage	94.4600	94.7300	93.3100	95.1900	92.6100	94.8600	91.8200	94.2600	91.2800	91.1000
M width	0.2801	0.2877	0.7057	0.7516	1.1406	1.2490	1.4511	1.6103	2.3851	2.6259
Median width	0.2800	0.2874	0.1520	0.1296	1.1327	1.2279	1.4338	1.5695	2.3097	2.4967
Empirical power × 100	79.5900	79.5400	78.7500	75.3600	79.4400	72.1800	79.1200	70.0800	82.2800	72.0400
Condition 3										
(skew = 0.25, kurtosis = –0.75)										
% of coverage	94.6400	95.4300	94.3000	94.6800	94.5600	94.3500	93.9000	94.1300	94.9600	90.4800
M width	0.2801	0.2876	0.7053	0.7521	1.1382	1.2487	1.4461	1.6090	2.3549	2.6320
Median width	0.2800	0.2873	0.7039	0.7458	1.1320	1.2285	1.4321	1.5735	2.2925	2.5066
Empirical power × 100	79.8600	78.8800	79.2500	75.2200	80.1900	72.8800	80.2100	69.8600	84.0400	72.4100

percentile method and the bootstrap bias-corrected and accelerated method in terms of the precision of the confidence intervals and the variability of the widths of the confidence intervals. This is no surprise, because statistical tests with the strongest assumptions are generally the most powerful tests when their assumptions are satisfied (Seigel, 1956, chap. 3)

The results of the simulations given in Table 1 showed that there was an undercoverage problem with the bootstrap percentile method for the particular conditions examined. As the sample size grew larger, the empirical percentage of coverage began to approach the nominal value specified. However, this method was uniformly outperformed by the bootstrap bias-corrected and accelerated procedure. For this reason, the percentile method is not further considered as a viable option, and it is not recommended as a method for forming confidence intervals around δ. Although each of the noncentral results provided confidence interval coverage near the nominal value, this was in large part because the statistical assumptions were met. The simulations in this condition were conducted to obtain summary statistics (e.g., mean and median width and the standard deviation of the upper and lower bounds) for the purpose of comparison with the bootstrap procedures in the ideal case in which the assumptions are satisfied. Although it will be shown that there is a real benefit when using the bias-corrected and accelerated method when normality does not hold, it is shown in Table 1 that it is effective when the assumptions are met, because the results closely approximate those of the parametric method.

When the assumptions of a statistical procedure are not met, the meaning of the parametric results can be misleading. Simulations were conducted to investigate how nonnormal parent populations would affect the performance of the noncentral confidence interval procedure, as well as to examine the effectiveness of the bias-corrected and accelerated procedure. Using the three nonnormal distributions illustrated in Figure 1, the results of the noncentral method were compared with the results of the bias-corrected and accelerated method for both d and d_u. In this set of simulations, the null hypothesis was again true, and both of the groups had the same nonnormal parent population.

As with the results in Table 1, the sample sizes per group were 5, 15, and 50. The results in Table 2 illustrate that when both distributions follow the same nonnormal distributional form and when the null hypothesis is true, the empirical and nominal confidence interval coverage are very close to each other for both the noncentral method and the bias-corrected and accelerated method. Notice that Table 2 is not as detailed as Table 1, but the results given in Table 2 are of more interest for comparative purposes given space requirements. Because the null hypothesis is true ($\lambda = 0$), it is no surprise that the noncentral method works well, because it is well known that the t test is robust to violations of normality when the other assumptions are met and the null hypothesis is true (Boneau, 1960; Sawilowsky & Blair, 1992). It should be pointed out that the bias-corrected and accelerated procedure also worked well in terms of the appropriate confidence interval coverage, yet the confidence intervals still tended to be wider than those of the noncentral method for smaller sample sizes. As Wilcox (1998) has pointed out, when differences do exist among groups, "standard methods are not robust" (p. 300). Thus, even though the noncentral method worked well when the null hypothesis was true, this means little under conditions in which the researcher is typically interested, that is, when the null hypothesis is false. However, it is noteworthy that the overall performance of the bias-corrected and accelerated method was similar to that of the noncentral method. The next subsection examines the results when the null hypothesis is false.

Results When the Null Hypothesis Is False

Most researchers are interested in cases in which it is believed that the null hypothesis is false. Thus, it is especially important to evaluate the integrity of statistical procedures when the assumptions they are based on are violated and when the null hypothesis is false. Because the bootstrap confidence intervals formed when using d and d_u are essentially equivalent, except for the center of the interval, for clarity, only the results for the noncentral method and the bias-corrected and accelerated method for d are presented. Table 3 provides a comparison of the noncentral method and the bias-corrected and accelerated method for five different effect sizes ($\delta = 0.20$, 0.50, 0.80, 1.00, and 1.60), for which sample size was chosen such that statistical power (based on normality) was 0.80 in each condition. The first row in each of the three conditions identifies the effect size and the within-group sample size. The effect sizes chosen corresponded to Cohen's (1988, sect. 2.2.4) definition of "small" ($\delta = 0.20$), "medium" ($\delta = 0.50$), and "large" ($\delta = 0.80$), and the remaining two effect sizes ($\delta = 1.00$ and 1.60) were chosen because they represent effects that are generally considered substantially large. In situations in which an effect size is hypothesized to be very large, the corresponding sample size for adequate power to detect a nonzero effect can be small. Thus, the within-group sample sizes used in Table 3 range from small to large.

Table 3 gives the results for which data from one sample were sampled from a standard normal distribution, while the second sample was sampled from a distribution in which the mean was equal to the effect size, because the variance is one, and whose skew and kurtosis were specified in the three conditions illustrated in Figure 1. As can be seen from Table 3, in the first two conditions, when the sample size decreased, so too did the coverage of the noncentral confidence interval procedure. The results from the third condition remained quite valid throughout the range of effect sizes and sample sizes that were studied. This is likely the case because Condition 3 corresponded to a distribution that did not markedly diverge from normality.

With the exception of the effect size of $\delta = 1.60$, for which the small sample size was only eight per group, the bias-corrected and accelerated procedure accomplished its goal of 95% coverage very well in each of the other 12 scenarios. However, when sample size was very small ($n_j = 8$), the procedure was not satisfactory, although it worked about as well as the noncentral method for the first two conditions. When making decisions within the resampling framework, the data should not only be representative, but there should also be an adequate number of independent pieces of information available. When sample size is not small, the bootstrap methodology works very well, especially in situations in which the statistical assumptions are violated. (Note that "not small" does not imply large sample theory. There simply has to be a large enough sample size so that the bootstrap samples are largely unique, because there are few ways the bootstrap replications can differ when sample size is small.) Apparently, when $n = 8$ and data are nonnormal, there are not enough independent pieces of information to adequately represent the distributional form the data follow, and thus confidence interval coverage is itself not adequate.

Notice in Table 3 that the noncentral method yielded statistical power of approximately 0.80 in each of the 15 scenarios. This was the case even though the confidence interval coverage decreased further and further from the nominal value as sample size decreased. Had the confidence interval coverage been the nominal value, power would have likely been smaller. Thus, even though the empirical power nearly equaled the nominal power, the confidence interval coverage tended to be wider than desired. Further notice that the statistical power of the bias-corrected and accelerated method was smaller than that of the noncentral method, albeit not much when sample size was large.

Discussion

Parametric procedures that rely on the often untenable assumption of normality are the dominant procedures used by researchers in the behavioral sciences as well as other disciplines that make use of inferential statistics. However, behavioral phenomena often do not follow a normal distribution (Micceri, 1989), in addition to the homogeneity assumption often being untenable. When the assumptions of parametric tests are violated, the integrity of the results based on parametric statistical techniques is suspect. If conclusions are drawn on the basis of parametric procedures for which the assumptions have likely been violated, they should be interpreted with caution.

Although there are reasons to use parametric methods in certain circumstances, as well as the corresponding nonparametric bootstrap method in other circumstances, it should be realized that each method has its own advantages and disadvantages. Rather than recommending only one method at the expense of another, the recommendation offered here is a moderate approach whereby results are presented from both methodologies.

It is unreasonable to assume that researchers are likely to abandon their elegant parametric procedures for the corresponding nonparametric procedures anytime soon. It is also unreasonable to ignore the advancements and advantages offered by nonparametric statistics, particularly bootstrap methods. The suggestion of performing and reporting the results of both procedures allows the available evidence to be weighed with two very different methodologies. The bootstrap theory says that if the parametric assumptions hold, the results of the BC*a* method will yield results consistent with the parametric results. The theory of the bootstrap methodology also says that when the assumptions of parametric statistical assumptions are false, the BC*a* method will provide a more realistic assessment of the phenomenon under study, provided the data represent a random and representative sample from the population of interest. When the results of the parametric test and the corresponding bootstrap procedure agree, the results can be taken to be very accurate and meaningful. If, however, the procedures yield different results, researchers should ask themselves if the statistical assumptions have likely been violated. If they have, it is more likely that the results obtained from the bias-corrected and accelerated bootstrap procedure are more valid (unless the sample size is too small for the bootstrap replications to produce many unique sets of observations).

The present Monte Carlo study examined the results of the noncentral confidence interval procedure for δ using three nonnormal distributions. Although the assumptions of parametric tests can be met under only one scenario (normality, homogeneity of variance, independent observations), they can be violated in an infinite number of ways. The distributions chosen to illustrate the problems when normality is violated may not represent the distributional forms from some areas of research, and they may not differ from the normal distribution enough to unambiguously show the benefits of the bootstrap procedures. Although this is a limitation of the current study, a study evaluating the effects of violating the assumption of normality (and combinations of violations) can potentially examine an arbitrarily large number of distributions for which several measures are simultaneously evaluated. Another limitation is that sample sizes were always equal across groups. It is known that the *t* test is more robust when sample sizes are nearly equal. Thus, it is likely that the bootstrap method would offer more advantages when parametric assumptions are violated and sample sizes across groups are not equal. Although the present article examines a limited number of situations in which the normality assumption was violated, the simulations were meant to support, not supplant, the theory and rationale of the bootstrap methodology.

The bootstrap approach, although presented for a specific procedure, albeit an important one, is very general and applicable to most parameters and statistical procedures (but see LePage & Billard, 1992, for instances in which the bootstrap method may fail). The assumptions behind the bootstrap approaches are minimal. The bootstrap approaches assume only that the data are a random and representative sample from some larger population.

Although Cohen's *d* is the most common estimate of δ, its absolute value is known to be a positively biased value, particularly when sample sizes are small. Therefore, it is recommended that d_u from equation 10 be used as the point estimate for δ (of course, $d \to d_u \to \delta$ as ν becomes large). It is further recommended that d_u be used in conjunction with the bias-corrected and accelerated procedure when forming a $100(1 - \alpha)\%$ confidence interval around δ. This bootstrap confidence interval can be used alone or in addition to the noncentral method. Using the methodology developed within the bootstrap framework can assist researchers who are interested in forming confidence bounds for some statistic, an effect size for example, but who are not comfortable or willing to base their conclusions on assumptions that they realize may be untenable.

Appendix A
Obtaining Confidence Limits for the Percentile and the Bias-Corrected and Accelerated Methods in R or S-Plus

The syntax given below provides a method to obtain confidence limits for the percentile and the bias-corrected and accelerated bootstrap methods. (Note that the bias-corrected and accelerated method is recommended for use over the percentile method; the percentile method is provided for completeness and is not necessarily recommended for applied applications.) The point estimate of δ that the syntax uses is the unbiased estimate, d_u, which is given in equation 10. Note that this syntax relies on the `Cohens.d` and `Unbiased.d` functions, which are given in Appendix B.

```
B <- 1000 # Number of bootstrap samples/replications.
alpha <- .05 # Type I error rate (or 1-alpha as confidence interval
       coverage)

Group.1 <- DATA VECTOR OF GROUP 1'S SCORES
Group.2 <- DATA VECTOR OF GROUP 2'S SCORES

n.1 <- length(Group.1)
n.2 <- length(Group.2)

Bootstrap.Results <- matrix(NA, B, 1)
for(b in 1:B)
{
```

```
Bootstrap.Results[b,1] <- Unbiased.d(sample(Group.1, size=n.1,
      replace=T),sample(Group.2,size=n.2, replace=T))
}

Jackknife.Results <- matrix(NA,n.1+n.2,1)
Marker.1 <- seq(1, n.1, 1)
for(sample.1 in 1:n.1)
{
Jackknife.Results[sample.1, 1] <- Unbiased.d(Group.1[Marker.1
      [-sample.1]],Group.2)
}

Marker.2 <- seq(1, n.2, 1)
for(sample.2 in 1:n.2)
{
Jackknife.Results[n.1+sample.2, 1] <- Unbiased.d(Group.1,
      Group.2[Marker.2[-sample.2]])
}
Mean.Jackknife <- mean(Jackknife.Results)
a <- (sum((Mean.Jackknife-Jackknife.Results)^3))/
      (6*sum((Mean.Jackknife-Jackknife.Results)^2)^(3/2))
z0 <- qnorm(sum(Bootstrap.Results Unbiased.d(Group.1, Group.2))/B)

CI.Low.BCa <-  pnorm(z0 + (z0+qnorm(alpha/2))/(1-a*(z0+qnorm(alpha/
      2))))
CI.Up.BCa <- pnorm(z0 + (z0+qnorm(1-alpha/2))/(1-a*(z0+qnorm(1-
      alpha/2))))
LINE SPACE
Percentile.Confidence.Limits <- c(quantile(Bootstrap.Results,
      alpha/2), quantile(Bootstrap.Results, 1-alpha/2))
BCa.Confidence.Limits <- c(quantile(Bootstrap.Results, CI.Low.BCa),
      quantile(Bootstrap.Results, CI.Up.BCa))
LINE SPACE
# Below are the confidence limits for the bootstrap Percentile
      method and the BCa method for the unbiased estimate of d.
Percentile.Confidence.Limits
BCa.Confidence.Limits
```

Appendix B
Obtaining an Unbiased Estimate of δ in R or S-Plus

The syntax given below provides functions to estimate δ with Cohen's *d* (from equation 2) and with the unbiased estimate, d_u (from equation 10). These functions are necessary to carry out the bootstrap procedures for obtaining confidence intervals given in Appendix A.

```
Cohens.d <- function(Group.1, Group.2)
{
n.1 <- length(Group.1)
n.2 <- length(Group.2)
SS1 <- var(Group.1)*(n.1-1)
SS2 <- var(Group.2)*(n.2-1)
pooled.sd <- sqrt((SS1 + SS2)/(n.1+n.2-2))
Result <- (mean(Group.1)-mean(Group.2))/pooled.sd
Result
}

Unbiased.d <- function(Group.1, Group.2)
{
nu <- length(Group.1)+length(Group.2)-2
G.nu <- gamma(nu/2)/(sqrt(nu/2)*gamma((nu-1)/2))
d <- Cohens.d(Group.1, Group.2)
Result <- ifelse(nu > 171, d, d*G.nu)
# Because of limitations of the S language, the gamma function
     cannot be applied to degrees of
# freedom greater than 171. When such a case arises, Cohen's d is
     used rather than the unbiased d.
# This does not pose a practical problem because the differences
     at such large degrees of freedom are trivial.
Result
}
```

References

Boneau, A. C. (1960). The effects of violations of the assumptions underlying the *t* test. *Psychological Bulletin*, *57*, 49-64.

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

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.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika*, *43*, 521-532.

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

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., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

LePage, R., & Billard, L. (Eds.). (1992). *Exploring the limits of the bootstrap*. New York: John Wiley.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*, 156-166.

Olejnik, S., & Algina, J. (2000). Measures of effect size for comparative studies: Applications, interpretations, and limitations. *Contemporary Educational Psychology*, *25*, 241-286.

Roberts, J. K., & Henson, R. K. (2002). Correction for bias in estimating effect sizes. *Educational and Psychological Measurement*, *62*, 241-253.

Sawilowsky, S. S., & Blair, R. (1992). A more realistic look at the robustness and type II error properties of the *t* test to departures from population normality. *Psychological Bulletin*, *111*, 352-360.

Siegel, S. (1956). *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill.

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

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

Stuart, A., & Ord, J. K. (1994). *Kendall's advanced theory of statistics: Distribution theory* (Vol. 1, 6th ed.). New York: John Wiley.

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

Wilcox, R. R. (1998). How many discoveries have been lost by ignoring modern statistical methods? *American Psychologist*, *53*, 300-314.

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