Journal of Modern Applied Statistical Methods

Volume 7 | Issue 1 Article 6

5-1-2008

Coverage Performance of the Non-Central F-based and Percentile Bootstrap Confidence Intervals for Root Mean Square Standardized Effect Size in One-Way Fixed-Effects ANOVA

Guili Zhang
East Carolina University, zhangg@ecu.edu

James Algina
University of Florida, algina@ufl.edu

Journal of Modern Applied Statistical Methods
May, 2008, Vol. 7, No. 1, 56-76

1538 – 9472/08/$95.00

Coverage Performance of the Non-Central *F*-based and Percentile Bootstrap Confidence Intervals for Root Mean Square Standardized Effect Size in One-Way Fixed-Effects ANOVA

Guili Zhang
East Carolina University

James Algina
University of Florida

The coverage performance of the confidence intervals (CIs) for the Root Mean Square Standardized Effect Size (RMSSE) was investigated in a balanced, one-way, fixed-effects, between-subjects ANOVA design. The noncentral *F* distribution-based and the percentile bootstrap CI construction methods were compared. The results indicated that the coverage probabilities of the CIs for RMSSE were not adequate.

Key words: confidence interval, effect size, ANOVA, root mean square standardized effect size, non-central *F*-distribution, percentile bootstrap, coverage probability, robustness.

Introduction

Reporting an effect size (ES) in addition to or in place of a hypothesis test has been recommended by some statistical methodologists since as early as the 1960s because ESs are recognized as being more appropriate and more informative (Cohen, 1965, 1994; Cumming & Finch, 2005; Finch et al., 2002; Hays, 1963; Meehl, 1967; Nickerson, 2000; Steiger, 2004; Steiger & Fouladi, 1997). In the last two decades, reporting an ES has become mandatory in some editorial policies (Murphy, 1997; Thompson, 1994) and is strongly recommended for American Psychological Association journals. The Publication Manual of the American Psychological Association (2001) states that it is almost always necessary to include some index of ES or strength of relationship in the results section of a research paper. The APA Task Force on Statistical Inference (Wilkinson and the Task Force on Statistical Inference, 1999) also supports the report of ESs as well as the obligation of researchers to provide confidence intervals (CI) for all principal outcomes. A CI for an ES is recommended as a superior replacement for significance testing because this CI contains all the information found in the significance tests and vital information not provided by the significance tests about the magnitude of effects and precision of estimates (Cohen, 1994; Cumming & Finch, 2001, 2005). A CI indicates the range of population ESs with which the data are consistent. By contrast, a hypothesis test merely indicates whether the data are consistent with a population ES of zero. Because of the obvious advantages of CIs, advocate on the use of ESs and CIs for ESs is "a rapidly rising tide" (Grissom & Kim, 2005).

Dr. Guili Zhang is Assistant Professor of Research and Evaluation Methodology in the Department of Curriculum and Instruction. Her research interests are in applied statistics. Email her at: zhangg@ecu.edu. Dr. James Algina is Professor of Educational Psychology. His research interests are in psychometric theory and applied statistics. Email him at: algina@ufl.edu.

Effect Size Indices and Confidence Intervals in the Two-Group Case

A large number of ES indices have been developed and proposed (Algina et al., 2005a). For example, the number of commonly used ESs measuring separation of two independent samples alone has almost reached a dozen: Cohen's *d* (Cohen, 1965), Glass's *d*, Hedges' *g* (Hedges & Olkin, 1985), two versions of

Cohens d based on trimmed means and Winsorized variances ($d_R^{\dagger}$ suggested by Hogarty & Kromrey, 2001 and d_R suggested by Algina and Keselman 2003b; Algina et al., 2005a), eta squared, omega squared, McGraw and Wong's (1992) common language ES (CL), Cliff's dominance statistic (1993, 1996), Kraemer and Andrews γ_1^* (1982), Wilcox and Muska's W (1999), and Vargha and Delaney's A (2000).

Research investigating the performance of the various ES measures is fairly limited. Hedges and Olkin (1985) suggested that Cohen's d evidenced a small sample bias. Hogarty and Kromrey (2001) compared the performance of nine ES indices when they were used in the context of populations with various levels of nonnormality and variance heterogeneity. The nine indices included Cohen's d, Cliff's dominance statistic, g, γ_1^*, CL, A, $d_R^{\dagger}$, a naïve estimator of W and a .632 bootstrap estimator of W. The results indicated that Cohen's d and Hedges' g showed nontrivial sensitivity to violations of normality and homogeneity of variance, which confirmed the concerns raised about the appropriateness of using these indices as indicators of effects in such populations (Kraemer & Andrews, 1982; Wilcox & Muska, 1999). In addition, $d_R^{\dagger}$ evidenced severe bias under small sample conditions. Indices CL, γ_1^* and the naïve estimator of W only appeared to be slightly less sensitive than Cohen's d and Hedges' g, but showed pronounced bias under small sample size condition or nontrivial sensitivity to violations of normality and homogeneity of variance. Cliff's dominance statistic and Vargha and Delaney's A showed better performances in producing relatively unbiased estimates and consistent standard errors.

Hess and Kromrey (2004) investigated the performance of the CIs for Cohen's d and Cliff's dominance statistic constructed by using seven CI construction methods: the normal theory Z band, the percentile bootstrap, the bias corrected bootstrap, the bias corrected and accelerated bootstrap (BCa), pivotal, Studentized pivotal, and the Steiger and Fouladi interval inversion band. Monte Carlo methods were used to compare CI estimates using random samples generated from populations under known and controlled conditions. Across all of the conditions, all of the CI construction methods provided better coverage probabilities for Cliff's dominance statistic than for Cohen's d, with the exception of the Pivotal Bootstrap method.

Cohen's d and Its Confidence Intervals

In the two-group independent samples case, Cohen's d is probably the most widely accepted ES index for a pairwise contrast on means and it is defined as follows:

$$d = \frac{\bar{Y}_2 - \bar{Y}_1}{S} \tag{1}$$

where $\bar{Y}_j$ is the mean for the jth level (j = 1, 2), and S is the square root of the pooled variance. The number of observations in a level is denoted by n_j. Cohen's d estimates:

$$\delta = \frac{\mu_2 - \mu_1}{\sigma} \tag{2}$$

where μ_j is the population mean for the jth (j=1,2) level, and σ is the population standard deviation, assumed to be equal for both levels.

Reporting a CI for the ES is important as was well put by Wilkinson et al. (1999), "it is hard to imagine a situation in which reporting a dichotomous reject-accept decision is better than reporting an actual p value or, better still, a confidence interval" (p. 599). Steiger and Fouladi (1997) asserted that "a confidence interval conveys more information, in a more naturally usable form, than a significance test." Interests in the accuracy and usefulness of the ESs have motivated explorations of the usefulness and effectiveness of CIs for ESs (Algina & Keselman, 2003a, 2003b; Bird, 2002; Cumming & Fitch, 2001).

An exact CI for δ can be obtained by using the noncentral t distribution when the sample data are normally distributed, the two population have equal variances, and the scores are independently distributed (Algina et al., 2005a; Cumming & Fitch, 2001; Johnson & Welch, 1940; Serlin & Lapsley, 1985; Steiger & Fouladi, 1997). This CI is the same CI that Hess & Kromrey (2004) referred to as the Steiger and Fouladi inversion method. In this situation, the noncentral *t* distribution has two parameters: the degrees of freedom, and the noncentrality parameter λ, which is given by

$$\lambda = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \left(\frac{\mu_2 - \mu_1}{\sigma} \right) = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \delta. \quad (3)$$

To find a 95% CI for δ, we first use the noncentral *t* distribution to find a 95% CI for λ, then multiply the two end points of the interval for λ by $\sqrt{(n_1 + n_2) / n_1 n_2}$ to obtain the two end points of a 95% CI for δ. The lower limit of the 95% CI for λ is the noncentrality parameter for the noncentral *t* distribution in which the calculated *t* statistic

$$t = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \left(\frac{\bar{Y}_2 - \bar{Y}_1}{S} \right) \quad (4)$$

is the .975 quantile, and the upper limit of the 95% CI for λ is the noncentrality parameter for the noncentral *t* distribution in which the calculated *t* statistic is the .025 quantile of the distribution (Algina et. al., 2005a, 2006; Steiger & Fouladi, 1997). Algina and Kesleman (2003a) adapted this procedure for the dependent samples case.

As noted previously when the population data are normally distributed, the two population have equal variances, and the scores are independently distributed, the noncentral *t* distribution-based CI is exact. However, when sampling from nonnormal data, the noncentral *t* distribution-based CI may not have adequate coverage probability in both the independent samples case (Algina & Keselman, 2003a; Kelley, 2005) and dependent samples case (Algina et al., 2005a). Failure to have adequate coverage probability means, for example, that if a nominal 95% CI for δ is computed, the actual probability that the CI contains the parameter will be different than .95.

Kelley (2005) compared three methods for constructing a CI around Cohen's ES. Specifically, he evaluated noncentral *t* distribution-based, the percentile bootstrap, and the BCa CIs through a set of simulation studies that involves three conditions of nonnormality, three cases of sample size, and six cases of population ES. Kelley's study indicated that the noncentral *t* distribution-based CI has inaccurate coverage probability when data are nonnormal. He concluded that when the assumptions of parametric tests are violated, the integrity of the results based on parametric statistical techniques is suspect. The study by Algina et al. (2006) detected the same problem with the noncentral *t* distribution-based CI in the dependent samples case. In addition, the results from the Hess and Kromrey (2003, 2004) studies also pointed to the inadequate coverage probability issue with the CIs for Cohen's *d*.

Results from recent studies indicated that in the two-group case, the bootstrap CI is preferable and should be used instead of the noncentral *t* distribution-based CI. Kelley (2005) asserted that when the normality assumption is false, a CI constructed with the BCa method is more valid than the noncentral *t* distribution-based CI. When the normality assumption holds, the BCa method will yield results consistent with the parametric results. Therefore, he recommends the use of the BCa method. Like Kelley, Algina et al. (2006) also found that under many conditions the BCa method worked best, although in some cases of data nonnormality, the BCa method did not control probability coverage. By including a wider range of nonnormality than was investigated by Kelly, they found that the BCa method for setting a CI around the population ES is indeed negatively affected by nonnormality. Additionally, they found that the coverage probability declines as sample size decreases and the population ES increases. It is apparent that even with the nonparametric bootstrap construction methods, problem still persists with CIs for Cohen's δ.

The work reported by Algina and Keselman (2003b), Algina et al. (in press, 2005a), and Kelly (2005) indicated that in both the independent samples and dependent samples cases, CIs for Cohen's δ may be misleading because of poor coverage probability when data are nonnormal. There is a second problem with using Cohen's δ: although it is intended as a measure of group separation, it is not always an adequate measure of group separation. This shortcoming was pointed out by Wilcox and Keselman (2003), and is due to the fact that δ can be dramatically affected by outliers and long-tailed distributions. Cohen's δ is defined by using the usual population means and variances, both of which are least-square parameters. Least-square parameters are not robust, meaning that a small change in the population distribution can strongly affect the parameters. In particular, the usual population mean and variance can be greatly influenced by the existence of extreme observations (outliers) in a distribution. Slight changes in the population distributions, changes that do not have much effect on the separation of the distributions, can substantially alter the value of δ. Therefore, δ can be a very poor measure of group separation, and can grossly misrepresent the degree to which two distributions differ (Algina et al., 2005b; Wilcox & Keselman, 2003).

Root Mean Square Standardized Effect Size and Its Confidence Intervals

Measures of ES in analysis of variance (ANOVA) are measures of the degree of association between a factor and the dependent variable. When it comes to the one-way, fixed-effects, between-subjects ANOVA case, the available generalized ES measures are, but not limited to, eta squared, omega squared, $d_{\max}$, Cohen's *f,* and the Mean Square Standardized Effect Size (RMSSE) (Olejnik & Algina, 2003; Steiger & Fouladi, 1997). Eta squared and omega squared are estimates of the degree of association. Eta squared is the proportion of the total sum of squares that is attributed to an effect. It is calculated as the ratio of the effect variance to the total variance. Omega squared is an estimate of the dependent variable variance accounted for by the independent variable in the population for a fixed-effects model. The effect size $d_{\max}$ is an overall ES that is calculated by utilizing the smallest and the largest means where $d_{\max} = \dfrac{\bar{Y}_{\max} - \bar{Y}_{\min}}{S}$ (Cohen, 1988), while Cohen's *f* and RMSSE are overall ESs that use all of the means and are measures of the standardized average effect in the population across all of the levels of the independent variable. Among these ES measures, the RMSSE, proposed by Steiger and Fouladi (1997), denoted by f^* in our study, was part of the focus of our investigation. RMSSE is a standardized mean difference measure, a generalization of Cohen's δ, and a variant of Cohen's *f*.

In a balanced, one-way, between-subjects, fixed-effects design, f^* is defined by Steiger & Fouladi (1997) as follows:

$$f^* = \sqrt{\frac{\sum_{j=1}^{J}(\mu_j - \mu)^2}{(J-1)\sigma^2}} \qquad (5)$$

where μ_j is the mean for the j th level, μ is the grand mean, and σ^2 is the within-level variance, which is assumed to be constant across levels. Recall that Cohen (1969) defined $f = \sqrt{\sum_{j=1}^{J}\left(\mu_j - \mu\right)^2 \Big/ J\left(n-1\right)\sigma^2}$,

so f^* is a variation of Cohen's f.

Consider a one-way, fixed-effects ANOVA with n_j observations in the *j*th group, and J groups. The F statistic is calculated by using

$$F = \frac{MS_B}{MS_W} \qquad (6)$$

where

$$MS_B = \frac{\sum_{j=1}^{J} n_j(\bar{Y}_j - \bar{Y})^2}{J-1} \qquad (7)$$

and

$$MS_W = \frac{\sum_{j=1}^{J}\sum_{i=1}^{n_i}(Y_{ij} - \bar{Y}_j)^2}{N-J} \qquad (8)$$

In Equation 7 and 8, Y_{ij} is the ith score in group j, $\bar{Y}_j$ is the sample mean for the group j, and $\bar{Y}$ is the grand mean and is calculated by using

$$\bar{Y} = \frac{\sum_{i=1}^{n_i}\sum_{j=1}^{J} Y_{ij}}{N} \quad (9)$$

Based on expected mean squares in a balanced design, f^* can be estimated by using

$$\hat{f}^* = \sqrt{\frac{MS_B - MS_W}{nMS_W}} = \sqrt{\frac{1}{n}\left(F-1\right)} \quad (10)$$

if $F \geq 1$ and by using $\hat{f}^* = 0$, otherwise. Alternatively, based on the expected value of F under normality f^* can be estimated by using

$$\hat{f}^* = \sqrt{\frac{\left(N-J-2\right)}{n\left(N-J\right)}\left(F-1\right)} \quad (11)$$

if $F \geq 1$ and $\hat{f}^* = 0$ otherwise. Both estimates are very similar but the estimate in Equation 10 was used in our study because it does not require the normality assumption in its derivation.

The CIs for Steiger and Fouladi's f^* can be constructed based on the noncentral F distribution (Steiger and Fouladi, 1997; Steiger, 2004). In a one-way, between-subjects, fixed-effects ANOVA, the F statistic with $J-1$ and $N-J$ degrees of freedom has noncentrality parameter

$$\lambda = \frac{\sum_{j=1}^{J} n_j(\mu_j - \mu)^2}{\sigma^2} \quad (12)$$

Clearly in a balanced design

$$f^* = \sqrt{\frac{\lambda}{n\left(J-1\right)}} \quad (13)$$

To find a 100(1 - α)% (95% in our study) CI for f^*, we first use the noncentral F distribution to find a 95% CI for λ. Once the CI on λ is found, we transform the endpoints of the CI for λ by dividing λ by $(J-1)n$ and then take the square root. The result is an exact CI for f^* in the analysis of variance, when its assumptions are met. The lower limit of the 95% CI for λ is the noncentrality parameter for the noncentral F distribution in which the calculated F statistic is the .975 quantile. The upper limit of the 95% CI for λ is the noncentrality parameter for the noncentral F distribution in which the calculated F statistic is the .025 quantile of the distribution.

Purposes of the Study

Constructing a CI for RMSSE by using the noncentral F distribution is based on the assumption that the data are drawn from normal distributions. If data are not normally distributed, the actual coverage probability of the CI may or may not match the nominal level. A method that may be useful for constructing CI for f^* is the percentile bootstrap (Efron and Tibshirani, 1993). Therefore, the performance of the percentile bootstrap on the construction of CIs for f^* was examined in our current study. The purpose of the study is to investigate the coverage performance of the noncental F distribution-based and the percentile bootstrap CI for f^*.

Methodology

The noncentral F distribution-based and the percentile bootstrap CIs were implemented for all combinations of the following five factors: (a) five population distributions including the normal distribution and four additional cases from the family of the g and h distributions that are nonnormal (Hoaglin, 1983, Martinez & Iglewicz, 1984); (b) two numbers of levels for treatment groups: J = 3 and J = 6; (c) three cell sample sizes in each treatment; (d) six values of population RMSSEs; (e) two mean configurations: the equally spaced mean configuration and the one extreme mean configuration. The nominal confidence level for all intervals investigated was .95 and each condition was replicated 2500 times. The number of bootstrap replications in the bootstrap procedure was 1000.

Conditions

Data for all five distributions were generated from the g and h distributions: (a) $g = h = 0$, the standard normal distribution ($\gamma_1 = \gamma_2 = 0$), where $\gamma_1 = \sqrt{\beta_1}$ and is the skewness, and $\gamma_2 = \beta_2$ and is the kurtosis, (b) $g = .76$ and $h = -.098$, a distribution with the skewness and kurtosis of an exponential distribution ($\gamma_1 = 2$, $\gamma_2 = 3$), (c) $g = 0$ and $h = .225$ ($\gamma_1 = 0$ and $\gamma_2 = 154.84$), (d) $g = h = .225$ ($\gamma_1 = 4.90$ and $\gamma_2 = 4673.80$), and (e) $g = 0$ and $h = .109$ ($\gamma_1 = 0$ and $\gamma_2 = 6$), a distribution with the skewness and kurtosis of a double exponential distribution. The four nonnormal distributions cover a wide range of nonnormality including distributions that are quite strongly nonnormal. Such a selection of distributions allows the researcher to investigate the performances of the CIs under a wide range of data conditions. The goal is to find which procedure or procedures are likely to work well over a wide range of distributions because it is impossible for any one simulation to include every possible distribution that might be encountered in real data or to anticipate what types of distributions are realistic in all of social and behavioral science fields.

The numbers of treatment groups investigated were 3 and 6, which cover the likely range encountered in most research in the social and behavioral sciences. The sample sizes in each treatment included were 20, 35, and 50. Such a range seems fairly typical of sample sizes used in social science research, although clearly does not cover sample sizes found in very small or very large studies.

The treatment group means followed two mean configurations: the equally spaced mean configuration and the one extreme mean configuration, which will allow determination of whether results tend to generalize over configurations.

Six values of f^* were investigated: 0, .1, .25, .40, .55, and .70. Defining

$$\delta_{\max} = \frac{\mu_{\max} - \mu_{\min}}{\sigma} \tag{14}$$

as Cohen's effect size for the largest and smallest means, under the equally spaced mean configurations, these population f^* values approximately correspond to $\delta_{\max}$ of 0, .2, .5, .8, 1.10, and 1.40, respectively. Under the one extreme mean configuration, these population f^* values roughly correspond to $\delta_{\max}$ of 0, .173, .433, .693, .952, and 1.212. Therefore, a f^* of .0 indicates no effect, .1 a small effect, .25 a medium effect, .40 a large effect, and .55 and .70 very large effect**s**.

The nominal confidence level for all intervals investigated was .95 and each condition was replicated 2500 times, assuring sufficient precision for an adequate initial investigation into the sampling behaviors of the CIs. The number of bootstrap replications in the bootstrap procedure was 1000.

Analyses Conducted

The study was designed to investigate the robustness of the noncentral F distribution-based CIs and the percentile bootstrap CIs for f^* to sampling from nonnormal distributions.

Variables conforming to a g and h distributions are transformations of a standard normal distribution. When g and h are both nonzero,

$$Y = \frac{\exp(gZ) - 1}{g} \exp\left(\frac{hZ^2}{2}\right) \tag{15}$$

where Z is a standard normal variable, and Y is the g and h distributed variable. When g is zero,

$$Y = Z \exp\left(\frac{hZ^2}{2}\right) \tag{16}$$

Standard normal variables (Z_{ij}) were generated by using RANNOR function in SAS (SAS, 1999). Then the Z_{ij} were converted to the desired g and h distributed random variable by using Equation 15 and 16. To create scores corresponding to the selected values of f^*, it is necessary to linearly transform the g and h distributed variables. Data were generated for three samples and six samples in each replication of each condition by the following steps: First, for the first sample n_1 scores were generated from the appropriate distribution. Then n_2 scores from the same distribution were generated and a constant was added to each score. Thirdly, n_3 scores from the same distribution were generated and a constant was added to each score and so forth until n_J scores from the same distribution were generated and a constant was added to each score. The constants were chosen such that the population RMSSE, f^* would equal to the following values: 0, .1, .25, .40, .55, and .70.

For the equally spaced mean configuration, the addition of the constant was accomplished by using

$$Y_{ij} = X_{ij} + (j-1)\sqrt{\frac{12}{J(J+1)}} f^* \sigma, \quad j = 1, \ldots, J. \tag{17}$$

For the configuration with one extreme mean, $Y_{ij} = X_{ij}$ for groups $j = 1, \ldots, J-1$. For group J the transformation was

$$Y_{ij} = X_{ij} + \sqrt{J} f^* \sigma. \tag{18}$$

To obtain a $(1-\alpha)$% (95% in the current study) CI for f^*, the noncentral F distribution is first used to obtain a 95% CI on λ, the noncentrality parameter of the F distribution. Given an observed F statistic with a value F and known degrees of freedoms, a $(1-\alpha)$% CI on λ can be obtained with the following steps (Steiger, 2004):

1. Calculate the cumulative probability of the value F in the central F distribution. This is 1 - p, where p is the probability level printed by most analysis of variance procedures. If 1 - p is below $\alpha/2$, then both limits of the CI are zero. If 1 - p is below $1-\alpha/2$, the lower limit of the CI is zero, and the upper limit must be calculated (go to step 3). Otherwise, calculate both limits of the CI for λ by using steps 2 and 3.
2. To calculate the lower limit of λ, find the unique value of λ that places the F statistic at the 1 - $\alpha/2$ probability point of a noncentral F distribution with the known degrees of freedom.
3. To calculate the upper limit of λ, find the unique value of λ that places the F statistic at the $\alpha/2$ cumulative probability point percentile of a noncentral F distribution.

In summary, calculating a CI for λ requires iterative calculation of the unique value of λ that places an observed value of F at a particular percentile of the noncentral F distribution. These procedures were implemented by using the "FNONCT" function in SAS. Notice the CI for f^* constructed by the noncentral F distribution-based method will result in coverage probability of .975 when $f^* = 0$ because the probability noncoverage from the lower side of the distribution will be 0 instead of .025.

Once the CI on λ is found, the endpoints of the CI for λ are transformed to endpoints for f^* by dividing by $(J-1)n$ and then taking the square root. The result is an exact CI for f^* in the analysis of variance, when the ANOVA assumptions are met.

To apply the percentile bootstrap method, the following steps are completed 1000 times within each replication of a condition.

1. A sample of size n_j is randomly selected with replacement from the scores for the group j, $j = 1, \ldots, J$.

These J samples are combined to form a bootstrap sample.

2. The parameter f^{*2} is estimated by using

$$\hat{f}^{*2} = \frac{1}{n}(F-1) \qquad (19)$$

1. The 1000 $\hat{f}^{*2}$ estimates are then ranked from low to high. The lower limit of the CI for f^{*2} is determined by finding the 26th estimate in the rank order [i.e., the (.025x1000+1)th estimate]; and the 975th estimate is the upper limit of the CI for f^{*} [i.e., the (.975x1000)th estimate].
2. The lower limit of the CI for f^{*} is equal to the square root of the lower limit of the CI for f^{*2} if the latter lower limit is larger than zero and is zero otherwise. The upper limit of the CI for f^{*} is equal to the square root of the upper limit of the CI for f^{*2}.

Results

The estimated coverage probabilities for and the average widths of the noncentral F distribution-based and bootstrap CIs for f^{*} are reported and compared for all conditions. The estimated coverage probabilities of the noncentral F distribution-based and bootstrap CIs for f^{*} are reported in Table 1 through Table 4. The average widths of the noncentral F distribution-based and bootstrap CIs for f^{*} are shown in Table 5 through Table 8.

Estimated Coverage Probabilities of Confidence Intervals for f^{*}

The interval [.925, .975] used by Algina et al. (2006) was used as a criterion for adequate coverage probability when the nominal confidence coefficient is .95. This interval corresponds to Bradley's (1978) liberal criterion for a nominal .05 Type I error rate. In addition, because this interval may be considered as too lenient, a more stringent interval, [.94, .96], was also used to judge the adequacy of the coverage probabilities. In Tables 1 through 4, estimates that are outside the [.94, .96] interval are bolded, while estimates that are outside of the interval [.925, .975] are bolded and underlined.

The patterns of results across Tables 1 to 4 for the noncentral F distribution-based CI for f^{*} are fairly similar. First, when sampling from a normal distribution, as stated earlier, the coverage probability of the noncentral F distribution-based CI should be .975 when $f^{*} = 0$, and the results in Tables 1 to 4 are consistent with the theory. When $f^{*} > 0$, the coverage probability of the noncentral F distribution-based CI is expected to be .95 under normality and the results in Tables 1 to 4 are consistent with this expectation.

Second, coverage probability for the noncentral F distribution-based CI tends to be better than for the bootstrap CI both when sampling from normal and nonnormal distributions. When $J = 3$ and samples are drawn from a normal distribution, coverage probability for the noncentral F distribution-based CI is outside [.925, .975] in 2 out of 36 total cases, while the bootstrap CI coverage probability is outside [.925, .975] in 13 cases. Under normality, when $J = 6$, although both CIs have 2 coverage probabilities that are outside [.925, .975], the noncentral F distribution-based CI has 6 coverage probabilities that are outside [.94, .96] while the bootstrap CI has 18 coverage probabilities that are outside this interval. When sampling from the nonnormal distributions, the noncentral F distribution-based CI has fewer coverage probabilities that are outside the criterion intervals than does the bootstrap CI under each of the four distribution conditions.

Third, the performances of the noncentral F distribution-based CIs for f^{*} under the four nonnormal distributions reveal some common characteristics across levels of J and types of mean configuration. When $f^{*} = 0$, coverage probability tends to be outside [.925, .975]. When $f^{*} = .10$, coverage probabilities of the noncentral F distribution-based CI for f^{*} are all inside the [.94, .96] interval. Coverage probability tends to be inside either the [.925, .975] interval or both intervals in most conditions when $f^{*} = .25$ with exceptions occurring principally when data are sampled from the g = .225 and h = .225 distribution.

Coverage probability of the noncentral F distribution-based CI for f^* tends to be inside either the [.925, .975] interval or both intervals in most conditions when $f^* \geq .40$ and g = .000, and h = .109. Coverage probability is outside [.925, .975] for only a few cases, all with J = 6. Coverage probabilities are mostly outside the [.925, .975] interval when $f^* \geq .40$ for the nonnormal distributions other than the g = .000, and h = .109 distribution.

Excluding $f^* = 0$, the coverage probability performance of the noncentral F distribution-based CI tends to decline as f^* increases, as the distributions become more long-tailed, and appears to be worse for skewed distributions. Overall, when data are sampled from the g = 0 and h = .109 distribution more estimates are within the [.925, .975] interval than when data are sampled from the three other nonnormal distributions.

The results of the bootstrap CIs for f^* in Tables 1 to 4 are also fairly similar across levels of J and mean configurations. First, when sampling from the normal distribution, when f^* = 0 and J = 3, the coverage probabilities of the bootstrap CI for f^* are all above .975. When f^* = 0 and J = 6, however, they are all inside [.94, .96]. When f^* = .10, the coverage probabilities of the bootstrap CI for f^* are all outside [.925, .975] when J = 3 and inside [.94, .96] when J = 6. When $f^* \geq$ .25, coverage probability tends to be inside either [.925, .975] or both intervals.

The coverage probabilities for the bootstrap CI for f^* under non-normality also have some common features across the mean configurations. When f^* = 0, the coverage probabilities of the bootstrap CIs for f^* tend to be outside [.925, .975] when J = 3 and inside [.94, .96] when J = 6. When f^* = .10, the coverage probability of the bootstrap CI for f^* tends to be inside [.925, .975] when J = 6 except when n = 20 and the mean configuration is equally spaced. Moreover, when f^* = .10, the coverage probabilities are mostly inside the [.925, .975] when J = 3 with exceptions occurring primarily, but not exclusively, when g = 0 and h = .109.

Coverage probability tends to be inside either the [.925, .975] interval or both intervals in most conditions when f^* = .25 and J = 3. When f^* = .25 and J = 6, more than half of the coverage probabilities are within the [.925, .975] interval. However, under the g = 0 and h = .225 and g = .225 and h = .225 data distributions, they are all outside this interval.

Coverage probability of the bootstrap CI for f^* tends to be inside either the [.925, .975] or both intervals in most conditions when $f^* \geq .40$ for the g = 0 and h = .109 distribution when J = 3. However, they have a tendency to be outside the [.925, .975] interval when J = 6, especially for the one extreme mean configuration. Coverage probabilities of the bootstrap CI for f^* are mostly outside the [.925, .975] interval when $f^* \geq .40$ for the nonnormal distributions other than g = .760 and h = −.098. Exceptions occur principally when $f^* = .40$, J = 3, and g = .760, h = −.098 under larger sample sizes (n = 35 or 50).

Excluding $f^* = 0$, the coverage probability performance of the bootstrap CI tends to decline as f^* increases, and as the distributions become more long-tailed. As f^* increases, the coverage probability of the bootstrap CI for f^* appears to be worse when J = 6 than when J = 3. The coverage probability for the bootstrap CI for f^* tends to be poorer than for the noncentral F distribution-based CI both when sampling from normal and nonnormal distributions.

Average Widths of Confidence Intervals for f^*

The average widths of the noncentral F distribution-based and bootstrap CIs for f^* under J = 3 and the equally spaced mean configuration are presented in Table 5. It is observed that generally the average widths of the noncentral F distribution-based CIs are shorter than those of the bootstrap CIs. The difference between the widths of the two CIs becomes smaller as sample size increases. Furthermore, the average width of both type of CIs gets narrower as the sample size increases and the population effect size f^* decreases. Holding f^* and sample size constant, across data distributions, there is very little difference in the width of the noncentral F distribution-based CIs, and there is also very little difference in that of the bootstrap CIs. Presented in Table 6, the average widths of the CIs for f^* under J = 3 and the one extreme mean configuration shows little difference from those for the equally spaced mean configuration. This suggests that the type of mean configuration does not substantially affect the width of the CIs and therefore to the precision with which f^* is estimated.

Table 7 shows the average widths of the CIs for f^* under J = 6 and the equally spaced mean configuration. It is quite obvious that, when J increases from 3 to 6, the intervals become narrower for all of the combinations of conditions. It is also observed that generally the average widths of the noncentral F distribution-based CIs are shorter than those of the bootstrap CIs. The difference between the widths of the two CIs gets smaller as the sample size increases. In addition, the average widths of both CIs get narrower as the sample size increases and the population ES f^* decreases. Across distributions, there is very little difference in the width of the noncentral F distribution-based CIs and there is also very little difference in that of the bootstrap CIs. The average widths of the CIs for f^* under J = 6 and the one extreme mean configuration are presented in Table 8. Again there is little difference between these widths and the widths from those occur for the equally spaced mean configuration, in terms of values as well as patterns observed. This again suggests that the type of mean configuration does not affect the accuracy with which f^* is estimated.

Conclusion

Confidence intervals for the ES have been strongly advocated by statistical methodologists to be used as a useful supplement to and maybe even a superior replacement for the traditional hypothesis testing. However, much investigation is needed to evaluate the robustness of the CIs in order to ensure their proper usage.

In the two group case, it has been reported that in both the independent samples and dependent samples case CIs for Cohen's δ may be misleading because of poor coverage probability when data are nonnormal (Algina & Keselman, 2003b; Algina et al., 2005a, 2006; Kelly, 2005). It has been further reported that the CIs for δ_R, a robust version of δ, have better coverage probability than do CIs for Cohen's δ and that the coverage probability is closer to the nominal level for the percentile bootstrap CIs than for the noncentral t distribution-based CIs (Algina & Keselman, 2003b).

Our study investigated the robustness of the CIs for RMSSE (f^*), in a one-way, fixed-effects, between-subjects ANOVA. The results indicated that the coverage probabilities of the CIs for f^* were not adequate. Under J = 3, the probability coverage of the CIs for f^* was acceptable only for (a) CIs constructed by using the noncentral F distribution-based method when data were sampled from the normal distribution and from the g = .000 and h = .109 distribution, and (b) CIs constructed by using the percentile bootstrap under normality when the population f^* was small (< .25). When J = 6, the probability coverage of the noncentral F distribution-based CIs was adequate only when the data were normally distributed. The bootstrap CI for f^* provided good probability coverage under normality for almost all values of f^* investigated.

Table 1.
Estimated coverage probabilities for nominal 95% noncentral F distribution-based (NCF) and percentile bootstrap (boot) CIs for f^*: J = 3, equally spaced mean configuration

f^*	n	Normal		$g = .000$ $h = .109$		$g = .000$ $h = .225$		$g = .760$ $h = -.098$		$g = .225$ $h = .225$	
		NCF	boot	NCF	boot	NCF	boot	NCF	boot	NCF	boot
.00	20	**.976**	**.984**	**.974**	**.981**	**.978**	**.984**	**.974**	**.977**	**.978**	**.980**
	35	**.965**	**.984**	**.973**	**.987**	**.972**	**.984**	**.976**	**.985**	**.978**	**.985**
	50	**.976**	**.990**	**.978**	**.990**	**.979**	**.985**	**.975**	**.985**	**.971**	**.983**
.10	20	.949	**.982**	.952	**.974**	.953	**.969**	.950	**.970**	.951	**.968**
	35	.951	**.982**	.953	**.984**	.948	**.973**	.951	**.978**	.956	**.975**
	50	.951	**.984**	.951	**.984**	.955	**.975**	.950	**.974**	.951	**.977**
.25	20	.948	**.965**	.947	**.964**	.953	**.938**	.949	.947	**.938**	**.924**
	35	.950	**.974**	.947	**.965**	**.935**	.952	.950	.946	**.932**	**.937**
	50	.950	**.968**	**.938**	.956	**.936**	.943	.946	.958	**.923**	**.931**
.40	20	.942	.948	.941	.946	**.925**	**.917**	**.933**	**.918**	**.913**	**.876**
	35	.959	.956	**.932**	**.938**	**.926**	**.923**	**.935**	**.925**	**.912**	**.894**
	50	.951	.950	**.932**	**.935**	**.908**	**.912**	**.934**	**.926**	**.900**	**.891**
.55	20	.946	**.932**	**.935**	**.923**	**.900**	**.865**	**.914**	**.886**	**.875**	**.830**
	35	.950	.943	**.928**	**.926**	**.901**	**.895**	**.915**	**.897**	**.859**	**.860**
	50	.951	.944	**.934**	**.934**	**.886**	**.902**	**.926**	**.919**	**.844**	**.856**
.70	20	.952	**.934**	**.928**	**.913**	**.880**	**.866**	**.909**	**.875**	**.848**	**.812**
	35	**.938**	**.922**	**.936**	**.925**	**.860**	**.865**	**.904**	**.903**	**.808**	**.812**
	50	.949	.941	**.938**	**.936**	**.846**	**.866**	**.899**	**.911**	**.786**	**.827**

Note. Bold values are estimates outside the interval $[.94, .96]$ and bold underlined values are outside the interval $[.925, .975]$.

Table 2.
Estimated coverage probabilities for nominal 95% noncentral F distribution-based (NCF) and percentile bootstrap (boot) CIs for f^* : J = 3, one extreme mean configuration

f^*	n	Normal		$g=.000$ $h=.109$		$g=.000$ $h=.225$		$g=.760$ $h=-.098$		$g=.225$ $h=.225$	
		NCF	boot	NCF	boot	NCF	boot	NCF	boot	NCF	boot
.00	20	**.973**	**.984**	**.970**	**.978**	**.980**	**.982**	**.986**	**.983**	**.974**	**.980**
	35	**.975**	**.991**	**.977**	**.989**	**.978**	**.986**	**.977**	**.986**	**.974**	**.982**
	50	**.972**	**.986**	**.972**	**.983**	**.976**	**.986**	**.976**	**.986**	**.977**	**.986**
.10	20	.956	**.978**	.950	**.976**	.958	**.978**	.956	**.971**	.954	**.970**
	35	.947	**.981**	.942	**.976**	.942	**.970**	.942	**.970**	.952	**.972**
	50	.945	**.981**	.946	**.976**	.954	**.980**	.953	**.979**	.952	**.975**
.25	20	.949	**.964**	.949	.960	.951	.942	.954	.953	.940	**.934**
	35	.948	**.968**	.943	.959	.944	.953	.942	.956	.940	**.930**
	50	.945	**.961**	.950	**.962**	**.936**	.940	.951	**.964**	**.938**	**.935**
.40	20	.948	.954	**.938**	**.936**	**.920**	**.899**	**.933**	**.922**	**.911**	**.886**
	35	.950	.950	.942	**.939**	**.922**	**.919**	**.933**	**.929**	**.912**	**.899**
	50	.950	.950	.942	.944	**.916**	**.918**	**.934**	**.933**	**.896**	**.894**
.55	20	.945	**.936**	**.933**	**.927**	**.908**	**.876**	**.931**	**.907**	**.881**	**.850**
	35	.944	**.938**	**.928**	**.922**	**.892**	**.880**	**.916**	**.912**	**.867**	**.864**
	50	.949	**.945**	**.935**	**.930**	**.885**	**.889**	**.923**	**.928**	**.836**	**.862**
.70	20	.949	**.932**	.940	**.921**	**.871**	**.845**	**.910**	**.888**	**.843**	**.811**
	35	.945	**.935**	**.934**	**.937**	**.850**	**.851**	**.896**	**.894**	**.807**	**.822**
	50	.950	.941	**.936**	**.936**	**.856**	**.867**	**.905**	**.922**	**.791**	**.828**

Note. Bold values are estimates outside the interval $[.94, .96]$ and bold underlined values are outside the interval $[.925, .975]$.

Table 3.
Estimated coverage probabilities for nominal 95% noncentral F distribution-based (NCF) and percentile bootstrap (boot) CIs for f^* : J = 6 , equally spaced mean configuration

f^*	n	Normal		$g=.000$ $h=.109$		$g=.000$ $h=.225$		$g=.760$ $h=-.098$		$g=.225$ $h=.225$	
		NCF	boot	NCF	boot	NCF	boot	NCF	boot	NCF	boot
.00	20	**.977**	.950	**.975**	.944	**.980**	.949	**.976**	.946	**.980**	.948
	35	**.974**	.955	**.976**	.950	**.976**	.953	**.975**	.949	**.980**	.954
	50	**.972**	.956	**.972**	.954	**.985**	**.966**	**.977**	.952	**.978**	.956
.10	20	.953	.951	.952	**.927**	.954	**.935**	.959	**.932**	.952	**.924**
	35	.951	.943	.942	.944	.953	**.938**	.950	.943	.944	**.928**
	50	.944	.945	.950	.945	.958	.946	.940	.942	.954	**.928**
.25	20	.948	**.927**	.948	**.921**	**.938**	**.892**	.950	**.905**	**.928**	**.871**
	35	.952	.944	.944	**.933**	**.937**	**.905**	**.938**	**.910**	**.919**	**.889**
	50	.954	.954	.943	**.933**	**.932**	**.910**	.944	**.926**	**.910**	**.880**
.40	20	.950	**.933**	.945	**.920**	**.917**	**.858**	**.922**	**.901**	**.880**	**.819**
	35	.953	**.937**	.940	**.927**	**.900**	**.877**	**.928**	**.906**	**.860**	**.837**
	50	.955	.947	.943	**.935**	**.904**	**.890**	**.932**	**.924**	**.860**	**.859**
.55	20	.949	**.923**	**.934**	**.904**	**.876**	**.825**	**.914**	**.874**	**.856**	**.800**
	35	.958	.940	**.931**	**.921**	**.872**	**.860**	**.902**	**.889**	**.818**	**.807**
	50	.954	**.939**	**.930**	**.928**	**.869**	**.884**	**.914**	**.910**	**.808**	**.840**
.70	20	.955	**.930**	**.932**	**.893**	**.849**	**.816**	**.893**	**.876**	**.790**	**.752**
	35	.942	**.930**	**.923**	**.914**	**.826**	**.837**	**.892**	**.893**	**.766**	**.784**
	50	.943	**.932**	**.918**	**.927**	**.820**	**.857**	**.895**	**.918**	**.752**	**.823**

Note. Bold values are estimates outside the interval $[.94, .96]$ and bold underlined values are outside the interval $[.925, .975]$.

Table 4.
Estimated coverage probabilities for nominal 95% noncentral F distribution-based and percentile bootstrap (boot) CIs for f^*: J = 6, one extreme mean configuration

f^*	n	Normal		$g = .000$ $h = .109$		$g = .000$ $h = .225$		$g = .760$ $h = -.098$		$g = .225$ $h = .225$	
		NCF	boot	NCF	boot	NCF	boot	NCF	boot	NCF	boot
.00	20	**.975**	.949	**.976**	.946	**.974**	.942	**.972**	**.936**	**.982**	.948
	35	**.976**	.956	**.968**	.947	**.978**	.951	**.978**	.952	**.976**	.954
	50	**.975**	.958	**.976**	.959	**.982**	.958	**.980**	.956	**.970**	.944
.10	20	.949	.944	.948	**.927**	.959	**.938**	.954	**.926**	.949	**.926**
	35	.954	.950	.954	.943	.948	**.932**	.947	**.930**	.955	**.935**
	50	.952	.948	.946	.950	.950	**.936**	.957	**.945**	.953	**.933**
.25	20	.954	**.934**	.947	**.920**	**.935**	**.894**	.941	**.911**	.942	**.884**
	35	.953	.940	.948	.946	**.939**	**.910**	.947	**.933**	**.927**	**.888**
	50	.953	.947	.945	**.929**	**.932**	**.909**	.948	**.939**	**.917**	**.898**
.40	20	.952	**.930**	.951	**.924**	**.918**	**.860**	.947	**.898**	**.890**	**.827**
	35	.946	**.932**	**.937**	**.924**	**.911**	**.892**	**.934**	**.917**	**.883**	**.862**
	50	.950	**.936**	**.938**	**.931**	**.900**	**.894**	**.932**	**.932**	**.856**	**.860**
.55	20	.955	**.931**	**.938**	**.902**	**.877**	**.838**	**.923**	**.886**	**.844**	**.793**
	35	.951	**.930**	**.929**	**.919**	**.863**	**.862**	**.916**	**.909**	**.821**	**.824**
	50	.949	**.936**	**.922**	**.925**	**.858**	**.879**	**.909**	**.908**	**.783**	**.820**
.70	20	.945	**.915**	**.929**	**.893**	**.848**	**.826**	**.914**	**.885**	**.794**	**.754**
	35	.947	**.935**	**.920**	**.911**	**.828**	**.834**	**.896**	**.908**	**.752**	**.790**
	50	.942	**.930**	**.926**	**.918**	**.817**	**.849**	**.902**	**.920**	**.740**	**.815**

Note. Bold values are estimates outside the interval $[.94, .96]$ and bold underlined values are outside the interval $[.925, .975]$.

Table 5.
Average widths of noncentral F distribution-based (NCF) and percentile bootstrap (boot) CIs for f^* :
J=3, equally spaced mean configuration

f^*	n	Normal		$g=.000$ $h=.109$		$g=.000$ $h=.225$		$g=.760$ $h=-.098$		$g=.225$ $h=.225$	
		NCF	boot	NCF	boot	NCF	boot	NCF	boot	NCF	boot
.00	20	.446	.534	.449	.529	.442	.515	.451	.520	.448	.513
	35	.338	.393	.334	.388	.339	.383	.335	.381	.340	.381
	50	.281	.323	.278	.321	.283	.318	.284	.320	.285	.316
.10	20	.467	.551	.470	.545	.479	.542	.476	.541	.479	.542
	35	.367	.416	.361	.409	.367	.408	.369	.409	.370	.407
	50	.309	.346	.313	.348	.314	.345	.314	.345	.317	.346
.25	20	.560	.628	.561	.627	.568	.627	.568	.629	.577	.640
	35	.453	.495	.452	.490	.457	.493	.456	.492	.457	.497
	50	.395	.425	.393	.424	.396	.427	.396	.426	.394	.429
.40	20	.641	.701	.641	.702	.642	.710	.638	.707	.648	.724
	35	.497	.533	.495	.533	.495	.547	.497	.543	.496	.555
	50	.413	.437	.413	.442	.413	.454	.414	.449	.413	.465
.55	20	.676	.726	.676	.739	.678	.764	.676	.754	.677	.781
	35	.504	.526	.504	.538	.506	.569	.505	.559	.507	.593
	50	.417	.429	.418	.444	.419	.477	.418	.461	.420	.498
.70	20	.693	.732	.692	.753	.696	.813	.696	.800	.702	.842
	35	.514	.527	.515	.550	.517	.612	.516	.590	.521	.640
	50	.428	.433	.428	.457	.430	.512	.428	.493	.433	.547

Table 6.
Average widths of noncentral F distribution-based (NCF) and percentile bootstrap (boot) CIs for f^* : J=3, one extreme mean configuration

f^*	n	Normal		$g = .000$ $h = .109$		$g = .000$ $h = .225$		$g = .760$ $h = -.098$		$g = .225$ $h = .225$	
		NCF	boot	NCF	boot	NCF	boot	NCF	boot	NCF	boot
.00	20	.448	.535	.452	.531	.453	.519	.442	.514	.452	.518
	35	.335	.392	.336	.387	.338	.383	.334	.381	.341	.380
	50	.280	.324	.287	.325	.283	.316	.280	.317	.284	.316
.10	20	.472	.552	.473	.549	.482	.546	.473	.539	.476	.540
	35	.361	.413	.365	.410	.369	.410	.367	.408	.371	.410
	50	.312	.349	.312	.346	.315	.346	.315	.345	.319	.348
.25	20	.562	.629	.566	.629	.573	.634	.564	.622	.578	.637
	35	.452	.493	.456	.493	.457	.498	.455	.489	.464	.503
	50	.394	.423	.394	.423	.395	.426	.395	.422	.400	.432
.40	20	.641	.703	.638	.698	.643	.713	.643	.701	.645	.716
	35	.496	.534	.496	.533	.496	.541	.496	.532	.496	.549
	50	.414	.437	.414	.440	.414	.456	.414	.442	.413	.459
.55	20	.676	.726	.675	.737	.678	.763	.679	.744	.679	.777
	35	.504	.527	.504	.542	.506	.568	.505	.551	.507	.586
	50	.417	.428	.418	.443	.420	.477	.418	.453	.420	.493
.70	20	.692	.729	.693	.756	.698	.805	.695	.782	.701	.840
	35	.514	.530	.514	.553	.518	.612	.516	.581	.521	.634
	50	.427	.433	.427	.457	.430	.515	.428	.484	.432	.544

Table 7.
Average widths of noncentral F distribution-based (NCF) and percentile bootstrap (boot) CIs for f^* :
J=6, equally spaced mean configuration

f^*	n	Normal		$g = .000$ $h = .109$		$g = .000$ $h = .225$		$g = .760$ $h = -.098$		$g = .225$ $h = .225$	
		NCF	boot	NCF	boot	NCF	boot	NCF	boot	NCF	boot
.00	20	.321	.434	.325	.433	.319	.427	.318	.428	.324	.428
	35	.239	.320	.242	.319	.241	.317	.241	.317	.241	.315
	50	.202	.266	.201	.265	.202	.263	.200	.263	.201	.263
.10	20	.344	.446	.350	.446	.348	.443	.350	.443	.351	.443
	35	.274	.337	.271	.335	.277	.336	.272	.334	.276	.333
	50	.238	.284	.237	.283	.239	.283	.238	.282	.238	.280
.25	20	.426	.479	.424	.475	.428	.473	.429	.474	.429	.474
	35	.332	.353	.332	.351	.331	.349	.332	.350	.331	.350
	50	.275	.285	.275	.286	.273	.285	.275	.286	.273	.287
.40	20	.450	.469	.448	.468	.446	.475	.447	.473	.444	.481
	35	.323	.330	.324	.334	.324	.347	.324	.341	.324	.357
	50	.265	.268	.265	.274	.265	.290	.265	.281	.266	.301
.55	20	.442	.452	.442	.463	.444	.492	.443	.480	.444	.510
	35	.324	.328	.324	.339	.325	.374	.325	.356	.326	.393
	50	.268	.270	.268	.282	.269	.317	.269	.298	.270	.340
.70	20	.448	.457	.449	.478	.453	.529	.449	.513	.455	.565
	35	.332	.336	.332	.357	.334	.412	.333	.386	.336	.438
	50	.276	.277	.276	.299	.277	.352	.276	.323	.278	.381

Table 8.
Average widths of noncentral F distribution-based (NCF) and percentile bootstrap (boot) CIs for f^* : J=6, one extreme mean configuration

f^*	n	Normal		$g = .000$ $h = .109$		$g = .000$ $h = .225$		$g = .760$ $h = -.098$		$g = .225$ $h = .225$	
		NCF	boot	NCF	boot	NCF	boot	NCF	boot	NCF	boot
.00	20	.322	.434	.319	.431	.322	.429	.323	.432	.321	.427
	35	.239	.319	.239	.318	.243	.317	.242	.318	.241	.315
	50	.202	.266	.199	.263	.202	.264	.199	.263	.203	.263
.10	20	.350	.449	.347	.444	.349	.442	.352	.444	.355	.445
	35	.274	.338	.274	.337	.275	.333	.274	.334	.278	.335
	50	.236	.284	.237	.283	.236	.281	.237	.282	.240	.282
.25	20	.425	.480	.425	.477	.427	.474	.429	.469	.432	.473
	35	.333	.354	.331	.351	.332	.351	.333	.344	.331	.345
	50	.276	.285	.275	.286	.274	.286	.276	.278	.274	.284
.40	20	.449	.469	.449	.468	.447	.475	.449	.453	.446	.475
	35	.324	.329	.324	.335	.324	.350	.323	.330	.324	.352
	50	.265	.268	.265	.274	.265	.290	.265	.273	.266	.296
.55	20	.442	.452	.442	.463	.444	.496	.443	.468	.445	.508
	35	.324	.327	.324	.340	.325	.375	.324	.349	.326	.391
	50	.268	.270	.268	.283	.269	.318	.269	.294	.270	.334
.70	20	.448	.458	.449	.481	.452	.534	.449	.502	.455	.557
	35	.332	.334	.332	.357	.335	.412	.333	.380	.336	.438
	50	.276	.277	.276	.297	.278	.350	.276	.320	.278	.381

However, for all other combinations of conditions, the bootstrap CI did not provide accurate probability coverage. Furthermore, excluding $f^* = 0$, the coverage performance of the noncentral F distribution-based CIs tended to decline as f^* increased, as the distributions became more long-tailed, and appeared to be worse for skewed distributions. Overall, the noncentral F distribution-based CIs for f^* yielded relatively better probability coverage than that of the bootstrap CIs for f^*. The type of mean configurations and the number of treatment groups did not appear to affect the coverage probability of the CIs for f^* considerably. Therefore, the coverage performance of the CIs for f^* might be generalizable over types of mean configuration and various numbers of treatment groups.

The widths of the noncentral F distribution-based CIs for f^* were all narrower than those of the bootstrap CIs under the same condition. The interval widths of the CIs for f^* were relatively unchanged across data distributions. The width of both estimated CIs became narrower as the number of levels for J increased, the sample size increased, and the population effect size f^* decreased.

In summary, both the noncentral F distribution-based and the bootstrap CIs for f^*, which are based on the least-square estimators, yielded inadequate coverage probabilities. Thus an important task to help researchers who want to set a CI around f^* is developing a better interval than the noncentral F distribution-based or percentile bootstrap CI. An improved measure of effect size might be attained by substituting robust estimators, e.g., trimmed means and Winsorized variances, for the least-square values. Thus, one of our future studies has set out to propose a robust version of f^*. A robust measure of effect size may yield better coverage probabilities and provide a measure that is not likely to be strongly affected by outlying data points.

References

Algina, J., & Keselman, H. J. (2003a). Approximate confidence intervals for effect sizes. *Educational and Psychological Measurement*, *63*, 537-553.

Algina, J., & Keselman, H. J. (2003b, May). Confidence Intervals for Effect Sizes. Paper presented at a conference in honor of H. Swaminathan, Amherst, MA.

Algina, J., Keselman, H. J., & Penfield, R. D. (2005a). Effect sizes and their intervals: the two-level repeated measures case. *Educational and Psychological Measurement, 65*, 241-258.

Algina, J., Keselman, H. J., & Penfield, R. D. (2005b). An alternative to Cohen's standardized mean difference effect size: A robust parameter and confidence interval in the two independent groups case. *Psychological Methods*, *10*, 317-328.

Algina, J., Keselman, H. J., & Penfield, R. D. (2006). Confidence interval coverage for Cohen's effect size Statistic. *Educational and Psychological Measurement*, *66*, 945-990

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

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical & Statistical Psychology, 31*, 144-152.

Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin, 114*, 494-509.

Cliff, N. (1996). Answering ordinal questions with ordinal data using ordinal statistics, *Multivariate Behavioral Research, 31*, 331-350.

Cohen, J. (1965). Some statistical issues in psychological research. In B.B. Wolman (Ed.), *Handbook of clinical psychology* (pp. 95-121). New York: Academic Press.

Cohen, J. (1969). *Statistical power analyses for the behavioral sciences*. NY: Academic Press, Inc.

Cohen, J. (1988). *Statistical power and analysis for the behavioral sciences* (2nd ed.). NY: Academic Press.

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

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.

Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist, 60*, 178-180.

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

Finch, S., Thomason, N., & Cumming, G. (2002). Past and future American Psychological Association guidelines for statistical practice. *Theory and Psychology, 36*, 312-324.

Grissom, R. J., & Kim, J. J. (2005). Effect sizes for research: A broad practical approach. Mahwah, NJ: Lawrence Erlbaum.

Hays, W. L. (1963). *Statistics*. NY, Rinehart & Winston.

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

Hess, M. R. & Kromrey, J. D. (April, 2003). *Confidence intervals for standardized mean differences: An empirical comparison of bootstrap methods under non-normality and heterogeneous variances.* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Hess, M. R. & Kromrey, J. D. (April, 2004). *Robust confidence intervals for effect sizes: A comparative study of Cohen's d and Cliff's delta under non-normality and heterogeneous variances.* Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.

Hoaglin, D. C. (1983). Summarizing shape numerically: The g-and h distributions. In D. D. Hoaglin, F. Mosteller, & Tukey, J. W. (Eds.), *Data analysis for tables, trends, and shapes: Robust and exploratory techniques*. New York: Wiley.

Hogarty, K. Y., & Kromrey, J. D. (April, 2001). *We've been reporting some effect sizes: Can you guess what they mean?* Paper presented at the Annual Meeting of the American Educational Research Association, Seattle, WA.

Johnson, N. L., & Welch, B. L. (1940). Applications of the non-central *t* distribution. *Biometrika, 34*, 362-389.

Kelley, K. (2005). The effects of nonnormal distributions on confidence intervals around the standardized mean difference: Bootstrap and parametric confidence intervals. *Educational and Psychological Measurement, 65*, 51-69.

Kraemer, H. C., & Andrews, G. A. (1982). A nonparametric technique for meta-analysis effect size calculation. *Psychological Bulletin, 91*, 404-412.

Martinez, J., & Iglewicz, B. (1984). Some properties of the Tukey *g* and *h* family of distributions. *Communications in Statistics, Theory and Methods, 13*, 353-369.

McGraw, K. O. & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin, 111*, 361-365.

Meehl, P. E. (1967). Theory testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34*, 103-115.

Murphy, K. R. (1997). Editorial. *Journal of Applied Psychology, 82*, 3-5.

Nickerson, R. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5*, 241-301.

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods, 8*, 434-447.

SAS Institute Inc. (1999). *SAS/IML user's guide, version 8*. Cary, NC: Author.

Serlin, R. C., & Lapsley, D. K. (1985). Rationality in psychological research: The good enough principle. *American Psychologist, 40,* 73-83.

Steiger, J. H. (2004). Beyond the *F* test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods, 9*, 164-182.

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

Thompson, B. (1994). Guidelines for authors. *Educational and Psychological Measurement, 54*, 837-847.

Vargha, A. & Delaney, H. D. (2000). A critique and improvement of the CL Common Language effect size statistics of McGraw and Wong. *Journal of Educational and Behavioral Statistics, 25*, 101-132.

Wilcox, R. R. (2003). *Applying contemporary statistical techniques*. San Diego: Academic Press.

Wilcox, R. R. & Keselman. (2003). Modern robust data analysis methods: Measures of central tendency. *Psychological Methods, 8*, 254-274.

Wilcox, R. R. & Muska, J. (1999). Measuring effect size: A non-parametric analogue of ω^2. *British Journal of Mathematical and Statistical Psychology, 52*, 93-110.

Wilkinson, L. and the Task force on Statistical Inference (1999). Statistical methods in psychology journals. *American Psychologist, 54*, 594-604.