Journal of Modern Applied Statistical Methods

Volume 10 | Issue 1 Article 9

5-1-2011

# A Robust 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

Recommended Citation
Zhang, Guili and Algina, James (2011) "A Robust Root Mean Square Standardized Effect Size in One-Way Fixed-Effects ANOVA," *Journal of Modern Applied Statistical Methods*: Vol. 10 : Iss. 1 , Article 9.
DOI: 10.22237/jmasm/1304222880
Available at: http://digitalcommons.wayne.edu/jmasm/vol10/iss1/9

Journal of Modern Applied Statistical Methods
May 2011, Vol. 10, No. 1, 77-96

# A Robust Root Mean Square Standardized Effect Size in One-Way Fixed-Effects ANOVA

Guili Zhang
East Carolina University

James Algina
University of Florida

A robust Root Mean Square Standardized Effect Size (RMSSE$_R$) was developed to address the unsatisfactory performance of the Root Mean Square Standardized Effect Size. The coverage performances of the confidence intervals (CI) for RMSSE$_R$ were investigated. The coverage probabilities of the non-central *F* distribution-based CI for RMSSE$_R$ were adequate.

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

## Introduction

Using an effect size (ES) in addition to or in place of a hypothesis test has been enthusiastically advocated by many statistical methodologists because ESs are regarded as 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; Zhang, 2009; Zhang & Algina, 2008). Reporting an ES has become mandatory or strongly recommended in some editorial policies in the last two decades (Murphy, 1997; Thompson, 1994). *The Publication Manual of the American Psychological Association* (2001) stated that it is almost always necessary to include some index of ES or strength of relationship in the results section of a research report.

The APA Task Force on Statistical Inference (Wilkinson and the Task Force on Statistical Inference, 1999) not only supports the use of ESs but also requires 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 it is argued that 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; Steiger & Fouladi, 1997; Wilkinson, et al., 1999; Cumming & Finch, 2001, 2005; Zhang, 2009).

The increased interests in ES and CI have motivated explorations of their usefulness and effectiveness within recent years (Algina & Keselman, 2003a, 2003b; Bird, 2002; Cumming & Fitch, 2001; Zhang & Algina, 2008). In the two group case, it has been reported that - in both the independent and dependent samples cases - CIs for Cohen's $\delta$, arguably the most widely accepted ES index for a pairwise contrast on means, may be misleading due to poor coverage probability when data are nonnormal and can grossly misrepresent the degree to which two distributions differ (Algina & Keselman, 2003b; Algina, et al., 2006; Algina, et al., 2005a; Kelly, 2005; Wilcox & Keselman, 2003). However, research has shown that the CIs for $\delta_R$, a robust version of $\delta$ based on trimmed

Guili Zhang is an Assistant Professor of Research and Evaluation Methodology in the Department of Curriculum and Instruction, East Carolina University. Her research interests are in applied statistics. Email her at: zhangg@ecu.edu. James Algina is a Professor of Research and Evaluation Methodology at the University of Florida. His research interests are in psychometric theory and applied statistics. Email him at: algina@ufl.edu.

means and Winsorized variances, have better coverage probability than do CIs for Cohen's $\delta$ under data nonnormality (Algina & Keselman, 2003b).

In the more than two group case, Zhang and Algina (2008) investigated the coverage performance of the noncentral *F* distribution-based CI and the percentile bootstrap CI for one of the most commonly used generalized effect size indices, the Root Mean Square Standardized Effect Size (RMSSE), proposed by Steiger and Fouladi (1997), denoted by

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

in a one-way, fixed-effects, between-subjects ANOVA. Both CIs were implemented for all combinations of the following five factors: (1) 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); (2) two numbers of levels for treatment groups: *J* = 3 and *J* = 6; (3) three cell sample sizes in each treatment; (4) six values of population RMSSEs; and (5) two mean configurations, the equally spaced mean configuration and the one extreme mean configuration. Each condition was replicated 2,500 times and the number of bootstrap replications in the bootstrap procedure was 1,000. Zhang and Algina found that both the noncentral *F* distribution-based CI and the percentile bootstrap CI for RMSSE yielded inadequate coverage probabilities under data nonnormality.

According to arguments in Wilcox and Keselman (2003) about the robustness of $\delta$ in the two-group case, it is not surprising that $f^*$ is not an entirely adequate measure of group separation because $f^*$ is formulated with least-square parameters which are affected by skewed data, long tails and/or outlying values. It is therefore imperative to develop a robust version of the RMSSE to ensure the appropriate and effective use of the ES in ANOVA.

## Methodology

The unsatisfactory coverage performance of the CIs for $f^*$ reported by Zhang and Algina (2008) is understandable: This is because the problems that trouble Cohen's $\delta$ and its CI are very likely to also haunt $f^*$ and its CI, as $f^*$ is a generalized $\delta$ and is formulated with the nonrobust least-square means and variances. It is well known that when the distribution of the data is not normal, the least-square means and standard deviations can work poorly because they are affected by the skewness of the data and by the outliers in the data; consequently $f^*$ may be misleading as a measure of population separation. Therefore, a robust version of $f^*$ that is parallel to $\delta_R$, the robust effect size in the two-group case, is strongly desired. The purposes of the study are:

a. To develop a robust RMSSE, $f_R^*$,
b. To develop a noncentral *F* distribution-based CI for $f_R^*$, and
c. To investigate the performance of the noncental *F* distribution-based and percentile bootstrap CI for $f_R^*$.

Note that $f^*$ and $f_R^*$ are two different parameters based on different measures of location and variability and, unless the data are normally distributed, $f^*$ and $f_R^*$ will not be equal. The parameter $f^*$ is used to characterize the amount of difference among the population means, while $f_R^*$ represents the amount of difference among the population trimmed means.

### Robust Root Mean Square Standardized Effect Size and Its Confidence Interval

To overcome the weaknesses in $f^*$, a robust version of the generalized effect size was developed, the Robust Root Mean Square Standardized Effect Size (RMSSE$_R$), denoted by $f_R^*$ in this study. The value of $f_R^*$ is defined by using robust parameters (20% trimmed means and Winsorized variances) as opposed to the least-square parameters (means and variances). Trimmed means are used because it has been shown that the impact of outliers on trimmed means can be much less disturbing than on the usual means (Wilcox, 2005). The Winsorized variance is used because the sample Winsorized variance is used in hypothesis testing based on trimmed means. Both the trimmed mean and the Winsorized variance are robust parameters as judged by the criteria of qualitative robustness, quantitative robustness and infinitesimal robustness (Wilcox, 2005, Section 2.1 describes these criteria).

In a balanced one-way between-subjects ANOVA design, $f_R^*$ is defined as

$$f_R^* = .642\sqrt{\frac{\sum_{j=1}^{J}\left(\mu_{Tj}-\mu_T\right)^2}{\left(J-1\right)\sigma_W^2}}\,, \tag{1}$$

where $\mu_{Tj}$ is the trimmed mean for the $j^{\text{th}}$ level, $\mu_T$ is the grand mean based on the trimmed means, and ${\sigma_W}^2$ is the within-level Winsorized variance, which is assumed to be constant across levels. The quantity 0.642 is the square root of the population Winsorized variance for a standard normal distribution, therefore, including 0.642 in the definition of the robust effect ensures that $f_R^*$ = $f^*$ when the data are drawn from normal distributions with equal variances.

An estimate of $f_R^*$ can be attained from sample statistics by applying the following formula:

$$\hat{f}_R^* = .642\times\sqrt{\frac{\sum_{j=1}^{J}\left(\bar{Y}_{Tj}-\bar{Y}_T\right)^2}{\left(J-1\right)S_{Wp}^2}}\,, \tag{2}$$

where $\bar{Y}_{Tj}$ is the trimmed sample mean for the $j^{\text{th}}$ level, $\bar{Y}_T$ is the sample grand trimmed mean, and ${S_{Wp}}^2$ is the sample pooled within-level Winsorized variance.

The quantity $S_{Wp}^2$ is obtained by using

$$S_{Wp}^2 = \frac{\sum_{j=1}^{J}(n_j-1)S_{Wj}^2}{\sum_{j=1}^{J}n_j-J}\,. \tag{3}$$

A CI for $f_R^*$ can be constructed based on the noncentral $F$ distribution. Consider a one-way, between-subjects, fixed-effects ANOVA with $n_j$ observations in the $j^{\text{th}}$ group and $J$ groups. The robust $F$ statistic is calculated by using (Yuen, 1974)

$$F_R = \frac{MS_{RB}}{MS_{RW}}, \tag{4}$$

where $MS_{RB}$ and $MS_{RW}$ are the robust mean square between and robust mean square within respectively, and are calculated by using:

$$MS_{RB} = \frac{\sum_{j=1}^{J}h_j(\bar{Y}_{Tj}-\bar{Y}_T)^2}{J-1} \tag{5}$$

and

$$MS_{RW} = \frac{\sum_{j=1}^{J}\sum_{i=1}^{n_i}(Y_{Wij}-\bar{Y}_{Wj})^2}{\sum_{j=1}^{J}h_j-J}, \tag{6}$$

where $Y_{Wij}$ is the $i^{\text{th}}$ Winsorized score in group $j$, and $\bar{Y}_{Wj}$ is the Winsorized mean for group $j$. The robust $F$ statistic has robust noncentrality parameter

$$\lambda_R = \frac{\sum_{j=1}^{J}h_j(\mu_{Tj}-\mu_T)^2}{\tilde{\sigma}_W^2}, \tag{7}$$

where $\tilde{\sigma}_W^2$ is an adjusted version of the population Winsorized variance:

$$\tilde{\sigma}_W^2 = \frac{N-J}{\sum_{j=1}^{J} h_j - J}\sigma_W^2 . \qquad (8)$$

The lower limit of the 95% CI for $\lambda_R$ is the robust noncentrality parameter for the noncentral *F* distribution in which the calculated robust *F* statistic is the 0.975 quantile. The upper limit of the 95% confidence interval for $\lambda_R$ is the robust noncentrality parameter for the noncentral *F* distribution in which the calculated robust *F* statistic is the 0.025 quantile of the distribution.

In a balanced one-factor between-subject design with equal *n*s, $f_R^*$ can be written as a function of $\lambda_R$:

$$f_R^* = \sqrt{\frac{.4129\times\left(\sum_{j=1}^{J} n_j - J\right)}{\left(\sum_{j=1}^{J} h_j - J\right)\times(J-1)h}\lambda_R} . \qquad (9)$$

To find a (1 − $\alpha$)% (95% in this study) CI for $f_R^*$, the noncentral *F* distribution was first used to find a 95% CI for $\lambda_R$. After the CI on $\lambda_R$ is found, equation 9 is applied to transform the endpoints of the CI for $\lambda_R$ to obtain the endpoints for the CI for $f_R^*$.

Although the noncentral *F* distribution can be used to obtain a CI for $f_R^*$, because this CI construction method is based on the assumption that the data are drawn from a normal distribution, when the data are nonnormal the coverage probability for this interval may be poor and the percentile bootstrap CI may have better coverage probability (Algina & Keselman, 2003b; Efron & Tibshirani, 1993). Therefore, the performances of the percentile bootstrap method for the construction of CIs for $f_R^*$ were examined and compared to the noncentral *F* distribution-based method in terms of the probability coverage and interval width.

## Coverage Performance of the Confidence Interval for Robust Root Mean Square Standardized Effect Size

To investigate the coverage performance of the CIs for $f_R^*$, the noncentral *F* distribution-based and the percentile bootstrap CIs were implemented for all combinations of the following five factors: (1) 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); (2) two numbers of levels for treatment groups: *J* = 3 and *J* = 6; (3) three cell sample sizes in each treatment; (4) six values of population RMSSE$_R$; (5) two mean configurations, the equally spaced mean configuration and the one extreme mean configuration. The nominal confidence level for all intervals investigated was 0.95 and each condition was replicated 2,500 times. The number of bootstrap replications in the bootstrap procedure was 1,000.

## Conditions

Data for all five distributions were generated from the *g* and *h* distributions: (1) $g = h = 0$, 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, (2) $g = .76$ and $h = -.098$, a distribution with the skewness and kurtosis of an exponential distribution ($\gamma_1 = 2$, $\gamma_2 = 6$), (3) $g = 0$ and $h = .225$ ($\gamma_1 = 0$ and $\gamma_2 = 154.84$), (4) $g = h = .225$ ($\gamma_1 = 4.90$ and $\gamma_2 = 4673.80$), and (5) $g = 0$ 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 strongly nonnormal. Such a selection of distributions allows the researcher to investigate the performances of the CIs under a wide range of the 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 of the simulations 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 inclusion of the normal distribution provides a reference for judgments on the CIs' performance under data that deviate from normality.

The numbers of treatment groups investigated were 3 and 6 ($J$ = 3 and $J$ = 6), and sample sizes in each treatment included were $n_j = 20$ to 50 in steps of 15. In other words, the treatment groups have equal sample size and the sample sizes investigated were 20, 35 and 50. The number of treatment groups equal to 3 and 6 was selected because this covers the likely range encountered in most research in the social and behavioral sciences. Sample sizes ranging from 20 to 50 are fairly typical of sample sizes used in social science research, although clearly do 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. A mean configuration is a specification of the arrangement of the treatment groups means. Denoting the smallest and the largest means by $\mu_{min}$ and $\mu_{max}$, if the means other than $\mu_{min}$ and $\mu_{max}$ are equally spaced between these two extremes, the configuration is referred to as an equally spaced configuration (Cohen, 1969). If one of the means is equal to $\mu_{min}$ and the rest of the means are all equal to $\mu_{max}$, or, if one of the means is equal to $\mu_{max}$ and the rest of the means are equal to $\mu_{min}$, then the configuration is called a one extreme mean configuration. Mean configurations are an artifice adopted because the actual configuration of means in social science research is quite variable. Nevertheless, the selected configurations cover a range of possibilities and will allow determination of whether results tend to generalize over configurations.

Six values of $f_R^*$ were investigated: 0, 0.1, 0.25, 0.40, 0.55 and 0.70. Defining

$$\delta_{\max} = \frac{\mu_{\max} - \mu_{\min}}{\sigma} \tag{10}$$

as Cohen's effect size for the largest and smallest means, under the equally spaced mean configurations, these population $f_R^*$ values approximately correspond to $\delta_{\max}$ of 0, 0.2, 0.5, 0.8, 1.10 and 1.40, respectively. Under the one extreme mean configuration, these population $f_R^*$ values roughly correspond to $\delta_{\max}$ of 0, 0.173, 0.433, 0.693, 0.952, and 1.212. Therefore, a $f_R^*$ of 0 indicates no effect, .1 a small effect, 0.25 a medium effect, 0.40 a large effect, and 0.55 and 0.70 very large effects.

The nominal confidence level for all intervals investigated was .95 and each condition was replicated 2,500 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 1,000.

### Analyses Conducted

The study was designed to investigate the robustness of the noncentral $F$ distribution-based CIs and the percentile bootstrap CIs for $f_R^*$ to sampling from nonnormal distributions. Coverage probabilities for the noncentral $F$ distribution-based and bootstrap CIs for $f_R^*$ were estimated. Additionally, the average width of the noncentral $F$ distribution-based and bootstrap CIs for $f_R^*$ were compared.

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

$$Y = \frac{\exp(gZ) - 1}{g} \exp\left(\frac{hZ^2}{2}\right) \tag{11}$$

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

$$Y = Z \exp\left(\frac{hZ^2}{2}\right). \tag{12}$$

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 Equations 11 and 12. To create scores corresponding to the selected values of $f_R^*$, 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. Secondly, $n_2$ 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$_R$, $f_R^*$ would equal the following values: 0, 0.1, 0.25, 0.40, 0.55, and 0.70.

For the equally spaced mean configuration, the *Y* variables were obtained by using

$$Y_{ij} = X_{ij} + (j-1)\sqrt{\frac{12}{J(J+1)}} f_R^* \frac{\sigma_W}{.642}, \quad j = 1, \ldots, J. \tag{13}$$

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

$$Y_{ij} = X_{ij} + \sqrt{J} f_R^* \frac{\sigma_W}{.642}. \tag{14}$$

To find a $(1-\alpha)$% (95% in the current study) confidence interval for $f_R^*$, the noncentral *F* distribution is first used to obtain a 95% confidence interval on $\lambda_R$, the robust noncentrality parameter of the *F* distribution. Once the CI for $\lambda_R$ is found, the endpoints of the CI for $\lambda_R$ are transformed to endpoints for $f_R^*$ by applying Equation 9. Notice the CI for $f_R^*$ constructed by the noncentral *F* distribution-based method will result in coverage probability of 0.975 when $f_R^* = 0$ because the probability noncoverage from the lower side of the distribution will be 0 instead of 0.025.

To apply the percentile bootstrap method, the following steps are completed 1,000 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 $f_R^{*2}$ is estimated by using

$$\widehat{f}_R^{*2} = \frac{.642^2}{n} \frac{\sum_{j=1}^{J} n_j - J}{\sum_{j=1}^{J} h_j - J} (F_R - 1). \tag{15}$$

3. The 1,000 $f_R^{*2}$ estimates are ranked from low to high. The lower limit of the CI for $f_R^{*2}$ is determined by finding the 26th estimate in the rank order [i.e., the (0.025 x 1,000+1) th estimate]; and the 975th estimate is the upper limit of the CI for $f_R^{*2}$ (i.e. the (0.975 x 1,000)th estimate].
4. The lower limit of the CI for $f_R^*$ is equal to the square root of the lower limit of the CI for $f_R^{*2}$ if the latter lower limit is larger than zero and is zero otherwise. The upper limit of the CI for $f_R^*$ is equal to the square root of the upper limit of the CI for $f_R^{*2}$.

## Results

The estimated coverage probabilities of the noncentral *F* distribution-based CIs for $f_R^*$ are reported in Tables 1-4. The average widths of the noncentral *F* distribution-based CIs for $f_R^*$ are shown in Tables 5-8.

### Estimated Coverage Probabilities of Confidence Intervals for $f_R^*$

In Tables 1 to 4, the estimated coverage probabilities of the noncentral *F* distribution-based and bootstrap CIs for $f_R^*$ are presented with estimates outside the [.94, .96] interval bolded, and estimates outside of the interval [.925, .975] bolded and underlined.

The pattern of results for the noncentral *F* distribution-based CI for $f_R^*$ looks strikingly similar across Tables 1 to 4. First, when sampling from a normal distribution, the coverage probability of the noncentral *F* distribution-based CIs should be 0.975 when $f_R^* = 0$, and the results in Tables 1 to 4 are consistent with the theory. When $f_R^* > 0$, the coverage probability of the noncentral *F* distribution-based CI is expected to be 0.95 under normality and the results presented in Tables 1-4 are consistent with this expectation.

Second, considering the results in all four tables, coverage probability for the noncentral *F* distribution-based CI for $f_R^*$ tends to be appreciably better than for the bootstrap CI both when sampling from normal and nonnormal distributions. When sampling from the normal distribution, when $J = 3$ the coverage probability for the noncentral *F* distribution-based CI is outside the [.925, .975] interval in only 1 case out of a total of 36, while the bootstrap CI has a total of 20 cases outside this interval. Under normality, when $J = 6$, the noncentral *F* distribution-based CI coverage probabilities are outside [.925, .975] in 2 out of 36 cases, while the bootstrap CI coverage probabilities are outside this interval in 6 out of 36 cases.

For the nonnormal distributions, the noncentral *F* distribution-based CI for $f_R^*$ has noticeably fewer coverage probabilities that are outside the criterion intervals than does the bootstrap CI under each of the four distribution conditions. The number of cases that are outside the [.925, .975] criterion interval, out of a total of 72 cases under each nonnormal distribution for the noncentral *F* distribution-based and bootstrap CIs for $f_R^*$, are: 7 versus 31 for the $g = 0$ and $h = 0.109$ distribution; 7 versus 40 for the $g = 0$ and $h = 0.225$ distribution; 20 versus 38 for the $g = 0.760$ and $h = -0.098$ distribution; and 6 versus 41 for the $g = 0.225$ and $h = 0.225$ distribution.

Third, the performance of the noncentral *F* distribution-based CI under the four nonnormal distributions reveals some common characteristics across Table 1 to Table 4. When $f_R^* = 0$, roughly 50% of the coverage probabilities tend to be outside [.925, .975]. Of the coverage probabilities that are inside the interval, most are for $J = 3$ when the data are sampled from either the $g = 0$ and $h = 0.225$ distribution or the $g = 0.225$ and $h = 0.225$ distribution.

The coverage probabilities of the noncentral *F* distribution-based CI for $f_R^*$ are all inside either [.925, .975] or both intervals when $f_R^*$ is 0.10, 0.25 or 0.40. The coverage probabilities of noncentral *F* distribution-based CI for $f_R^*$ are also all inside either the [.925, .975] interval or both intervals when $f_R^*$ is 0.55 except when $n = 35$ and the data are sampled from the $g = 0.760$ and $h = -0.098$ distribution with the means following the equally spaced mean configuration. Even when $f_R^* = 0.70$, the coverage probabilities still tend to be inside the [.925, .975] interval. The exceptions occur mostly for the $g = 0.760$ and $h = -0.098$ distribution in combination with the equally spaced mean configuration. Other exceptions involve the $g = 0$ and $h = 0.225$ distribution when $n = 35$, $J = 6$, and the $g = 0.225$ and $h = 0.225$ distribution when $n = 35$ with the group means following the equally spaced mean configuration.

Overall, under all data distributions, the coverage probabilities of the noncentral *F* distribution-based CI for $f_R^*$ are adequate by the [.925, .975] criterion except for some cases of $f_R^* = 0$ and a few cases when $f_R^* = 0.70$. When $f_R^* = 0$ the probability coverage of the noncentral *F* distribution-based CI for $f_R^*$ tends to exceed 0.975, and when $f_R^* = 0.70$ the probability coverage of the noncentral *F* distribution-based CI for $f_R^*$ tends to go below 0.925. It is observed that, excluding $f_R^* = 0$, the coverage performance of the noncentral *F* distribution-based CIs for $f_R^*$ becomes less satisfactory when $f_R^*$ gets larger.

The results of the bootstrap CIs for $f_R^*$ are also presented in Tables 1 to 4. When sampling from normal distributions, when $f_R^*$= 0 and *J* = 3, the coverage probabilities of the bootstrap CI for $f_R^*$ are all above 0.975, but when $f_R^*$= 0 and *J* = 6 they are outside the [.94, .96] interval only when *n* = 20. Under normality, when $f_R^*$= 0.10 or 0.25, the coverage probabilities of the bootstrap CI for $f_R^*$ are all outside the [.925, .975] criterion interval when *J* = 3, but all inside the [.94, .96] interval when *J* = 6 except when $f_R^*$= 0.10 and *n* = 20. When $f_R^* \geq$ 0.40, coverage probabilities tend to be inside [.925, .975] for both levels of *J*, except when $f_R^*$ = 0.40 and *n* = 20 for *J* = 3, when $f_R^*$ = 0.70, *n* = 20 for *J* = 6 and the equally spaced mean configuration, and when $f_R^*$ = 0.55, *n* = 20 for *J* = 6 and the one extreme configuration.

Under the four nonnormal distributions, when $f_R^*$= 0, the coverage probability of the bootstrap CI for $f_R^*$ tends to be outside the [.925, .975] criterion interval when *J* = 3. Roughly 50% are inside [.925, .975] when *J* = 6, mostly associated with larger sample sizes. When $f_R^*$= 0.10 or 0.25, the coverage probability of bootstrap CI for $f_R^*$ tends to be outside the [.925, .975] criterion interval when *J* = 3, and inside the [.925, .975] criterion interval when *J* = 6 except when sample size is small for some data distributions. For example, for the *g* = 0 and *h* = 0.109 distribution, when *J* = 6 and $f_R^*$= 0.10 or 0.25, the coverage probabilities of the bootstrap CI are all within [.925, .975] except when *n* = 35 and the mean configuration is the one extreme mean configuration. For the other three nonnormal distributions, the coverage probabilities are outside the [.925, .975] interval mostly when *n* = 20 and *J* = 3.

The coverage probability tends to be inside either the [.925, .975] interval or both intervals in most conditions when $f_R^*$ ≥ 0.40, except when *n* = 20 and a few cases when *n* = 35. The inadequate coverage probabilities under *n* = 35 mostly occur in the conditions with *J* = 6. Overall, the performance of the coverage probability of the bootstrap CI for $f_R^*$ is much less adequate than is the performance of the noncentral *F* distribution-based CIs for $f_R^*$. Typically the coverage probability of the bootstrap CI is too high.

Average Widths of Confidence Intervals for $f_R^*$

The average widths of the noncentral *F* distribution-based and bootstrap CIs for $f_R^*$ 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 for $f_R^*$ are shorter than those of the bootstrap CIs for $f_R^*$. The difference between the widths of the two kinds of CIs has a tendency to become smaller when sample size gets larger. For both the noncentral *F* distribution-based and the bootstrap CIs for $f_R^*$, the average width of the CIs gets narrower as the sample size increases and as the population effect size $f_R^*$ decreases.

Across distributions, there is only a very trivial difference in the width of the noncentral *F* distribution-based CIs for $f_R^*$. Similar to the pattern in the widths of the bootstrap CIs for $f^*$ observed and reported by Zhang and Algina (2008), the widths of the bootstrap CIs for $f_R^*$ fluctuate very little across data distribution conditions.

Presented in Table 6, the average widths of the noncentral *F* distribution-based and bootstrap CIs for $f_R^*$ under *J* = 3 and the one extreme mean configuration shows little difference from those from the widths for the equally spaced mean configuration in Table 5.

This suggests that the type of mean configuration does not affect the precision of estimation for $f_R^*$.

Table 7 shows the average widths of the noncentral $F$ distribution-based and bootstrap CIs for $f_R^*$ under $J$ = 6 and the equally spaced mean configuration. It is fairly apparent that, when $J$ increases from 3 to 6, the intervals become narrower. This is observed for all combinations of conditions. It is also observed that, generally, the average widths of the noncentral $F$ distribution-based CIs for $f_R^*$ are shorter than those of the bootstrap CIs for $f_R^*$. This difference is consistent across all combinations of conditions. Furthermore, for both the noncentral $F$ distribution-based and the bootstrap CIs for $f_R^*$, the average width of the CIs gets narrower as the sample size increases and the population effect size $f_R^*$ decreases. Across distributions, there is very little difference in the widths of the noncentral $F$ distribution-based CIs, and the widths of the bootstrap CIs for $f_R^*$ also remain quite constant across data distribution conditions.

The average widths of the noncentral $F$ distribution-based and bootstrap CIs for $f_R^*$ 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 in the equally spaced mean configuration in Table 7, in terms of values as well as patterns observed. This suggests that the type of mean configuration does not strongly affect the estimation accuracy for $f_R^*$.

## Conclusion

Confidence intervals for effect size 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. Despite the increasing need for using CIs, much remains to be known about the robustness of the CIs in order to ensure their proper usage. Investigation and evaluation of the performance of the CIs and their robustness under various conditions are urgently needed. In the two-group case, it has been reported that in both the independent samples and dependent samples case CIs for Cohen's $\delta$ may be misleading because of poor coverage probability when data are nonnormal (Algina & Keselman, 2003b; Algina, et al., 2005a, Algina, et al., 2006; Kelly, 2005). A second problem with using Cohen's $\delta$ is that, although it is intended as a measure of group separation, it is not always an adequate measure of group separation due to the fact $\delta$ can be dramatically affected by outliers and long-tailed distributions (Keselman & Wilcox, 2003). Algina, et al. (2005b) recommended a robust version of Cohen's $\delta$ defined by

$$\delta_R = .642\left(\frac{\mu_{t2}-\mu_{t1}}{\sigma_W}\right).$$

Algina and Keselman (2003b) and Algina, et al. (2005b) reported that CIs for $\delta_R$ have better coverage probability than do CIs for Cohen's $\delta$, and that the actual coverage probability is closer to the nominal coverage probability for CIs constructed by using the percentile bootstrap than for the CIs constructed by using the noncentral $t$ distribution-based method.

In the more than two group cases, Zhang and Algina (2008) examined the coverage performance of the CIs for the Root Mean Square Standardized Effect (RMSSE, $f^*$) proposed by Steiger and Fouladi (1997), which is one of the generalized ES measures in ANOVA. The findings of their study indicated that the coverage probabilities of the CIs for $f^*$ were not adequate under data nonnormality. This is not surprising because $f^*$ is formulated with least-square parameters which are affected by skewed data, long tails and/or outlying values.

This study proposed a robust version of $f^*$, $f_R^*$, by substituting robust estimators, i.e., trimmed means and Winsorized variances, for the least-square values. The coverage performances of the noncentral $F$ distribution-based and the percentile bootstrap CIs for $f_R^*$ were examined in this investigation.

Table 1: Estimated Coverage Probabilities for Nominal 95% Noncentral F Distribution-Based (NCF) and Percentile Bootstrap (Boot) CIs for $f_R^*$: J = 3, Equally Spaced Mean Configuration

| $f_R^*$ | $n$ | Normal | | $g = .000$<br>$h = .109$ | | $g = .000$<br>$h = .225$ | | $g = .760$<br>$h = -.098$ | | $g = .225$<br>$h = .225$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | NCF | Boot | NCF | Boot | NCF | Boot | NCF | Boot | NCF | Boot |
| .00 | 20 | **.973** | **.994** | **.968** | **.993** | **.977** | **.994** | **.975** | **.992** | **.975** | **.994** |
| | 35 | **.968** | **.990** | **.977** | **.993** | **.972** | **.990** | **.979** | **.993** | **.974** | **.996** |
| | 50 | **.972** | **.993** | **.976** | **.994** | **.974** | **.992** | **.977** | **.992** | **.974** | **.990** |
| .10 | 20 | .943 | **.991** | .951 | **.994** | .948 | **.993** | .953 | **.990** | .955 | **.993** |
| | 35 | .957 | **.987** | .951 | **.989** | .954 | **.988** | .949 | **.988** | .954 | **.990** |
| | 50 | .943 | **.983** | .956 | **.988** | .954 | **.988** | .958 | **.988** | .952 | **.989** |
| .25 | 20 | .942 | **.981** | .952 | **.993** | .947 | **.989** | **.961** | **.988** | .943 | **.988** |
| | 35 | .945 | **.981** | .950 | **.982** | .946 | **.981** | .954 | **.984** | .952 | **.986** |
| | 50 | .953 | **.978** | .940 | **.981** | .951 | **.987** | .950 | **.978** | .945 | **.982** |
| .40 | 20 | .942 | **.976** | .951 | **.988** | .945 | **.990** | .940 | **.980** | **.939** | **.991** |
| | 35 | .954 | **.970** | **.934** | **.964** | .944 | **.969** | .949 | **.974** | .954 | **.978** |
| | 50 | .950 | **.961** | .943 | **.962** | **.935** | .960 | **.939** | **.961** | .949 | **.968** |
| .55 | 20 | .943 | **.973** | **.939** | **.978** | **.937** | **.979** | **.932** | **.972** | **.938** | **.977** |
| | 35 | .944 | **.968** | .946 | **.963** | .940 | **.969** | **.924** | .960 | .940 | **.966** |
| | 50 | .947 | .960 | **.934** | .952 | .940 | .959 | **.929** | **.963** | **.929** | .958 |
| .70 | 20 | .945 | **.968** | .940 | **.972** | **.938** | **.977** | **.916** | **.973** | **.936** | **.980** |
| | 35 | **.935** | .958 | .944 | **.964** | **.935** | **.965** | **.923** | **.969** | **.924** | **.964** |
| | 50 | .942 | **.962** | .944 | **.967** | **.928** | **.967** | **.923** | **.968** | **.935** | **.963** |

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 $f_R^*$: J = 3, One Extreme Mean Configuration

| $f_R^*$ | $n$ | Normal | | $g = .000$ <br> $h = .109$ | | $g = .000$ <br> $h = .225$ | | $g = .760$ <br> $h = -.098$ | | $g = .225$ <br> $h = .225$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | NCF | Boot | NCF | Boot | NCF | Boot | NCF | Boot | NCF | Boot |
| .00 | 20 | **.973** | **.994** | **.972** | **.992** | **.974** | **.996** | **.983** | **.996** | **.972** | **.992** |
| | 35 | **.981** | **.993** | **.979** | **.993** | **.975** | **.991** | **.980** | **.991** | **.969** | **.991** |
| | 50 | **.974** | **.991** | **.976** | **.994** | **.973** | **.991** | **.980** | **.995** | **.970** | **.993** |
| .10 | 20 | .952 | **.990** | .944 | **.991** | .946 | **.991** | .951 | **.990** | .953 | **.994** |
| | 35 | .949 | **.986** | .948 | **.990** | .944 | **.990** | **.938** | **.986** | .944 | **.987** |
| | 50 | .945 | **.987** | .949 | **.986** | .948 | **.990** | .958 | **.986** | .956 | **.987** |
| .25 | 20 | .946 | **.982** | .945 | **.985** | .952 | **.987** | .954 | **.991** | .954 | **.990** |
| | 35 | .952 | **.984** | .942 | **.982** | .950 | **.986** | .948 | **.981** | .947 | **.986** |
| | 50 | **.938** | **.976** | .953 | **.981** | .951 | **.982** | .950 | **.984** | .949 | **.983** |
| .40 | 20 | .949 | **.980** | .943 | **.984** | .942 | **.988** | **.938** | **.984** | .942 | **.992** |
| | 35 | .943 | **.966** | .949 | **.972** | .946 | **.976** | .948 | **.972** | .948 | **.973** |
| | 50 | .952 | **.962** | .950 | **.964** | .952 | **.970** | .946 | **.965** | .945 | **.966** |
| .55 | 20 | .952 | **.975** | .943 | **.975** | .941 | **.980** | .940 | **.981** | .943 | **.984** |
| | 35 | .943 | .958 | .947 | **.966** | **.938** | **.963** | **.936** | **.968** | .943 | **.970** |
| | 50 | .942 | **.961** | **.931** | .953 | .943 | **.962** | **.935** | **.964** | **.934** | .958 |
| .70 | 20 | .944 | **.970** | **.937** | **.976** | **.931** | **.972** | **.930** | **.982** | **.936** | **.982** |
| | 35 | .941 | .960 | **.938** | **.964** | **.932** | **.965** | **.924** | **.966** | **.934** | **.966** |
| | 50 | **.939** | .957 | .940 | **.962** | **.938** | **.966** | **.932** | **.965** | **.935** | **.967** |

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 $f_R^*$: J = 6, Equally Spaced Mean Configuration

| $f_R^*$ | $n$ | Normal | | $g = .000$<br>$h = .109$ | | $g = .000$<br>$h = .225$ | | $g = .760$<br>$h = -.098$ | | $g = .225$<br>$h = .225$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | NCF | Boot | NCF | Boot | NCF | Boot | NCF | Boot | NCF | Boot |
| .00 | 20 | **.976** | **.978** | **.977** | **.982** | **.976** | **.986** | **.980** | **.983** | **.978** | **.986** |
| | 35 | **.972** | **.969** | **.969** | **.972** | **.969** | **.974** | **.979** | **.977** | **.977** | **.974** |
| | 50 | **.974** | **.966** | **.972** | **.965** | **.975** | **.974** | **.979** | **.970** | **.975** | **.972** |
| .10 | 20 | .950 | **.977** | .950 | **.974** | .948 | **.979** | .947 | **.980** | .944 | **.982** |
| | 35 | .948 | **.963** | .952 | **.969** | .950 | **.970** | .951 | **.974** | .951 | **.972** |
| | 50 | .948 | **.962** | .952 | **.965** | .954 | **.971** | .944 | **.967** | .947 | **.966** |
| .25 | 20 | .945 | **.966** | .947 | **.975** | **.938** | **.983** | .952 | **.982** | .943 | **.983** |
| | 35 | .943 | .960 | .947 | **.970** | .948 | **.981** | .942 | **.969** | .950 | **.978** |
| | 50 | .950 | **.962** | .940 | **.961** | .944 | **.970** | .944 | **.970** | .944 | **.969** |
| .40 | 20 | .947 | **.972** | .943 | **.980** | .948 | **.990** | **.938** | **.981** | **.935** | **.989** |
| | 35 | .946 | **.968** | .946 | **.971** | .945 | **.978** | **.938** | **.972** | .942 | **.976** |
| | 50 | .954 | **.968** | .944 | **.966** | .942 | **.971** | **.936** | **.969** | .944 | **.970** |
| .55 | 20 | .949 | **.974** | **.939** | **.980** | **.936** | **.986** | **.926** | **.985** | **.936** | **.989** |
| | 35 | .955 | **.971** | .947 | **.973** | .942 | **.973** | **.915** | **.969** | **.934** | **.982** |
| | 50 | .946 | **.962** | .943 | **.967** | .948 | **.973** | **.927** | **.969** | **.928** | **.967** |
| .70 | 20 | .949 | **.980** | **.938** | **.984** | **.930** | **.983** | **.902** | **.984** | **.939** | **.988** |
| | 35 | .943 | **.967** | **.934** | **.970** | **.921** | **.968** | **.907** | **.972** | **.921** | **.971** |
| | 50 | .941 | **.962** | .944 | **.972** | **.933** | **.968** | **.914** | **.970** | **.928** | **.966** |

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 (NCF) and Percentile Bootstrap (Boot) CIs for $f_R^*$: J = 6, One Extreme Mean Configuration

| $f_R^*$ | $n$ | Normal | | $g = .000$<br>$h = .109$ | | $g = .000$<br>$h = .225$ | | $g = .760$<br>$h = -.098$ | | $g = .225$<br>$h = .225$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | NCF | Boot | NCF | Boot | NCF | Boot | NCF | Boot | NCF | Boot |
| .00 | 20 | **.972** | **.976** | **.979** | **.981** | **.977** | **.983** | **.974** | **.984** | **.977** | **.989** |
| | 35 | **.974** | **.969** | **.971** | **.968** | **.976** | **.976** | **.975** | **.972** | **.978** | **.976** |
| | 50 | **.978** | **.973** | **.976** | **.970** | **.977** | **.974** | **.982** | **.976** | **.971** | **.965** |
| .10 | 20 | **.962** | **.976** | .945 | **.975** | .953 | **.982** | .943 | **.978** | .954 | **.981** |
| | 35 | .954 | **.971** | .955 | **.970** | .955 | **.969** | .945 | **.962** | .948 | **.975** |
| | 50 | .948 | .960 | .948 | **.965** | .951 | **.964** | .954 | **.971** | .956 | **.972** |
| .25 | 20 | .951 | **.974** | **.937** | **.974** | .954 | **.987** | .949 | **.978** | .951 | **.988** |
| | 35 | .952 | **.967** | .950 | **.976** | .945 | **.974** | .955 | **.973** | .951 | **.974** |
| | 50 | .953 | **.965** | .946 | **.961** | .943 | **.970** | .948 | **.970** | .951 | **.973** |
| .40 | 20 | .945 | **.972** | .950 | **.984** | .945 | **.983** | .952 | **.983** | **.938** | **.988** |
| | 35 | **.939** | .958 | .944 | **.969** | .942 | **.977** | .952 | **.979** | **.938** | **.972** |
| | 50 | .941 | .956 | .945 | **.968** | **.936** | **.971** | .951 | **.975** | .945 | **.975** |
| .55 | 20 | .944 | **.976** | .943 | **.982** | **.936** | **.987** | .942 | **.989** | **.938** | **.988** |
| | 35 | .949 | **.970** | .940 | **.973** | **.934** | **.970** | **.937** | **.981** | **.935** | **.970** |
| | 50 | .950 | **.963** | **.939** | **.961** | **.932** | **.967** | **.927** | **.968** | **.931** | **.966** |
| .70 | 20 | **.935** | **.972** | **.938** | **.982** | **.929** | **.985** | **.925** | **.992** | **.928** | **.987** |
| | 35 | .946 | **.970** | **.936** | **.972** | **.917** | **.964** | **.920** | **.973** | **.926** | **.969** |
| | 50 | .946 | **.966** | **.932** | **.961** | **.930** | **.972** | **.908** | **.961** | **.930** | **.968** |

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 $f_R^*$: J=3, Equally Spaced Mean Configuration

| $f_R^*$ | $n$ | Normal | | $g = .000$<br>$h = .109$ | | $g = .000$<br>$h = .225$ | | $g = .760$<br>$h = -.098$ | | $g = .225$<br>$h = .225$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | NCF | Boot | NCF | Boot | NCF | Boot | NCF | Boot | NCF | Boot |
| .00 | 20 | .491 | .618 | .494 | .611 | .488 | .596 | .490 | .601 | .487 | .595 |
| | 35 | .366 | .437 | .361 | .430 | .365 | .428 | .364 | .425 | .368 | .428 |
| | 50 | .301 | .354 | .304 | .354 | .306 | .353 | .305 | .351 | .304 | .352 |
| .10 | 20 | .512 | .637 | .509 | .625 | .515 | .618 | .514 | .625 | .515 | .617 |
| | 35 | .397 | .461 | .390 | .453 | .395 | .450 | .396 | .453 | .386 | .444 |
| | 50 | .332 | .375 | .334 | .378 | .332 | .374 | .335 | .376 | .332 | .374 |
| .25 | 20 | .609 | .725 | .607 | .714 | .601 | .701 | .615 | .725 | .604 | .703 |
| | 35 | .487 | .550 | .485 | .540 | .480 | .533 | .485 | .540 | .481 | .533 |
| | 50 | .423 | .466 | .421 | .464 | .421 | .462 | .421 | .463 | .421 | .461 |
| .40 | 20 | .702 | .828 | .702 | .823 | .692 | .813 | .698 | .840 | .699 | .818 |
| | 35 | .545 | .613 | .543 | .611 | .543 | .613 | .544 | .623 | .542 | .613 |
| | 50 | .454 | .499 | .454 | .501 | .454 | .506 | .454 | .511 | .453 | .505 |
| .55 | 20 | .760 | .899 | .758 | .905 | .755 | .906 | .755 | .937 | .756 | .914 |
| | 35 | .562 | .625 | .562 | .634 | .561 | .641 | .562 | .662 | .561 | .647 |
| | 50 | .463 | .501 | .463 | .508 | .463 | .513 | .463 | .532 | .463 | .518 |
| .70 | 20 | .794 | .947 | .791 | .961 | .789 | .980 | .793 | 1.026 | .789 | .987 |
| | 35 | .579 | .644 | .578 | .655 | .577 | .668 | .579 | .698 | .578 | .679 |
| | 50 | .479 | .517 | .478 | .526 | .477 | .536 | .478 | .565 | .478 | .545 |

Note: Results are based on 2,500 replications.

Table 6: Average Widths Of Noncentral F Distribution-Based (NCF) and Percentile Bootstrap (Boot) CIs for $f_R^*$: J=3, One Extreme Mean Configuration

| $f_R^*$ | $n$ | Normal | | $g=.000$<br>$h=.109$ | | $g=.000$<br>$h=.225$ | | $g=.760$<br>$h=-.098$ | | $g=.225$<br>$h=.225$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | NCF | Boot | NCF | Boot | NCF | Boot | NCF | Boot | NCF | Boot |
| .00 | 20 | .487 | .616 | .497 | .613 | .489 | .598 | .484 | .596 | .494 | .598 |
| | 35 | .366 | .436 | .364 | .432 | .367 | .430 | .363 | .425 | .370 | .430 |
| | 50 | .304 | .357 | .309 | .357 | .305 | .352 | .305 | .350 | .303 | .351 |
| .10 | 20 | .518 | .638 | .517 | .631 | .515 | .620 | .510 | .619 | .510 | .613 |
| | 35 | .391 | .457 | .390 | .452 | .390 | .450 | .391 | .450 | .392 | .450 |
| | 50 | .334 | .381 | .333 | .378 | .334 | .376 | .333 | .373 | .337 | .378 |
| .25 | 20 | .611 | .726 | .607 | .715 | .608 | .709 | .606 | .715 | .598 | .696 |
| | 35 | .485 | .546 | .486 | .543 | .484 | .539 | .485 | .542 | .485 | .538 |
| | 50 | .420 | .464 | .422 | .464 | .421 | .461 | .422 | .462 | .423 | .461 |
| .40 | 20 | .705 | .834 | .696 | .815 | .697 | .813 | .700 | .833 | .695 | .813 |
| | 35 | .544 | .612 | .544 | .611 | .543 | .614 | .544 | .614 | .542 | .610 |
| | 50 | .454 | .501 | .454 | .501 | .454 | .503 | .454 | .503 | .454 | .503 |
| .55 | 20 | .760 | .899 | .756 | .901 | .756 | .912 | .759 | .926 | .755 | .910 |
| | 35 | .562 | .627 | .561 | .635 | .561 | .642 | .563 | .652 | .561 | .642 |
| | 50 | .463 | .501 | .463 | .508 | .463 | .513 | .463 | .520 | .462 | .514 |
| .70 | 20 | .793 | .942 | .791 | .956 | .790 | .979 | .792 | 1.006 | .789 | .982 |
| | 35 | .580 | .646 | .578 | .652 | .579 | .671 | .580 | .692 | .578 | .670 |
| | 50 | .478 | .518 | .478 | .526 | .478 | .539 | .478 | .557 | .478 | .540 |

Note: Results are based on 2,500 replications.

Table 7: Average Widths of Noncentral F Distribution-Based (NCF) and Percentile Bootstrap (Boot) CIs for $f_R^*$: J = 6, Equally Spaced Mean Configuration

| $f_R^*$ | $n$ | Normal | | $g = .000$<br>$h = .109$ | | $g = .000$<br>$h = .225$ | | $g = .760$<br>$h = -.098$ | | $g = .225$<br>$h = .225$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | NCF | Boot | NCF | Boot | NCF | Boot | NCF | Boot | NCF | Boot |
| .00 | 20 | .352 | .487 | .356 | .484 | .345 | .474 | .345 | .480 | .348 | .475 |
| | 35 | .259 | .351 | .261 | .350 | .262 | .348 | .256 | .346 | .260 | .346 |
| | 50 | .218 | .289 | .219 | .289 | .216 | .286 | .215 | .286 | .219 | .287 |
| .10 | 20 | .372 | .499 | .379 | .501 | .375 | .493 | .374 | .499 | .373 | .491 |
| | 35 | .294 | .370 | .290 | .367 | .294 | .368 | .290 | .367 | .290 | .364 |
| | 50 | .253 | .310 | .252 | .308 | .251 | .307 | .251 | .306 | .249 | .305 |
| .25 | 20 | .464 | .559 | .460 | .554 | .459 | .551 | .463 | .561 | .457 | .550 |
| | 35 | .363 | .406 | .362 | .404 | .361 | .405 | .360 | .406 | .360 | .405 |
| | 50 | .302 | .326 | .301 | .326 | .301 | .326 | .301 | .328 | .301 | .326 |
| .40 | 20 | .506 | .577 | .505 | .582 | .503 | .590 | .503 | .598 | .502 | .591 |
| | 35 | .361 | .389 | .362 | .392 | .361 | .396 | .361 | .405 | .361 | .398 |
| | 50 | .293 | .309 | .293 | .311 | .293 | .314 | .293 | .321 | .293 | .315 |
| .55 | 20 | .506 | .572 | .506 | .583 | .506 | .600 | .506 | .619 | .507 | .610 |
| | 35 | .363 | .392 | .363 | .399 | .363 | .405 | .364 | .422 | .363 | .411 |
| | 50 | .299 | .318 | .299 | .323 | .299 | .328 | .299 | .340 | .298 | .331 |
| .70 | 20 | .517 | .590 | .516 | .605 | .516 | .629 | .517 | .666 | .516 | .641 |
| | 35 | .377 | .412 | .376 | .421 | .376 | .433 | .377 | .461 | .376 | .442 |
| | 50 | .311 | .336 | .311 | .342 | .310 | .350 | .311 | .372 | .310 | .356 |

Note: Results are based on 2,500 replications.

Table 8: Average Widths of Noncentral F Distribution-Based (NCF) and Percentile Bootstrap (Boot) CIs for $f_R^*$ : J = 6, One Extreme Mean Configuration

| $f_R^*$ | $n$ | Normal | | $g = .000$<br>$h = .109$ | | $g = .000$<br>$h = .225$ | | $g = .760$<br>$h = -.098$ | | $g = .225$<br>$h = .225$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | NCF | Boot | NCF | Boot | NCF | Boot | NCF | Boot | NCF | Boot |
| .00 | 20 | .353 | .488 | .348 | .482 | .350 | .477 | .349 | .484 | .346 | .476 |
| | 35 | .259 | .350 | .260 | .349 | .260 | .347 | .258 | .348 | .259 | .346 |
| | 50 | .219 | .290 | .215 | .287 | .218 | .288 | .215 | .286 | .217 | .287 |
| .10 | 20 | .381 | .504 | .377 | .499 | .375 | .492 | .377 | .502 | .379 | .494 |
| | 35 | .295 | .372 | .294 | .369 | .292 | .365 | .292 | .367 | .293 | .366 |
| | 50 | .251 | .308 | .252 | .308 | .251 | .306 | .252 | .307 | .251 | .306 |
| .25 | 20 | .463 | .558 | .459 | .553 | .460 | .552 | .462 | .558 | .461 | .553 |
| | 35 | .362 | .406 | .361 | .404 | .361 | .405 | .363 | .401 | .361 | .403 |
| | 50 | .302 | .325 | .301 | .325 | .301 | .326 | .303 | .321 | .302 | .325 |
| .40 | 20 | .505 | .576 | .505 | .582 | .505 | .589 | .506 | .583 | .504 | .589 |
| | 35 | .361 | .388 | .362 | .393 | .362 | .398 | .361 | .389 | .361 | .395 |
| | 50 | .293 | .308 | .293 | .311 | .293 | .315 | .293 | .312 | .293 | .312 |
| .55 | 20 | .506 | .570 | .505 | .584 | .506 | .599 | .506 | .600 | .506 | .603 |
| | 35 | .363 | .393 | .363 | .399 | .363 | .405 | .363 | .411 | .363 | .407 |
| | 50 | .299 | .317 | .299 | .321 | .299 | .327 | .299 | .335 | .298 | .328 |
| .70 | 20 | .518 | .591 | .517 | .607 | .516 | .634 | .517 | .643 | .516 | .640 |
| | 35 | .376 | .410 | .376 | .423 | .376 | .433 | .376 | .452 | .376 | .440 |
| | 50 | .311 | .334 | .310 | .341 | .310 | .349 | .311 | .368 | .310 | .353 |

Note: Results are based on 2,500 replications.

Comparisons were made to the CIs for $f_R^*$ constructed by using the noncentral $F$ distribution-based and the bootstrap methods in terms of the probability coverage and interval width.

The robustness of the CIs for $f_R^*$ was investigated in a one-way, fixed-effects, between-subjects ANOVA. The study conditions incorporated five population distributions including the normal distribution and four additional cases from the family of the $g$ and $h$ distributions that are nonnormal; two number of levels for the number of treatment groups: $J = 3$ and $J = 6$; three cell sample sizes in each treatment ($n$ = 20, 35 and 50); six values of population $\text{RMSSE}_\text{R}$ (0.00, 0.10, 0.25, 0.40, 0.55 and 0.70); and two mean configurations: the equally spaced mean configuration and the one extreme mean configuration. The nominal confidence level for all intervals investigated was 0.95 and each condition was replicated 2,500 times. The number of bootstrap replications in the bootstrap procedure was 1,000.

The results indicated that the coverage probabilities of the noncentral $F$ distribution-based CIs for $f_R^*$ introduced in this study, which was formulated with robust trimmed means and Winsorized variances, were generally adequate, that is, generally either within our lenient criterion of robustness [.925, .975], or both the lenient criterion of robustness and the strict criterion interval [.94, .96]. There were only a few cases in which the noncentral $F$ distribution-based CIs for $f_R^*$ broke down. These include some cases of $f_R^* = 0$, and when $f_R^* = .70$ for small sample sizes under nonnormal data distributions, especially under the $g = .760$, $h = -.098$ distribution.

For the bootstrap CIs for $f_R^*$, the probability coverage were not adequate when $J = 3$ and $f_R^* \leq .25$ or when $J = 6$ and sample size was small, especially when sample size was 20. In particular, when $J = 3$, over half of the estimated coverage probabilities were outside of the [.925, .975] interval. These probability coverages mostly occurred when $f_R^* \leq .25$. When $J = 6$, the bootstrap CIs were mostly inside the [.925, .975] criterion interval under normality. However, under all other data distribution conditions, they were outside of the interval when sample size was small: most cases for $n = 20$ as well as some cases for $n = 35$.

For both the noncentral $F$ distribution-based and the bootstrap CIs for $f_R^*$, the mean configuration did not appear to alter the pattern of the probability coverage performance. However, sample sizes seem to be slightly positively related to probability coverage. The widths of the noncentral $F$ distribution-based CIs for $f_R^*$ were shorter than those of the bootstrap CIs under the same condition. Therefore, not only does the noncentral $F$ distribution-based CI for $f_R^*$ have better coverage probability than the bootstrap CIs for $f_R^*$, they are also narrower than those of the bootstrap CI. Both the widths of the noncentral $F$ distribution-based and bootstrap CIs for $f_R^*$ remained relatively unchanged across data distributions. In other words, the widths of the bootstrap CIs for $f_R^*$ fluctuated very little across data distribution conditions.

For both the noncentral $F$ distribution-based and the bootstrap CIs for $f_R^*$, as the number of levels of $J$ increases, the width of the estimated CIs becomes narrower. For both the noncentral $F$ distribution-based and the bootstrap CIs for $f_R^*$, under the same condition, the average width of the CIs becomes narrower as the sample size increases and the population effect size $f_R^*$ decreases.

In summary, both the noncentral $F$ distribution-based and the bootstrap CIs for $f^*$, which are based on the usual 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. The noncentral $F$ distribution-based CIs for $f_R^*$, which was proposed in the current study and was formulated with the robust parameters including the trimmed means and Winsorized variances, yielded fairly adequate coverage probabilities and better coverage probability than the percentile bootstrap CI. Accordingly, researchers who want to set a CI for $f_R^*$ can use the CI constructed by using the noncentral $F$ distribution and will enjoy the additional benefit of using a robust measure of effect size, that is, 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. (2006). Confidence interval coverage for Cohen's effect size Statistic. *Educational and Psychological Measurement*, *66*, 945-960.

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.

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

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

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

Cohen, J. (1994). The earth is round (p < 0.05). *American Psychologist*, *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*. New York: Chapman and Hall.

Hays, W. L. (1963). *Statistics*. New York: Holt, Rinehart & Winston.

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

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.

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.

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

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.

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

Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing* (*2nd Ed.*). San Diego: Elsevier Academic Press.

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

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

Zhang, G. (2009). t Test: The good, the bad, the ugly, & the remedy. *Middle Grades Research Journal*, *4*(*2*), 25-34.

Zhang, G., & Algina, J. (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. *Journal of Modern Applied Statistical Methods*, *7*(*1*), 56-76.

Yuen, K. K. (1974). The two-sample trimmed t for unequal population variances. *Biometrika*, *61*, 165-170.