Journal of Modern Applied Statistical Methods

Volume 5 | Issue 1 Article 2

5-1-2006

# Confidence Intervals For An Effect Size When Variances Are Not Equal

James Algina
*University of Florida*, algina@ufl.edu

H. J. Keselman
*University of Manitoba*, kesel@ms.umanitoba.ca

Randall D. Penfield
*University of Miami*, penfield@miami.edu

 Part of the Applied Statistics Commons, Social and Behavioral Sciences Commons, and the Statistical Theory Commons

Recommended Citation

Algina, James; Keselman, H. J.; and Penfield, Randall D. (2006) "Confidence Intervals For An Effect Size When Variances Are Not Equal," *Journal of Modern Applied Statistical Methods*: Vol. 5 : Iss. 1 , Article 2.

Journal of Modern Applied Statistical Methods
May, 2006, Vol. 5, No. 1, 2- 13
Copyright © 2006 JMASM, Inc.
1538 – 9472/06/$95.00

*Invited Articles*

# Confidence Intervals For An Effect Size When Variances Are Not Equal

James Algina
University of Florida

H. J. Keselman
University of Manitoba

Randall D. Penfield
University of Miami

Confidence intervals must be robust in having nominal and actual probability coverage in close agreement. This article examined two ways of computing an effect size in a two-group problem: (a) the classic approach which divides the mean difference by a single standard deviation and (b) a variant of a method which replaces least squares values with robust trimmed means and a Winsorized variance. Confidence intervals were determined with theoretical and bootstrap critical values. Only the method that used robust estimators and a bootstrap critical value provided generally accurate probability coverage under conditions of nonnormality and variance heterogeneity in balanced as well as unbalanced designs.

Key words: Effect size, confidence interval, trimmed means, Winsorized variance, noncentral distribution

## Introduction

Estimating effect size (ES) and setting intervals for such estimates has become a requirement in many scientific journals as a result of the American Psychological Association's (APA) Task Force on Statistical Inference (Wilkinson & APA Task Force on Statistical Inference, 1999). Indeed, according to Thompson (2003, personal communication) at least 23 journals require authors to follow the recommendation put forth by the task force.

James Algina (algina@ufl.edu) is Professor of Educational Psychology. His research interests are in applied statistics and psychometrics. H. J. Keselman (kesel@ms.umanitoba.ca) is Professor of Psychology. His research interests are in applied statistics. Randall D. Penfield (penfield@miami.edu) is Assistant Professor of Education. His research interests are in educational measurement and psychometrics.

Not surprisingly, there has been a renewed interest in ES estimates and accompanying confidence intervals (CIs). See, for example, Algina and Keselman (2003), Bird (2002), Cumming and Finch (2001), and Steiger and Fouladi (1997).

Glass (1976) used a control group standard deviation (in a two-group problem) to standardize the difference between the group means. However, other values have been used to standardize the mean difference. For example, Hedges (1981) used the square root of the pooled variance, which is referred to as the pooled standard deviation. If the variance equality assumption is not met, then the standard deviation for either one of the groups could be used as the standardizer. In the context of comparing an experimental and control treatment, Glass, McGaw, and Smith (1981) recommended using the standard deviation for the control group, but pointed out that the experimental group standard deviation could be used. Glass et al. (1981) presented an example demonstrating that the value of the ES estimate

can vary depending on which group's standard deviation is used as the standardizer. As well, they point out that both ES estimates would be correct. As Glass et al. (1981) noted, "These facts are not contradictory; they are two distinct features of a finding which cannot be expressed by one number" (p 107).

Thus, Olejnik and Algina (2000) noted that when the equality of variance assumption is violated, the researcher will have to select one standard deviation that expresses the contrast (i.e., the effect) on the scale the researcher imagines is most important, or will have to report the mean difference standardized by several standard deviations and discuss the implications of these ESs. Before turning to methods that can be used when variances appear to be heterogeneous, it is important to point out that heterogeneity of variance can occur due to some additional factor in the data that is not modeled in the analysis. It is better to model such factors than to uncritically use methods that are appropriate for heterogeneous variances.

When the population variances are assumed to be equal for the two levels of the factor, the population ES (PES) is

$$\delta_{Pooled} = \frac{\mu_2 - \mu_1}{\sigma}$$

where $\mu_j$ is the population mean for level *j* and $\sigma$ is the population standard deviation, which is assumed to be equal for the two levels of the factor. The PES can be estimated by

$$\hat{\delta}_{Pooled} = \frac{\bar{Y}_2 - \bar{Y}_1}{S_{Pooled}}$$

where $\bar{Y}_j$ $(j = 1,2)$ is a treatment level group mean, $n_j$ $(n_1 + n_2 = N)$ is the sample size for the jth group, and $S_{Pooled}$ is the pooled standard deviation.

According to Steiger and Fouladi (1997), a CI for the PES, which is exact under the assumptions for the independent samples t test, can be derived by using the noncentral t distribution with N – 2 degrees of freedom. First, a CI for the noncentrality parameter

$$\lambda = \sqrt{\frac{n_1 n_2}{n_1 + n_2}}\left(\frac{\mu_2 - \mu_1}{\sigma}\right) = \sqrt{\frac{n_1 n_2}{n_1 + n_2}}\,\delta_{Pooled}$$

is obtained. Then, by multiplying the limits of the interval for $\lambda$ by the inverse of

$$\sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

a CI for $\delta_{Pooled}$ is obtained. The lower limit of the CI for $\lambda$ is the noncentrality parameter for the noncentral t distribution in which the calculated t statistic

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

is the $1 - \alpha/2$ quantile. For example, if $t = 2.131$ and $N - 2 = 15$, the lower limit of the 95% CI for $\lambda$ is zero, because 2.131 is the .975 quantile of the t distribution with a noncentrality parameter equal to zero. The upper limit of the $100(1-\alpha/2)\%$ interval for $\lambda$ is the noncentrality parameter for the noncentral t distribution in which the calculated t statistic is the $\alpha/2$ quantile of the distribution (See Steiger & Fouladi, 1997).

The PES based on the standard deviation for the jth group is

$$\delta_j = \frac{\mu_2 - \mu_1}{\sigma_j}$$

and can be estimated by

$$\hat{\delta}_j = \frac{\bar{Y}_2 - \bar{Y}_1}{S_j}$$

where $S_j$ is the square root of the usual unbiased sample variance. With this ES, the noncentral t-based interval for $\delta$ is no longer correct. However, under the assumptions that the data in each group are normally distributed and all data are distributed independently, a

noncentral t-based approximate CI for $\delta_j$ can be derived. Thus, the CI does not assume equal variances, but the interval is based on normal distribution theory. This normality assumption is likely to be problematic because $\bar{Y}_2 - \bar{Y}_1$ and $S_j$ are not distributed independently when the distribution is skewed for the *j*th treatment. For example, if the distribution is positively skewed for the first treatment, the sampling correlation between $\bar{Y}_2 - \bar{Y}_1$ and $S_1$ will be negative.

Therefore, large values for $\bar{Y}_2 - \bar{Y}_1$ will tend to be associated with small values for $S_1$ and $\hat{\delta}_1$ will tend to be positively biased. Moreover, the distribution theory used in deriving the CI will no longer apply. As a result the CI may not have the correct probability coverage. In fact, in an investigation of CIs for ESs in *dependent* samples designs, Algina, Keselman, and Penfield (2005a) showed that nonnormality has a negative impact on coverage probability for a noncentral t based approximate CI for $\delta_j$.

Purposes of this article

Therefore, one purpose of the research was to investigate coverage probability for the noncentral t-based CI for $\delta_j$ when data are sampled in an *independent* samples design from a nonnormal distribution. Considering the prediction that the noncentral t-based CI for $\delta_j$ is likely to be negatively impacted by nonnormality, a second purpose of the article was to investigate alternatives to the interval.

One reasonable alternative is to use the percentile bootstrap to construct a CI for $\delta_j$. A second alternative is to replace the least squares estimates in $\hat{\delta}_j$ with robust estimates. This approach was recommended by Algina et al. (2005a) in the context of CIs for $\delta_j$ in repeated measures designs and by Algina, Keselman, and Penfield (2005b) in the context of CIs for $\delta$ in independent samples and is consistent with the observation in Wilcox and Keselman (2003) that the common population definition and sample estimate of ES (i.e., $\delta_{Pooled}$ and $\hat{\delta}_{Pooled}$ or $\delta_j$ and $\hat{\delta}_j$ for the two-group problem), based on least squares estimators, are not robust to distribution shape. That is, skewed distributions and distributions containing outliers can cause the PES value and its estimate to be grossly misleading (Wilcox, 2003, Sec 8.11). Accordingly, in place of $\hat{\delta}_j$, the following is used

$$\hat{\delta}_{R_j} = .642\left(\frac{\bar{Y}_{t2} - \bar{Y}_{t1}}{S_{W_j}}\right) \qquad (1)$$

where $\bar{Y}_{tj}$ is the 20% trimmed mean for the *j*th group $(j = 1, 2)$ and $S^2_{W_j}$ is the 20% Winsorized variance for group *j*. Twenty percent refers to the percentage trimmed from each tail. The constant .642 is the population value for the Winsorized standard deviation for a standard normal distribution for 20% trimming. (See Wilcox, 2003, for a justification of 20% trimming and computational definitions of the trimmed mean and Winsorized variance). For a normal distribution, both $\hat{\delta}_{R_j}$ and $\hat{\delta}_j$ converge to $\delta_j$ as the sample sizes increase. Probability coverage for a noncentral t-based CI and for a percentile bootstrap CI for $\delta_{R_j}$ was investigated (defined later in equation (2)).

A Noncentral t-Based CI for $\delta_j$

If the variances are unequal, in a two-group independent samples design, the population and sample ES is defined as

$$\delta_1 = \frac{\mu_2 - \mu_1}{\sigma_1}$$

and

$$\hat{\delta}_1 = \frac{\bar{Y}_2 - \bar{Y}_1}{S_1},$$

respectively. (The standard deviation for the second group could also be used. Glass et al. (1981) pointed out that these ESs provide different information.)

It is well known that if $U \sim N(\mu,1)$, $V \sim \chi^2(k)$, and $U$ and $V$ are independently distributed, then

$$\frac{U}{\sqrt{\frac{V}{k}}} \sim t(k,\mu)$$

where $t(k,\mu)$ is the noncentral t distribution with degrees of freedom k and noncentrality parameter μ. Using this result with

$$U = \frac{\bar{Y}_2 - \bar{Y}_1}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

and

$$V = \frac{(n_1 - 1)S_i^2}{\sigma_1^2}$$

then

$$\frac{\frac{\bar{Y}_2 - \bar{Y}_1}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}}{\sqrt{\frac{S_1^2}{\sigma_1^2}}} = \frac{\bar{Y}_2 - \bar{Y}_1}{S_1\sqrt{\frac{1}{n_1} + \frac{\sigma_2^2}{n_2\sigma_1^2}}} \sim t(n_1 - 1, \lambda)$$

where

$$\lambda = \frac{\mu_2 - \mu_1}{\sigma_1\sqrt{\frac{1}{n_1} + \frac{\sigma_2^2}{n_2\sigma_1^2}}}.$$

If the estimate of $\lambda$ is calculated as

$$\hat{\lambda} = \frac{\bar{Y}_2 - \bar{Y}_1}{S_1\sqrt{\frac{1}{n_1} + \frac{S_2^2}{n_2 S_1^2}}} = \frac{\hat{\delta}_1}{\sqrt{\frac{1}{n_1} + \frac{S_2^2}{n_2 S_1^2}}}$$

the noncentral t distribution, with $n_1 - 1$ degrees of freedom, can be used to find a CI on $\lambda$. Specifically, the upper limit of a $100(1-\alpha)\%$ interval for $\lambda$ is the noncentrality parameter for the noncentral t distribution with $n_1 - 1$ degrees of freedom in which $\hat{\lambda}$ is the $\alpha/2$ quantile of the distribution; the lower limit is the noncentrality parameter for the noncentral t distribution in which $\hat{\lambda}$ is the $(1-\alpha/2)$ quantile. Then, multiplying the lower and upper limit by $\sqrt{\frac{1}{n_1} + \frac{S_2^2}{n_2 S_1^2}}$, an approximate CI for $\delta_1$ is obtained. The interval is approximate because the limits of the CI for $\lambda$ are multiplied by a random variable.

To obtain an estimate of the robust ES, let $\lfloor .2n_j \rfloor$ indicate that $.2n_j$ is rounded down to the nearest integer, $g_j = \lfloor .2n_j \rfloor$, $h_j = n_j - 2g_j$, and then let

$$\tilde{S}_j^2 = \frac{(n_j - 1)S_{W_j}^2}{h_j - 1}$$

and

$$\tilde{\sigma}_j^2 = \frac{(n_j - 1)\sigma_{W_j}^2}{h_j - 1}$$

where $\sigma_{W_j}^2$ is the population Winsorized variance for treatment *j*. To obtain a CI for

$$\delta_{R_1} = .642\left(\frac{\mu_{t2} - \mu_{t1}}{\sigma_{w_1}}\right) \qquad (2)$$

define

$$\lambda_R = \frac{\mu_{t2} - \mu_{t1}}{\tilde{\sigma}_1\sqrt{\frac{1}{h_1} + \frac{\tilde{\sigma}_2^2}{h_2\tilde{\sigma}_1^2}}} = \frac{\delta_{R_1}}{.642\sqrt{\frac{n_1 - 1}{h_1 - 1}\left(\frac{1}{h_1} + \frac{\tilde{\sigma}_2^2}{h_2\tilde{\sigma}_1^2}\right)}} \quad (3)$$

where $\mu_{tj}$ is the population trimmed mean. Also define

$$\hat{\lambda}_R = \frac{\bar{Y}_{t2} - \bar{Y}_{t1}}{\tilde{S}_1\sqrt{\frac{1}{h_1} + \frac{\tilde{S}_2^2}{h_2\tilde{S}_1^2}}} = \frac{\hat{\delta}_{R_1}}{.642\sqrt{\frac{n_1 - 1}{h_1 - 1}\left(\frac{1}{h_1} + \frac{\tilde{S}_2^2}{h_2\tilde{S}_1^2}\right)}}. \quad (4)$$

The upper limit of a $100(1-\alpha)\%$ interval for $\lambda_R$ is the noncentrality parameter for the noncentral *t* distribution, with $h_1 - 1$ degrees of freedom, in which $\hat{\lambda}_R$ is the $\alpha/2$ quantile of the distribution; the lower limit is the noncentrality parameter for the noncentral *t* distribution in which $\hat{\lambda}_R$ is the $(1-\alpha/2)$ quantile. An approximate CI for $\delta_{R_1}$ is obtained by multiplying the lower and upper limit by

$$.642\sqrt{\left(\frac{n_1 - 1}{h_1 - 1}\right)\left(\frac{1}{h_1} + \frac{\tilde{S}_2^2}{h_2\tilde{S}_1^2}\right)}.$$

The interval is approximate for two reasons. First, when trimmed means and Winsorized variances are used, there is no guarantee that the noncentral t distribution is the appropriate distribution for calculating a CI for $\lambda_R$. Second, the interval is approximate because the limits of the CI for $\lambda_R$ are multiplied by a random variable.

The investigations of these intervals were carried out in three studies.

Study 1

Methodology

Probability coverage of CIs for $\delta_1$ and $\delta_{R_1}$ based on the noncentral t distribution were investigated. It is important to recognize that $\delta_1$ and $\delta_{R_1}$ are different parameters. When applied to normal distributions, the parameters will be equal, but otherwise will most likely be unequal. Thus, there is no attempt to compare the interval estimates of the $\delta_1$ and $\delta_{R_1}$.

Probability coverage was investigated for all combinations of the following three factors: $n_1 = n_2 = 20$ to 100 in steps of 20, PESs $(\delta_1$ and $\delta_{R_1})$ ranging from 0 to 1.6 in steps of .4, and population distribution (four cases from the family of g and h distributions). The nominal confidence level for all intervals was .95 and each condition was replicated 5000 times.

The data were generated from the g and h distribution (Hoaglin, 1985). Specifically, four g and h distributions were chosen for investigation: (a) $g = h = 0$, a standard normal distribution; (b) $g = .76$ and $h = -.098$, a distribution with skew and kurtosis equal to that for an exponential distribution $(\gamma_1 = 2,\ \gamma_2 = 6)$; (c) $g = 0$ and $h = .225$, a long-tailed symmetric distribution $(\gamma_1 = 0$ and $\gamma_2 = 154.84)$; and (d) $g = .225$ and $h = .225$, a long-tailed skewed distribution ($\gamma_1 = 4.90$ and $\gamma_2 = 4673.80$). To generate data from a g and h distribution, standard unit normal variables $Z_{ij}$ were converted to g and h distributed random variables via

$$Y_{ij} = \frac{\exp(gZ_{ij}) - 1}{g}\exp\left(\frac{hZ_{ij}^2}{2}\right)$$

when both g and h were non-zero. When g was zero

$$Y_{ij} = Z_{ij}\exp\left(\frac{hZ_{ij}^2}{2}\right).$$

$Z_{ij}$ scores were generated by using RANNOR in SAS (SAS, 1999). For simulees in treatment 2, the $Y_{i2}$ scores were transformed to

$$\sqrt{PVR}(Y_{i2} - \mu_2) + \mu_2 + \sigma_1 \times \delta_1 \quad (5)$$

where PVR is the ratio of the population variance for the transformed $Y_{i2}$ scores to the variance of the $Y_{i1}$ scores and was set equal to 4 for all conditions in Study 1. The scores generated by using equation (5) were used in the

CI for $\delta_1$. Additional levels of PVR were planned for investigation. Because the results for $PVR = 4$ indicated poor probability coverage in some conditions and the focus should be to find intervals that work well in a wide variety of conditions, the intervals being estimated were dismissed.

To facilitate reporting of results for the CI for $\delta_{R_1}$, the $Y_{i2}$ scores were transformed to

$$\sqrt{PVR}\left(Y_{i2} - \mu_{t2}\right) + \mu_{t2} + \frac{\sigma_{W_1}}{.642}\delta_1 . \qquad (6)$$

This method of generating the scores in treatment 2 results in $\delta_1 = \delta_{R_1}$. The CI for $\delta_{R_1}$ was also investigated using equation (5) to generate $Y_{i2}$ scores, $\delta_1 \neq \delta_{R_1}$. The general pattern of results was the same in the two sets of conditions.

## Results

Estimated coverage probability for the two CIs are reported in Table 1 for the four g and h distributions, all sample size values, and all values of the PES (The CI for $\delta_{R_1}$ is based on $Y_{i2}$ generated by using equation (6)). The results show that both CIs had estimated probability coverage near the nominal confidence level when the data were normally distributed $(g = h = 0)$, but both could have poor probability coverage when the data were nonnormal. As the PES increased, both CIs had increasingly worse coverage probability. Coverage probability appeared to be largely unaffected by sample size.

### Study 2

Both noncentral *t*-based CIs had good coverage probability when the data were normal despite the fact that both CIs are only approximately correct. However, both could have poor coverage probability when the data were nonnormal. Therefore, the use of a percentile bootstrap CI to construct an interval on $\delta_1$ was investigated.

## Methodology

Probability coverage of a percentile bootstrap CI for all combinations of the following $n_1 = n_2 = 20$ to 100 in steps of 20, population distribution (four cases from the family of g and h distributions), and $\delta_1$ ranging from 0 to 1.6 in steps of .4 was investigated. In all conditions, $PVR = 4$. The distributions from Study 1 were investigated and the data was generated by using the procedure described for Study 1. Because a CI for $\delta_1$ was being investigated, the data for treatment 2 were generated by using Equation (5). As in Study 1, 5000 replications were conducted for each condition combination. 600 bootstrap replications were used. In all conditions, the nominal confidence level was .95.

## Results

Estimated coverage probability for the bootstrap CI for $\delta_1$ is reported in Table 2 for all sample size values and all levels of PES. The results show that the percentile CI for $\delta_1$ can have poor coverage probability and therefore should not be used. These intervals were particularly poor when the sample size was small and $\delta_1$ was large.

### Study 3

The results indicate that each of the noncentral *t*-based and percentile bootstrap CIs for $\delta_1$ and the noncentral t-based CI for $\delta_{R_1}$ can have poor coverage probability with nonnormal data. Therefore, coverage probability for a percentile bootstrap interval for $\delta_{R_1}$ was investigated.

Table 1. Estimated Coverage Probabilities for Noncentral t Distribution-Based CIs for $\delta_1$ and $\delta_{R_1}$

| | | $g = .000$, $h = .000$ | | $g = .000$, $h = .225$ | | $g = .760$, $h = -.098$ | | $g = .225$, $h = .225$ | |
|---|---|---|---|---|---|---|---|---|---|
| | | $\delta_1$ | $\delta_{R_1}$ | $\delta_1$ | $\delta_{R_1}$ | $\delta_1$ | $\delta_{R_1}$ | $\delta_1$ | $\delta_{R_1}$ |
| 0.00 | 20 | .954 | .955 | .954 | .954 | .943 | .949 | .956 | .962 |
| | 40 | .959 | .955 | .955 | .954 | .948 | .951 | .957 | .957 |
| | 60 | .954 | .957 | .956 | .955 | .947 | .950 | .954 | .958 |
| | 80 | .953 | .954 | .952 | .948 | .949 | .953 | .951 | .953 |
| | 100 | .954 | .951 | .955 | .952 | .948 | .948 | .952 | .949 |
| 0.40 | 20 | .948 | .950 | .955 | .955 | .924 | .932 | .940 | .954 |
| | 40 | .955 | .952 | .949 | .951 | .920 | .925 | .932 | .952 |
| | 60 | .957 | .953 | .943 | .951 | .928 | .928 | .931 | .943 |
| | 80 | .945 | .943 | .937 | .952 | .930 | .932 | .921 | .948 |
| | 100 | .948 | .946 | .937 | .953 | .920 | .926 | .918 | .944 |
| 0.80 | 20 | .949 | .949 | .936 | .948 | .900 | .913 | .906 | .937 |
| | 40 | .948 | .947 | .927 | .948 | .894 | .907 | .891 | .927 |
| | 60 | .952 | .951 | .919 | .949 | .895 | .911 | .874 | .933 |
| | 80 | .949 | .943 | .915 | .951 | .895 | .915 | .872 | .931 |
| | 100 | .953 | .948 | .913 | .948 | .893 | .902 | .859 | .934 |
| 1.20 | 20 | .951 | .943 | .914 | .940 | .871 | .890 | .876 | .925 |
| | 40 | .953 | .943 | .893 | .941 | .867 | .892 | .843 | .925 |
| | 60 | .953 | .948 | .885 | .940 | .858 | .894 | .825 | .922 |
| | 80 | .950 | .939 | .877 | .938 | .859 | .887 | .809 | .920 |
| | 100 | .946 | .940 | .871 | .933 | .858 | .886 | .799 | .914 |
| 1.60 | 20 | .956 | .949 | .883 | .931 | .836 | .866 | .837 | .915 |
| | 40 | .948 | .941 | .862 | .920 | .836 | .872 | .802 | .911 |
| | 60 | .953 | .945 | .843 | .932 | .831 | .875 | .773 | .909 |
| | 80 | .948 | .939 | .836 | .933 | .823 | .860 | .764 | .915 |
| | 100 | .947 | .941 | .834 | .928 | .830 | .865 | .749 | .917 |

Note: $PVR = 4$.

Table 2. Estimated Coverage Probabilities for the Bootstrap Percentile CI for $\delta_1$

| $\delta_1$ | $n_1 = n_2$ | $g = .000$, $h = .000$ | $g = .000$, $h = .225$ | $g = .760$, $h = -.098$ | $g = .225$, $h = .225$ |
|---|---|---|---|---|---|
| 0.0 | 20 | .936 | .929 | .920 | .921 |
| | 40 | .942 | .937 | .939 | .935 |
| | 60 | .939 | .935 | .935 | .938 |
| | 80 | .948 | .946 | .935 | .940 |
| | 100 | .945 | .939 | .940 | .941 |
| 0.4 | 20 | .934 | .922 | .926 | .915 |
| | 40 | .939 | .929 | .930 | .928 |
| | 60 | .942 | .935 | .937 | .932 |
| | 80 | .950 | .941 | .940 | .933 |
| | 100 | .948 | .936 | .947 | .931 |
| 0.8 | 20 | .931 | .904 | .915 | .900 |
| | 40 | .934 | .921 | .928 | .904 |
| | 60 | .943 | .921 | .933 | .916 |
| | 80 | .945 | .933 | .940 | .907 |
| | 100 | .944 | .929 | .938 | .916 |
| 1.2 | 20 | .929 | .882 | .905 | .862 |
| | 40 | .937 | .901 | .922 | .874 |
| | 60 | .943 | .905 | .925 | .884 |
| | 80 | .938 | .918 | .930 | .880 |
| | 100 | .949 | .913 | .934 | .892 |
| 1.6 | 20 | .926 | .861 | .883 | .824 |
| | 40 | .940 | .881 | .911 | .838 |
| | 60 | .945 | .889 | .908 | .850 |
| | 80 | .943 | .895 | .927 | .850 |
| | 100 | .942 | .893 | .927 | .848 |

Note: $PVR = 4$

## Methodology

Probability coverage was investigated for all combinations of: sample size $n_1$ = 20, 40, and 60 in combination with $n_2 = n_1$ and $n_2 = n_1 + 20$; population distribution (four cases from the family of g and h distributions), various PESs, $\delta_{R_1} = .00$, .40, .80, 1.20 and 1.60, and $PVR = .25$, .5, 1, 4, and 8. As in Study 2, $g = h = 0$, $g = .76$ and $h = -.098$, $g = 0$ and $h = .225$, and $g = .225$ and $h = .225$ were investigated. Because a CI for $\delta_{R_1}$ was being investigated, the data for treatment 2 were generated by using Equation (6). In all conditions the nominal confidence level was .95. As in the previous study, 5,000 replications and 600 bootstrap replications were used.

## Results

Table 3 contains estimated coverage probabilities for the percentile bootstrap CI for all conditions with $PVR = 8$. Estimated coverage

Table 3. Estimated Coverage Probabilities for the Percentile Bootstrap CI for $\delta_{R_1}$

| $n_1, n_2$ | $\delta_{R_1}$ | $g = .000$, $h = .000$ | $g = .000$, $h = .225$ | $g = .760$, $h = -.098$ | $g = .225$, $h = .225$ |
|---|---|---|---|---|---|
| 20, 20 | .00 | .943 | .945 | .945 | .950 |
| | .40 | .950 | .956 | .954 | .951 |
| | .80 | .948 | .955 | .952 | .954 |
| | 1.20 | .961 | .964 | .957 | .966 |
| | 1.60 | .960 | .966 | .962 | .960 |
| 20, 40 | .00 | .949 | .957 | .949 | .952 |
| | .40 | .951 | .954 | .956 | .958 |
| | .80 | .953 | .959 | .951 | .961 |
| | 1.20 | .967 | .964 | .958 | .965 |
| | 1.60 | .959 | .969 | .957 | .963 |
| 60, 60 | .00 | .949 | .947 | .947 | .948 |
| | .40 | .953 | .944 | .943 | .952 |
| | .80 | .949 | .950 | .948 | .957 |
| | 1.20 | .952 | .951 | .952 | .949 |
| | 1.60 | .947 | .959 | .954 | .958 |
| 60 80 | .00 | .945 | .952 | .944 | .950 |
| | .40 | .952 | .949 | .946 | .951 |
| | .80 | .949 | .959 | .951 | .959 |
| | 1.20 | .955 | .954 | .953 | .956 |
| | 1.60 | .955 | .961 | .954 | .953 |
| 100,100 | .00 | .950 | .948 | .949 | .947 |
| | .40 | .947 | .948 | .953 | .951 |
| | .80 | .950 | .946 | .949 | .957 |
| | 1.20 | .951 | .953 | .951 | .952 |
| | 1.60 | .953 | .956 | .953 | .956 |
| 100,120 | .00 | .948 | .955 | .947 | .948 |
| | .40 | .939 | .951 | .948 | .948 |
| | .80 | .955 | .949 | .950 | .948 |
| | 1.20 | .951 | .947 | .955 | .955 |
| | 1.60 | .956 | .960 | .959 | .959 |

Note. $PVR = 8$.

probabilities for other values of PVR were not noticeably different from those in Table 3. Over the 120 conditions reported in Table 3, empirical coverage ranged from .939 to .969, with an average coverage value of .953. The results suggest coverage probability increased as $\delta_{R_1}$ increased, but was largely unaffected by the sampled distribution and whether the sample sizes were equal.

## Conclusion

Estimating the magnitude of a treatment effect has become a required mode of analysis for many scientific journals in the social and behavioral sciences as a result of recommendations made by the APA Task Force regarding statistical inference. Not surprisingly, issues related to estimating the magnitude of an effect have become of paramount interest to applied researchers. One issue is what standard deviation to use in the denominator of the ES statistic. That is, since Glass's (1976), which used the control group's standard deviation to standardize the mean difference, other approaches have been recommended. Hedges (1981) recommended using the pooled standard deviation when the variances are homogeneous. Glass et al. (1981) recognized that if homogeneity of variances is not a reasonable assumption, the standard deviation for either group could be used as the denominator. This applies regardless of whether one of the treatment groups is a control group.

A second issue is how to use the ES measures to construct a CI. It is well known that when the pooled standard deviation is used in the denominator, CIs can be constructed by using the noncentral t distribution and will be exact when the scores are independently drawn from normal distributions and with equal variances. As shown in this article, an alternative interval based on the noncentral t distribution can be used when the standard deviation for one of the groups is used in the denominator, as would be done if Glass's (1976) ES were used or if the recommendation of Glass et al. (1981) were used when the variances are not homogeneous. However, the theory underlying this interval assumes data that are normal in form, which implies that the numerator and denominator of the ES are independently distributed. Independence does not hold when the data for the group that contributes the standard deviation are skewed. Accordingly, the interval could not be recommended without first examining its operating characteristics under nonnormality

As Wilcox and Keselman (2003) indicated, ES measures can be inaccurate when the data are drawn from nonnormal distributions because of the effects of nonnormality on means and standard deviations. Therefore, CIs calculated from a robust effect size $\left(\hat{\delta}_{R_1}\right)$ in which trimmed means replace means and the square root of the Winsorized variance replaces the standard deviation were also investigated. An additional issue considered was whether one could obtain accurate probability coverage for CIs for ES when coverage was based on theoretically obtained critical values (i.e., based on the noncentral *t* distribution) or obtained through a bootstrapping method. This was an important issue because others have demonstrated the benefits of using bootstrapping methodology (See, e.g., Keselman et al., 2002).

It this article, it was found that: (1) the classical approach, which divides the mean difference by a standard deviation from one group $\left(\text{i.e., } \hat{\delta}_1\right)$ in combination with the interval based on the noncentral *t* distribution had poor probability coverage when data were skewed, (2) the robust approach, which divides the difference of the trimmed means by the square root of the Winsorized variance from one group $\left(\text{i.e., } \hat{\delta}_{R_1}\right)$ in combination with the interval based on the noncentral *t* distribution also had poor probability coverage when data were nonnormal, (3) bootstrap CIs for $\delta_1$ can perform poorly, and (4) the percentile bootstrap interval for $\delta_{R_1}$ was very little affected by nonnormality, providing a very good interval for $\delta_{R_1}$.

An emphasis must be placed on the belief that it is important to estimate a robust parameter, that is, the robust PES, rather than the usual parameter of ES, when data are nonnormal. Researchers should be interested in

estimates of a parameter that is robust to conditions of skewness and outlying values. Inferences pertaining to robust parameters may be more valid than inferences pertaining to the least squares derived parameters when dealing with populations that are nonnormal (e.g., Hample, Ronchetti, Rousseeuw & Stahel, 1986; Huber, 1981; Staudte & Sheather, 1990). Hogg (1974, p. 919) maintained that most distributions are skewed in practice, and Tukey (1960) argued that most distributions will have heavy tails. Therefore, according to this perspective, the justification for (testing hypotheses and) setting robust intervals for robust parameters is that (testing the usual hypotheses and) setting intervals around the usual parameters is a mistake or at least shortsighted when other robust methods are available, methods that are not generally affected by a relatively few data points in a distribution or some minor characteristic of the distribution, points and characteristics that need not affect the quantity researchers are interested in.

As well, it was found that the natural sample estimate of the robust parameter, one based on trimmed means and a Winsorized variance, provides probability coverage that is fairly close to the target value of .95, when upper and lower critical values for the interval were obtained through a percentile bootstrap method. Despite the preference for a robust parameter, others may feel that, given a hypothesis about the least square means (which is not recommended with nonnormal data), $\delta$ is the appropriate effect size measure. These researchers must face the fact that neither the noncentral *t* distribution-based CI nor the percentile bootstrap CI will necessarily have coverage probability near the nominal value.

## References

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

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.

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

Cumming G. & Finch S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement*, *61*, 532-574

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

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

Glass, G., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.

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

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

Hogg, R. V. (1974). Adaptive robust procedures: A partial review and some suggestions for future applications and theory. *Journal of the American Statistical Association, 69*, 909-927.

Huber, P. J. (1981). *Robust statistics*. New York: Wiley.

Keselman, H. J., Wilcox, R., R., Othman, A. R., & Fradette, K. (2002). Trimming, transforming statistics, and bootstrapping: Circumventing the biasing effects of heteroscedasticity and non normality. *Journal of Modern Applied Statistical Methods, 1(2)*, 288-309.

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

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

Staudte, R. G., & Sheather, S. J. (1990). *Robust estimation and testing*. New York: Wiley.

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.

Tukey, J. W. (1960). A survey of sampling from contaminated normal distributions. In I. Olkin et al. (Eds.) *Contributions to probability and statistics*. Stanford, CA: Stanford University Press.

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

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

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