Journal of Modern Applied Statistical Methods

Volume 4 | Issue 2 Article 3

11-1-2005

Robust Confidence Intervals for Effect Size in the Two-Group Case

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

James Algina
University of Florida, algina@ufl.edu

Katherine Fradette
University of Manitoba, umfradet@cc.umanitoba.ca

Recommended Citation
Keselman, H. J.; Algina, James; and Fradette, Katherine (2005) "Robust Confidence Intervals for Effect Size in the Two-Group Case," *Journal of Modern Applied Statistical Methods*: Vol. 4 : Iss. 2 , Article 3.
DOI: 10.22237/jmasm/1130803320

Journal of Modern Applied Statistical Methods
November, 2005, Vol. 4, No. 2, 353-371

1538 – 9472/05/$95.00

INVITED ARTICLES
Robust Confidence Intervals for Effect Size in the Two-Group Case

H. J. Keselman
University of Manitoba

James Algina
University of Florida

Katherine Fradette
University of Manitoba

The probability coverage of intervals involving robust estimates of effect size based on seven procedures was compared for asymmetrically trimming data in an independent two-groups design, and a method that symmetrically trims the data. Four conditions were varied: (a) percentage of trimming, (b) type of nonnormal population distribution, (c) population effect size, and (d) sample size. Results indicated that coverage probabilities were generally well controlled under the conditions of nonnormality. The symmetric trimming method provided excellent probability coverage. Recommendations are provided.

Key words: Robust Intervals, effect size statistics, symmetric and asymmetric trimmed means, nonnormality

Introduction

Journal editorial policies in medicine and psychology encourage researchers to supplement significance testing by reporting confidence intervals (CIs) as well as effect size (ES) statistics. As Fidler, Thomason, Cumming, Finch, and Leeman (2004) note, this movement started in medicine as early as the 1980s (see Rothman 1975, 1978a, 1978b). In psychology, in the past 15 years or so, there has been renewed emphasis on reporting ESs because of editorial policies requiring ESs (e.g., Murphy, 1997; Thompson, 1994) and official support for the practice. According to *The Publication Manual of the American Psychological Association* (2001), "it is almost always necessary to include some index of ES or strength of relationship in your Results section." (p. 25). The practice of reporting ESs has also received support from the APA Task Force on Statistical Inference (Wilkinson and the Task Force on Statistical Inference, 1999). An interest in reporting CIs for ESs has accompanied the emphasis on ESs. Cumming and Finch (2001), for example, presented a primer of CIs for ESs. The purpose of this article is to bring to the attention of researchers in medicine and psychology, and other interested researchers, who set CIs around an ES parameter, a better approach than currently adopted methods.

Algina and Keselman (2003) and Algina, Keselman and Penfield (2005) investigated two two-group ES statistics, looking, in particular, at the confidence coefficient of two intervals associated with each. One of the ES statistics was Cohen's (1965) standardized mean difference statistic

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

H. J. Keselman is Professor of Psychology. Email: kesel@ms.umanitoba.ca. James Algina is Professor of Educational Psychology. Email algina@ufl.edu. Katherine H. Fradette is a doctoral student in the Department of Psychology. Email: umfradet@cc.umanitoba.ca.

where $\bar{Y}_j$ is the mean for the jth level ($j = 1, 2$) of a treatment factor and S is the square root of the pooled variance. The second was

$$d_R = .643\left(\frac{\bar{Y}_{t2} - \bar{Y}_{t1}}{S_W}\right),$$

where $\bar{Y}_{tj}$ denotes the jth 20% trimmed mean, S_W is the square root of the pooled 20% Winsorized variance and .643 is the population 20% Winsorized standard deviation for a standard normal distribution. These authors included .643 in the definition of their robust effect so that the population values of d_R (δ_R) and d (δ) would be equal when data are drawn from normal distributions with equal variances.

However, these authors also pointed out that it is not obligatory to include the .643 multiplier in the definition of d_R and δ_R. Accordingly, the multiplier is excluded in this article. Using each ES statistic, CIs were constructed by using critical values obtained from theory or through a bootstrap method. Algina and Keselman (2003) found that probability coverage for intervals of the usual statistic based on least squares estimators was inaccurate whether or not the interval's critical values were obtained from a theoretical or bootstrap distribution. They also reported that probability coverage was inaccurate when the interval was set around a robust parameter of ES and the critical values for the interval were obtained from a theoretical probability distribution. However, probability coverage was by in large accurate (e.g., .940-.971 for a .95 confidence coefficient) when the interval for the robust parameter of ES was based on critical values obtained through a bootstrap method (see Algina et al., 2005).

Keselman, Wilcox, Lix, Algina and Fradette (in press) found that tests of treatment group equality based on robust estimators performed very well, with respect to Type I error control and power to detect effects in nonnormal heteroscedastic distributions, when adopting robust estimators based on asymmetric trimming of the data. That is, rather than trim a predetermined fixed amount of data from each tail of the empirical distribution, as frequently is recommended in the literature (e.g., 20% from each tail; see Wilcox, 1997; Wilcox & Keselman, 2003), Keselman et al. used nine adaptive procedures that empirically determined the amounts of data that should be trimmed in the right and left tails of each of the nonnormal distributions that they examined in their Monte Carlo investigation. The rationale behind asymmetric trimming is to remove more of the offending data (i.e., data that does not represent the bulk of the observations, that is, the typical score) from the tail containing more of the outlying values.

Based on the two aforementioned studies, it is believed that more accurate confidence coefficients for Algina and Keselman's (2003) and Algina et al.'s (2005) robust parameter of ES could be obtained by adopting the asymmetric trimming procedures enumerated in Keselman et al. (in press). Accordingly, this issue will be investigated in this article.

Theoretical Background

ES Statistics and Accompanying CIs

In the two independent-groups paradigm, Cohen's (1965) standardized mean difference statistic, d, is a popular choice for estimating ES. His ES statistic is defined as

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

Cohen's d estimates

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

where μ_j is the jth population mean and σ is the population standard deviation, assumed to be equal for both groups.

When the scores are independently distributed and are drawn from normal distributions having equal variances, an exact CI for the population ES (i.e., δ) can be constructed by using the noncentral t distribution (see, e.g., Cumming & Finch, 2001 or Steiger & Fouladi, 1997). The noncentral t distribution is the sampling distribution of the t statistic when δ is not equal to zero; it has two parameters. The first is the degrees of freedom and equals $N-2$ in the two independent-groups set-up ([$N=n_1+n_2$] and the number of observations in a level is denoted by n_j). The second parameter is the noncentrality parameter

$$\lambda=\sqrt{\frac{n_1n_2}{n_1+n_2}}\left(\frac{\mu_2-\mu_1}{\sigma}\right)=\sqrt{\frac{n_1n_2}{n_1+n_2}}\,\delta .$$

The noncentrality parameter controls the location of the noncentral t distribution. The mean of the noncentral t distribution is $\approx\lambda$ (Hedges, 1981); the accuracy of the approximation improves as N increases.

To find a 95% (for example) CI for δ, one would first use the noncentral t distribution to find a 95% CI for λ. A CI for δ can then be obtained by multiplying the limits of the interval for λ by $\sqrt{(n_1+n_2)/n_1n_2}$. The lower limit of the CI for λ is the noncentrality parameter for the noncentral t distribution in which the calculated t statistic

$$t=\sqrt{\frac{n_1n_2}{n_1+n_2}}\left(\frac{\bar{Y}_2-\bar{Y}_1}{S}\right)$$

is the .975 quantile. The upper limit of the interval for λ is the noncentrality parameter for the noncentral t distribution in which the calculated t statistic is the .025 quantile of the distribution (see Steiger & Fouladi, 1997).

The use of the noncentral t distribution is based on the assumption that the data are drawn from normal distributions. If this assumption is not true, there is no guarantee that the actual probability coverage for the interval will match the nominal probability coverage, as was demonstrated by Algina and Keselman (2003). In addition, as noted by Wilcox and Keselman (2003), when data are not normal, the usual population ES can be misleading because the (least squares) means and standard deviations can be affected by skewed data and by outliers. A better strategy, they maintain, is to replace the least squares values by robust estimates, such as trimmed means and Winsorized variances, and, accordingly, estimate a robust population ES.

As an alternative to d, Algina and Keselman (2003) and Algina et al. (2005) (hereafter referred to as A&K) proposed

$$d_R=\left(\frac{\bar{Y}_{t2}-\bar{Y}_{t1}}{S_W}\right).$$

(Remember, the .643 multiplier is not used.)

The robust population ES is

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

where μ_{tj} is the jth population 20% trimmed mean and σ_W is the population analogue of S_W. (See appendix 1.)

As Algina and Keselman (2003) and Algina et al. (2005) indicated, an approximately correct CI for δ_R can also be constructed by using the noncentral t distribution. However, as previously noted, this approach to forming intervals did not provide satisfactory probability coverage when data were obtained from nonnormal distributions. However, Algina et al. did find that probability coverage, under conditions of nonnormality, was generally reasonably good when critical values were obtained through a percentile bootstrap empirical sampling distribution, not from the noncentral t distribution.

Adaptive Trimming Methods

The theoretical background to the asymmetric trimming methods investigated by Keselman et al. (in press) is now discussed. Based on the work of Hogg (1974, 1982) and others, Reed and Stark (1996) defined seven adaptive location estimators based on measures of tail-length and skewness for a set of n observations. To define these estimators the measures of tail-length and skewness must first be defined. By adopting the notation of Hogg (1974, 1982) and Reed and Stark (1996), based on the ordered values, we let $L_\alpha=$ the mean of the smallest [αn] observations, where [αn] denote the greatest integer less than αn and $U_\alpha =$ the mean of the largest $\left[\alpha n\right]$ observations. When $\alpha = .05$, and, therefore, $L_{(.05)}$ is the mean of the smallest [.05n] observations, $B =$ the mean of the next largest .15n observations, $C =$ the mean of the next largest .30n observations, $D =$ the mean of the next largest .30n observations, and $E =$ the mean of the next largest .15n observations.

Tail-length measures. Hogg (1974) defined two measures of tail-length, Q and Q_1, where

$$Q = \left(U_{(.05)} - L_{(.05)}\right) \Big/ \left(U_{(.5)} - L_{(.5)}\right) \text{ and}$$

$$Q_1 = \left(U_{(.2)} - L_{(.2)}\right) \Big/ \left(U_{(.5)} - L_{(.5)}\right).$$

Q and Q_1 can be used to classify symmetric distributions as light-tailed, medium-tailed or heavy-tailed. Q and Q_1 are location free statistics and, moreover, are uncorrelated with location statistics such as trimmed means (Reed & Stark, 1996, p. 12). According to Hogg and Reed and Stark, values of $Q < 2$ imply a light-tailed distribution, $2.0 \le Q \le 2.6$ a medium-tailed distribution, $2.6 < Q \le 3.2$ a heavy-tailed distribution and $Q > 3.2$ a very heavy-tailed distribution. The cutoffs for Q_1 are: $Q_1 < 1.81$ (light-tailed), $1.81 \le Q_1 \le 1.87$ (medium-tailed) and $Q_1 > 1.87$ (heavy-tailed).

Hogg (1982) introduced another measure of tail-length:

$$H_3 = \left(U_{(.05)} - L_{(.05)}\right) \Big/ (E - B).$$

With this measure, values of $H_3 < 1.26$ suggest that the tails of the distribution are similar to a uniform distribution, values of 1.26 through 1.76 suggest a normal distribution and values greater than 1.76 suggest the tails are similar to those of a double exponential distribution.

Measures of skewnesss

Reed and Stark (1996) defined four measures of skewness as:

$$Q_2 = \left(U_{(.05)} - T_{(.25)}\right) \Big/ \left(T_{(.25)} - L_{(.05)}\right),$$

$$H_1 = \left(U_{(.05)} - D\right) \Big/ \left(C - L_{(.05)}\right),$$

$$SK_2 = \left(Y_{(1)} - YMD\right) \Big/ \left(YMD - Y_{(n)}\right) \text{ and}$$

$$SK_5 = \left(Y_{(1)} - YM\right) \Big/ \left(YM - Y_{(n)}\right),$$

where YMD is the median, YM is the arithmetic mean, $T_{(.25)}$ is the .25- trimmed mean (T_α) given below and $Y_{(1)}$ and $Y_{(n)}$ are, respectively the first and last ordered observations. According to Reed (1998), the α-trimmed mean is defined as

$$T_\alpha = \frac{1}{n(1-2\alpha)}\left[\sum_{i=k+1}^{n-k} Y_i + (k - \alpha n)\left(Y_k + Y_{n-k+1}\right)\right].$$

(In this definition a proportion, α, has been trimmed from each tail) and the accompanying Winsorized variance S^2 is defined as

$$S^2 = \frac{1}{(n-1)(1-2\alpha)^2}$$

$$\left[\sum_{i=k+1}^{n-k}\left(Y_i - T_\alpha\right)^2 + k\left(Y_k - T_\alpha\right)^2 + k\left(Y_{n-k+1} - T_\alpha\right)^2\right]$$

where $k = [\alpha n] + 1$.

Based on the former definitions of tail-length and skewness, Reed and Stark (1996, p. 13) proposed a set of adaptive linear estimators “that have the capability of asymmetric trimming.” These authors defined a general scheme for their approach as follows:

1. Set the value for the total amount of trimming from the sample, α.

1) Determine the proportion to be trimmed from the lower end of the sample (α_l) by the following proportion: $\alpha_l = \alpha\left[UW_X / \left(UW_X + LW_X\right)\right]$, where UW_X and LW_X are the numerator and

denominator portions of the previously defined selector statistics (i.e., tail-length and skewness).

2) The upper trimming proportion is then given by $\alpha_u = \alpha - \alpha_l$.

Based on this general schema, Reed and Stark (1996) defined seven hinge estimators, which are trimmed means:

1. $HQ \quad \alpha_l = \alpha\left[UW_Q / \left(UW_Q + LW_Q\right)\right]$,
2. $HQ_1 \quad \alpha_l = \alpha\left[UW_{Q_1} / \left(UW_{Q_1} + LW_{Q_1}\right)\right]$,
3. $HH_3 \quad \alpha_l = \alpha\left[UW_{H_3} / \left(UW_{H_3} + LW_{H_3}\right)\right]$,
4. $HQ_2 \quad \alpha_l = \alpha\left[UW_{Q_2} / \left(UW_{Q_2} + LW_{Q_2}\right)\right]$,
5. $HH_1 \quad \alpha_l = \alpha\left[UW_{H_1} / \left(UW_{H_1} + LW_{H_1}\right)\right]$,
6. $HSK_2 \quad \alpha_l = \alpha\left[UW_{SK_2} / \left(UW_{SK_2} + LW_{SK_2}\right)\right]$, and
7. $HSK_5 \quad \alpha_l = \alpha\left[UW_{SK_5} / \left(UW_{SK_5} + LW_{SK_5}\right)\right]$.

Keselman et al. (in press), investigating Type I error rates and power of procedures for testing equality of two trimmed means when variances are not assumed to be equal, examined the Reed and Stark (1996) procedure with various values for α because the literature varies on the amount of recommended (symmetric) trimming. Rosenberger and Gasko (1983) recommended 25% when sample sizes are small, though they thought generally 20% suffices. Wilcox (1997) also recommended 20%, and Mudholkar, Mudholkar and Srivastava (1991) suggested 15%. Ten percent has been considered by Hill and Dixon (1982), Huber (1977), Stigler (1977) and Staudte and Sheather (1990); results reported by Keselman, Wilcox, Othman and Fradette (2002) also support 10% trimming.

Reed and Stark (1996) found, based on a simulation study, that $T_{.10}$, $T_{.15}$, HSK_2 and HSK_5 were the most efficient estimators when the distribution was symmetric. When the distribution was asymmetric, they found that "**HQ**, $\mathbf{HQ}_1$, $\mathbf{HQ}_2$, HH_1, HSK_2 and HSK_5 [were] consistently among the top four estimators, with HQ_1 and HQ_2 in the top three" (p. 661).

According to Keselman et al. (in press), one can modify Reed and Stark's (1996) tail-length and skewness measures for the multi-group problem and then apply the modified multi-group measures to the hinge estimators. In particular, they indicated that each of the measures can be modified by taking weighted averages (in a manner analogous to the modifications of tail-length and symmetry measures suggested by Babu, Padmanaban and Puri, 1999) of each numerator and denominator term. For example, for the multi-group problem, where n_j represents the number of observations in each group, Q_1 and Q_2 can be defined as

$$Q_1 = \left[\sum_j n_j\left(U_{(.2)} - L_{(.2)}\right) \Big/ \sum_j n_j\right] \Bigg/ \left[\sum_j n_j\left(U_{(.5)} - L_{(.5)}\right) \Big/ \sum_j n_j\right],$$

and

$$Q_2 = \left[\sum_j n_j\left(U_{(.05)} - T_{(.25)}\right) \Big/ \sum_j n_j\right] \Bigg/ \left[\sum_j n_j\left(T_{(.25)} - L_{(.05)}\right) \Big/ \sum_j n_j\right].$$

The other measures would be similarly modified and these multi-group measures of tail-length and skewness are the measures that are applied to the general scheme proposed by Reed and Stark (1996).

Based on these multi-group tail-length and skewness measures, and their application to the hinge estimators, Keselman et al. (in press) reported that over the 288 empirical values they collected for each method investigated, in which they varied the total percent of data trimmed, sample size, degree of variance heterogeneity, pairing of variances and group sizes and population shape, five methods resulted in exceptionally good control of Type I error rates (HH3, HQ2, HH1, HSK2 and HSK5). With regard to the power to detect nonnull treatment effects, they found that HH3 was uniformly more powerful than the remaining ones.

Robust Estimation

In this study, the methods for constructing CIs for a robust ES, defined by using robust measures of central tendency and variability are investigated. It is important to note that α-trimmed means and Winsorized variances can be defined in a number of different ways (Hogg, 1974; Reed, 1998; Keselman et al., in press; Wilcox, 2003). Suppose n_j independent random observations $Y_{1j}, Y_{2j}, \ldots, Y_{n_j j}$ are sampled from population j ($j = 1, 2$). Let $Y_{(1)j} \le Y_{(2)j} \le \cdots \le Y_{(n_j)j}$ represent the ordered observations associated with the jth group. The approach taken by Reed (1998) is based on the work of Hogg (1974). For Hogg, the α-trimmed mean is

$$m(\alpha) = (1/h) \sum_{i=g+1}^{n_j - g} Y_{(i)},$$

where α is usually selected so that $g = \lfloor n_j \alpha \rfloor$ and $h = n_j - 2g = n_j - 2[n_j \alpha]$. The standard error of $m(\alpha)$ that Hogg suggests is based on the work of Tukey and McLaughlin (1963) and Huber (1970) and, according to these authors, is estimated by

$$S_{m(\alpha)} = \sqrt{SS(\alpha)/h(h-1)},$$

where $SS(\alpha)$ is the Winsorized sum of squares, defined as

$$\begin{aligned}&(g+1)\left[Y_{(g+1)} - m(\alpha)\right]^2 \\ &+ \left[Y_{(g+2)} - m(\alpha)\right]^2 + \ldots \\ &+ \left[Y_{(n_j-g-1)} - m(\alpha)\right]^2 + (g+1)\left[Y_{(n_j-g)} - m(\alpha)\right]^2.\end{aligned}$$

When allowing for different amounts of trimming in each tail of the distribution, Hogg (1974) defines the trimmed mean as

$$m(\alpha_1, \alpha_1) = (1/h) \sum_{i=g_1+1}^{n_j - g_2} Y_{(i)},$$

where $g_1 = \lfloor n_j \alpha_1 \rfloor$ and $g_2 = \lfloor n_j \alpha_2 \rfloor$ and $h = n_j - g_1 - g_2$. Hogg suggests that the standard deviation of $m(\alpha_1, \alpha_2)$ can be estimated as

$$S_{m(\alpha_1, \alpha_2)} = \sqrt{SS(\alpha_1, \alpha_2)/h(h-1)},$$

where $SS(\alpha_1, \alpha_2)$ can be calculated as

$$\begin{aligned}&(g_1+1)\left[Y_{(g_1+1)} - m(\alpha_1, \alpha_2)\right]^2 \\ &+ \left[Y_{(g_1+2)} - m(\alpha_1, \alpha_2)\right]^2 + \ldots \\ &+ \left[Y_{(n_j-g_2-1)} - m(\alpha_1, \alpha_2)\right]^2 + \\ &(g_2+1)\left[Y_{(n_j-g_2)} - m(\alpha_1, \alpha_2)\right]^2 \\ &- \frac{\left\{\begin{aligned}&(g_1)\left[Y_{(g_1+1)} - m(\alpha_1, \alpha_2)\right] + \\ &(g_2)\left[Y_{(n_j-g_2)} - m(\alpha_1, \alpha_2)\right]\end{aligned}\right\}^2}{n_j}\end{aligned}$$

Based on the preceding, our robust estimate of ES for asymmetrically trimmed data is defined as

$$d_R = \frac{m_1(\alpha_1, \alpha_2) - m_2(\alpha_1, \alpha_2)}{\sqrt{\dfrac{SS_1(\alpha_1, \alpha_2) + SS_2(\alpha_1, \alpha_2)}{N-2}}},$$

where $m_j(\alpha_1, \alpha_2)$ and $SS_j(\alpha_1, \alpha_2)$ are the jth asymmetrically trimmed mean and sum of squares, respectively. (See Appendix 2.)

Methodology

Probability coverage for seven ES statistics (based on seven hinge estimators: HQ, HQ1, HH3, HQ2, HH1, HSK2, and HSK5) was estimated for all combinations of the following four factors: (a) four values of total trimming, namely 10%, 15%, 20% and 25%, (b) population distribution (four cases from the family of g and h distributions), (c) sample size: $n_1 = n_2 = 20,\ 40,\ 60,\ 80, \text{ and } 100$, and (d) population ES ($\text{PES} = \delta_R$) of 0, .2, .5, .8, 1.1, and 1.2. The A&K statistic was also included, where the values of symmetric trimming investigated were 5%, 10%, 15% and 20%.

The data were generated from the family of g and h distributions (Hoaglin, 1985). Specifically, it was chosen to investigate four g and h distributions:

(a) $g = h = 0$, the standard normal distribution ($\gamma_1 = \gamma_2 = 0$),

(b) $g = 0 \text{ and } h = .225$, a long-tailed distribution ($\gamma_1 = 0, \gamma_2 = 154.84$),

(c) $g = .76 \text{ and } h = -.098$, a distribution with skew and kurtosis equal to that for an exponential distribution ($\gamma_1 = 2, \gamma_2 = 6$), and

(d) $g = .225 \text{ and } h = .225$, a long-tailed skewed distribution ($\gamma_1 = 4.90, \gamma_2 = 4673.80$).

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, $Y_{ij} = Z_{ij} \exp\left(\frac{hZ_{ij}^2}{2}\right)$. The Z_{ij} scores were generated by using RANNOR from SAS (1999). In particular, the following method to generate our data was used:

1. The original Y_{ij} data (for both groups) were generated from a desired population distribution (e.g., $g = .225 \text{ and } h = .225$). (NOTE: The original Y_{i2} data are not yet transformed)
2. A bootstrap sample (Y_{ij}^*) was obtained from the original sample by sampling n_1 observations with replacement from Y_{i1} and n_2 observations with replacement from Y_{i2}.
3. With the bootstrap data, we determined α_1 and α_2 for the desired total trimming percentage (e.g., 15%) for each of the seven hinge estimators.
4. The bootstrapped data for group 2 (Y_{i2}^*) were then transformed according to $Y_{i2}^* + \sigma_W \times \delta_R$, where σ_W depended on the hinge estimator, the total % of trimming, and the population distribution under investigation. For a particular population distribution and total % of trimming, σ_W was determined prior to conducting the study. That is, 1,000,000 observations were first generated from the population distribution in question and then the population trimming strategy was determined for each of the hinge estimators under the desired total % of trimming. The σ_W values for the seven different hinge estimators were then determined by computing the Winsorized standard deviation of the 1,000,000 observations, using the trimming strategies of each of the estimators.
5. The transformed bootstrap data was then used to compute the trimmed means ($\bar{Y}_{t1}^*$ and $\bar{Y}_{t2}^*$) and the pooled Winsorized standard deviation (S_W^*) for each of the 7 different hinge estimator methods, based on the trimming strategies previously determined.

6. For each estimator, the following was computed $d_R^* = \dfrac{\bar{Y}_{t2}^* - \bar{Y}_{t1}^*}{S_W^*}$.
7. Steps 1 through 6 were repeated 600 times.
8. For each hinge estimator, the 600 bootstrap ES estimates (d_R^*) were ranked and the upper and lower limits of the CIs were determined in the following manner. Letting $l = .025B$, rounded to the nearest integer, and $u = B - l$, an estimate of the .025 and .975 quantiles of the distribution of d_R is $d_{R_{(l+1)}}^*$ and $d_{R_{(u)}}^*$.
9. Finally, steps 1 through 8 were repeated 5000 times.

The nominal confidence level for all intervals was .95.

Results

Table 1 contains average probability coverage rates for the seven hinge estimator methods as well as A&K for setting intervals around the PES for the effects investigated. Bradley's (1978) liberal criterion will be used to judge the robustness of the methods.

Coverage probabilities within the interval .925-.975 are deemed well controlled, while those outside this range are regarded as substantially affected by an investigated effect(s). Values outside the interval will be demarcated with boldface type in the tables. The grand mean coverage probabilities were obtained over 480 conditions and most apparent is that the empirical values are not only contained in Bradley's interval, but, moreover, are actually quite close to the nominal .95 value, with the largest deviation between nominal and empirical values equaling .004. Indeed, the range of empirical values extends from .946 to .949. Similarly, none of the remaining Table 1 values fell outside the Bradley liberal criterion.

Thus, by this standard of robustness, all hinge estimator methods for setting intervals around the robust PES can be regarded as not adversely affected by the effects of percentage of trimming, sample size, PES, and shape of distribution. Indeed, the number of times each of the methods' empirical values fell outside the liberal interval were tabulated and it was found that, over the 3840 estimates (480 conditions X 8 procedures), only 56 were not contained in the interval (less than 1.5% of the values!).

Not surprisingly, 51 of these values occurred when $n = 20$; the remaining five values occurred when $n = 40$. From this tabulation it was also found that, of the hinge estimator procedures, only HSK2 and HSK5 never had a value outside the Bradley interval. However, if the $n = 20$ results are excluded, then HQ, HQ1, and HH3 can be added to this list of procedures that never had a value over the 480 conditions outside the Bradley interval. Also noteworthy is that all 480 of the A&K values were in the Bradley interval.

Nonetheless, one can observe from the tabled values that there are variations in coverage probabilities due to the investigated effects. That is, it appears that coverage probabilities were closer to .95 when the: (a) percentage of total trimming was at least 20% (for A&K the empirical estimates were equal across percentages of symmetric trimming), (b) sample size was at least 80 per group, and (c) nonnormal distribution was not $g = .76 \text{ and } h = -.098$.

Accordingly, exemplars of these empirical coverage probabilities are presented in Tables 2-5, where the four tables are for the four distributions investigated. When $PES = 0$, all empirical coverage probabilities (not tabled) were contained within Bradley's (1978) interval across all sample size and population distributions investigated. In Tables 2-5, 28 of the 1152 empirical values $(\simeq 2.4\%)$ were not contained in the .925-.975 interval. Twenty-five of the affected values occurred when data were obtained from the $g = .76 \text{ and } h = -.098$ distribution and when $n = 20$ (Table 4).

The remaining three liberal values also occurred when $n = 20$ but in these instances the data were $g = .225 \text{ and } h = .225$ distributed. One should also notice that empirical values for the A&K procedure were always in Bradley's (1978) interval across the

Table 1. Summary Data for Estimated Coverage Probabilities for Nominal 95% Bootstrap Intervals

Condition	A&K	HQ	HQ1	HH3	HQ2	HH1	HSK2	HSK5
Grand Mean	.949	.947	.948	.947	.947	.946	.948	.948
% Trimming								
10		.943	.945	.944	.944	.942	.948	.948
15		.946	.949	.946	.947	.946	.949	.948
20		.949	.949	.948	.948	.947	.948	.948
25		.949	.949	.948	.949	.948	.947	.948
5 (Symmetric)	.949							
10 (Symmetric)	.949							
15 (Symmetric)	.949							
20 (Symmetric)	.949							
Sample Size								
20	.950	.939	.943	.937	.938	.936	.948	.949
40	.951	.948	.950	.948	.948	.946	.949	.949
60	.946	.949	.949	.949	.949	.948	.947	.947
80	.950	.950	.950	.949	.950	.950	.948	.948
100	.948	.950	.949	.950	.950	.950	.947	.947
PES								
0	.946	.945	.945	.945	.947	.946	.946	.946
0.2	.947	.946	.947	.946	.948	.947	.948	.948
0.5	.949	.946	.947	.946	.947	.946	.947	.947
0.8	.949	.948	.949	.947	.947	.946	.948	.948
1.1	.951	.949	.950	.948	.948	.946	.949	.949
1.4	.953	.948	.949	.947	.947	.944	.949	.948
Distribution								
g=0/h=0	.947	.946	.946	.946	.947	.947	.946	.947
g=0/h=.225	.951	.944	.946	.944	.941	.936	.946	.944
g=.76/h=-.098	.947	.950	.950	.949	.950	.950	.950	.951
g=.225/h=.225	.951	.949	.950	.948	.951	.951	.950	.950

Notes: Based on definitions of tail-length and skewness, Reed and Stark (1996, p. 13) defined seven hinge estimators that have the capability of asymmetric trimming: HQ, HQ1, HH3, HQ2, HH1, HSK2, HSK5; Sample Size $(n_1 = n_2)$; PES-Population Effect Size; $g = X / h = Y$ specifies a particular g and h distribution with specific values of skewness and kurtosis.

Table 2. Estimated Coverage Probabilities for Nominal 95% Bootstrap Intervals ($g = 0$ & $h = 0$).

PES	n	Trimming	Test							
			A&K	HQ	HQ1	HH3	HQ2	HH1	HSK2	HSK5
0.2	20	5%	.942							
		10%	.943	.935	.935	.935	.938	.937	.939	.940
		15%	.944	.940	.941	.939	.942	.942	.941	.942
		20%	.945	.942	.943	.941	.944	.944	.942	.942
		25%		.942	.942	.942	.944	.944	.942	.942
	60	5%	.940							
		10%	.939	.945	.944	.945	.945	.945	.944	.944
		15%	.940	.946	.945	.945	.945	.945	.945	.945
		20%	.938	.946	.945	.946	.946	.946	.944	.945
		25%		.945	.946	.945	.946	.946	.945	.946
	100	5%	.948							
		10%	.949	.945	.944	.946	.946	.946	.945	.945
		15%	.948	.947	.946	.947	.947	.947	.946	.945
		20%	.947	.946	.945	.945	.947	.947	.946	.946
		25%		.945	.945	.945	.946	.946	.946	.946
0.8	20	5%	.946							
		10%	.950	.939	.939	.939	.940	.940	.943	.944
		15%	.951	.946	.947	.943	.946	.946	.946	.946
		20%	.953	.951	.951	.950	.950	.949	.949	.951
		25%		.949	.950	.948	.952	.952	.950	.952
	60	5%	.943							
		10%	.943	.947	.949	.949	.950	.950	.949	.949
		15%	.943	.949	.950	.950	.949	.949	.947	.947
		20%	.947	.951	.951	.950	.951	.951	.950	.951
		25%		.950	.949	.950	.953	.953	.951	.952
	100	5%	.944							
		10%	.944	.949	.949	.949	.949	.949	.949	.949
		15%	.945	.949	.949	.948	.948	.948	.947	.947
		20%	.945	.950	.950	.949	.949	.950	.949	.949
		25%		.949	.948	.948	.948	.948	.947	.948
1.4	20	5%	.943							
		10%	.951	.939	.939	.939	.940	.940	.942	.943
		15%	.952	.946	.950	.944	.947	.947	.949	.949
		20%	.954	.951	.948	.952	.952	.951	.954	.953
		25%		.950	.951	.950	.954	.953	.953	.955
	60	5%	.945							
		10%	.946	.947	.948	.947	.950	.951	.948	.947
		15%	.946	.948	.947	.948	.949	.949	.948	.947
		20%	.945	.951	.950	.949	.948	.948	.948	.948
		25%		.950	.950	.949	.950	.950	.950	.950
	100	5%	.946							
		10%	.949	.948	.949	.949	.949	.949	.948	.948
		15%	.949	.950	.950	.950	.949	.949	.948	.949
		20%	.950	.949	.951	.950	.950	.950	.947	.948
		25%		.949	.949	.949	.949	.948	.950	.950

Table 3. Estimated Coverage Probabilities for Nominal 95% Bootstrap Intervals ($g = 0$ & $h = .225$).

PES	N	Trimming	Test							
			A&K	HQ	HQ1	HH3	HQ2	HH1	HSK2	HSK5
0.2	20	5%	.944							
		10%	.950	.936	.937	.937	.934	.933	.942	.942
		15%	.949	.935	.946	.933	.943	.942	.946	.947
		20%	.946	.944	.947	.943	.946	.945	.946	.947
		25%		.947	.947	.944	.948	.947	.945	.948
	60	5%	.942							
		10%	.943	.948	.948	.948	.953	.952	.948	.948
		15%	.941	.950	.950	.950	.950	.951	.950	.949
		20%	.940	.948	.949	.948	.949	.948	.946	.946
		25%		.949	.949	.948	.950	.950	.945	.947
	100	5%	.950							
		10%	.951	.951	.950	.950	.949	.950	.946	.947
		15%	.950	.949	.948	.949	.948	.948	.948	.948
		20%	.950	.949	.948	.947	.949	.950	.949	.949
		25%		.948	.947	.947	.949	.948	.949	.946
0.8	20	5%	.949							
		10%	.959	.937	.937	.937	.935	.934	.946	.948
		15%	.958	.943	.953	.940	.944	.943	.952	.951
		20%	.958	.952	.953	.949	.950	.950	.955	.955
		25%		.953	.953	.952	.954	.953	.955	.957
	60	5%	.953							
		10%	.948	.949	.949	.947	.952	.952	.951	.951
		15%	.946	.951	.956	.951	.950	.952	.953	.952
		20%	.948	.957	.952	.955	.953	.953	.950	.950
		25%		.954	.951	.954	.953	.953	.950	.952
	100	5%	.950							
		10%	.946	.954	.955	.955	.958	.959	.953	.954
		15%	.944	.955	.954	.956	.953	.955	.953	.953
		20%	.947	.953	.950	.953	.953	.953	.951	.950
		25%		.952	.951	.952	.951	.951	.943	.951
1.4	20	5%	.952							
		10%	.965	.934	.933	.933	.929	.928	.948	.947
		15%	.963	.941	.958	.938	.939	.937	.954	.952
		20%	.963	.954	.946	.946	.943	.942	.957	.957
		25%		.950	.948	.946	.949	.948	.962	.958
	60	5%	.960							
		10%	.955	.950	.947	.945	.954	.951	.956	.957
		15%	.951	.949	.959	.948	.950	.951	.954	.954
		20%	.949	.960	.953	.957	.954	.953	.952	.953
		25%		.959	.953	.955	.954	.954	.950	.953
	100	5%	.956							
		10%	.955	.957	.956	.956	.959	.959	.954	.954
		15%	.953	.954	.951	.953	.957	.957	.951	.952
		20%	.950	.956	.952	.952	.954	.954	.953	.953
		25%		.954	.954	.954	.954	.955	.935	.951

Table 4. Estimated Coverage Probabilities for Nominal 95% Bootstrap Intervals ($g = .76$ & $h = -.098$).

PES	N	Trimming	Test							
			A&K	HQ	HQ1	HH3	HQ2	HH1	HSK2	HSK5
0.2	20	5%	.940							
		10%	.946	.927	.927	.927	.926	.926	.943	.943
		15%	.947	.932	.941	.932	.930	.929	.946	.946
		20%	.947	.941	.942	.939	.935	.932	.945	.946
		25%		.943	.944	.942	.940	.935	.945	.945
	60	5%	.936							
		10%	.938	.944	.948	.944	.944	.938	.947	.948
		15%	.938	.948	.947	.949	.945	.944	.946	.947
		20%	.938	.948	.949	.949	.949	.946	.948	.947
		25%		.947	.949	.949	.948	.947	.949	.949
	100	5%	.948							
		10%	.944	.950	.949	.950	.947	.946	.949	.948
		15%	.948	.949	.950	.950	.949	.948	.949	.949
		20%	.949	.950	.949	.948	.951	.949	.948	.947
		25%		.950	.948	.948	.950	.949	.947	.948
0.8	20	5%	.934							
		10%	.948	**.909**	**.914**	**.909**	**.905**	**.895**	.940	.941
		15%	.948	**.921**	.934	**.922**	**.912**	**.906**	.948	.949
		20%	.950	.934	.939	.935	**.921**	**.909**	.948	.949
		25%		.939	.942	.941	.926	**.917**	.951	.948
	60	5%	.949							
		10%	.949	.946	.947	.946	.941	.933	.948	.948
		15%	.944	.948	.947	.951	.946	.941	.947	.947
		20%	.944	.950	.950	.951	.949	.943	.945	.941
		25%		.951	.951	.951	.947	.947	.945	.941
	100	5%	.946							
		10%	.948	.952	.950	.951	.954	.948	.946	.947
		15%	.945	.949	.949	.950	.951	.952	.946	.944
		20%	.946	.948	.947	.947	.947	.949	.944	.936
		25%		.947	.948	.946	.949	.949	.941	.937
1.4	20	5%	.929							
		10%	.957	**.903**	**.907**	**.903**	**.892**	**.878**	.942	.943
		15%	.953	**.912**	.932	**.913**	**.905**	**.894**	.955	.954
		20%	.956	.931	.939	.931	**.917**	**.898**	.956	.952
		25%		.938	.945	.938	.924	**.911**	.948	.942
	60	5%	.955							
		10%	.953	.943	.951	.942	.939	**.921**	.944	.946
		15%	.950	.952	.951	.953	.944	.938	.948	.943
		20%	.949	.953	.952	.953	.948	.940	.944	.933
		25%		.951	.954	.952	.950	.946	.939	.932
	100	5%	.953							
		10%	.952	.951	.951	.949	.946	.935	.953	.953
		15%	.952	.950	.950	.951	.949	.945	.952	.945
		20%	.951	.950	.951	.953	.952	.944	.948	.932
		25%		.947	.953	.950	.947	.948	.936	.931

Table 5. Estimated Coverage Probabilities for Nominal 95% Bootstrap Intervals ($g = .225$ & $h = .225$).

PES	N	Trimming	Test							
			A&K	HQ	HQ1	HH3	HQ2	HH1	HSK2	HSK5
0.2	20	5%	.946							
		10%	.951	.929	.930	.930	.932	.931	.943	.944
		15%	.950	.931	.944	.930	.941	.940	.946	.947
		20%	.949	.941	.946	.938	.946	.944	.948	.949
		25%		.947	.947	.945	.949	.948	.946	.947
	60	5%	.944							
		10%	.942	.946	.946	.945	.948	.948	.948	.948
		15%	.942	.947	.948	.949	.951	.951	.947	.948
		20%	.939	.949	.950	.950	.953	.953	.947	.947
		25%		.950	.950	.950	.952	.952	.946	.946
	100	5%	.948							
		10%	.950	.950	.951	.952	.952	.953	.947	.948
		15%	.949	.951	.948	.950	.952	.952	.948	.948
		20%	.950	.950	.949	.949	.950	.951	.950	.950
		25%		.950	.947	.948	.948	.948	.950	.950
0.8	20	5%	.950							
		10%	.957	.926	.928	.928	.932	.931	.943	.944
		15%	.956	.934	.950	.934	.944	.943	.949	.951
		20%	.956	.947	.951	.942	.949	.947	.953	.953
		25%		.948	.948	.946	.954	.952	.955	.955
	60	5%	.955							
		10%	.949	.949	.949	.947	.950	.950	.951	.951
		15%	.947	.950	.955	.952	.955	.957	.948	.948
		20%	.945	.957	.953	.957	.954	.957	.952	.952
		25%		.956	.953	.955	.956	.954	.953	.952
	100	5%	.949							
		10%	.949	.954	.956	.956	.956	.955	.951	.951
		15%	.946	.956	.952	.954	.954	.956	.950	.951
		20%	.948	.954	.951	.953	.951	.954	.950	.951
		25%		.951	.950	.949	.951	.951	.950	.950
1.4	20	5%	.950							
		10%	.965	**.924**	.926	.926	**.924**	**.923**	.946	.947
		15%	.964	.930	.955	.927	.939	.940	.954	.952
		20%	.963	.950	.948	.939	.946	.944	.958	.955
		25%		.953	.945	.943	.953	.950	.957	.959
	60	5%	.961							
		10%	.955	.949	.948	.944	.951	.949	.953	.953
		15%	.952	.951	.961	.949	.956	.958	.952	.952
		20%	.951	.960	.958	.961	.955	.958	.951	.949
		25%		.963	.956	.956	.957	.958	.953	.951
	100	5%	.958							
		10%	.957	.957	.957	.955	.957	.958	.954	.954
		15%	.952	.957	.955	.957	.956	.958	.952	.953
		20%	.952	.956	.955	.956	.953	.956	.953	.952
		25%		.954	.954	.956	.956	.956	.951	.952

Table 6. Ranks

N	Test	PES=0	PES=.2	PES=.5	PES=.8	PES=1.1	PES=1.4	**Total**
20	HQ	1	2	5	6	3	6	**23**
	HQ1	5	5	9	8	7	9	**43**
	HH3	0	0	3	3	4	3	**13**
	HQ2	6	4	8	7	4	5	**34**
	HH1	4	3	7	6	6	5	**31**
	HSK2	7	8	12	10	10	8	**55**
	HSK5	12	9	13	10	10	7	**61**
	Total	**35**	**31**	**57**	**50**	**44**	**43**	**260**
40	HQ	5	11	10	7	7	8	**48**
	HQ1	9	15	12	10	11	13	**70**
	HH3	7	13	13	5	10	10	**58**
	HQ2	8	5	7	9	8	11	**48**
	HH1	9	6	6	5	9	8	**43**
	HSK2	6	12	15	10	13	11	**67**
	HSK5	7	12	15	9	9	8	**60**
	Total	**51**	**74**	**78**	**55**	**67**	**69**	**394**
60	HQ	14	14	8	12	8	10	**66**
	HQ1	13	15	12	14	10	11	**75**
	HH3	13	15	9	10	8	6	**61**
	HQ2	12	14	10	10	9	10	**65**
	HH1	10	13	8	9	11	8	**59**
	HSK2	9	10	3	14	7	9	**52**
	HSK5	11	13	4	13	9	8	**58**
	Total	**82**	**94**	**54**	**82**	**62**	**62**	**436**
80	HQ	7	12	13	9	10	9	**60**
	HQ1	3	16	12	11	13	10	**65**
	HH3	8	16	15	11	8	11	**69**
	HQ2	14	9	8	10	12	14	**67**
	HH1	11	8	6	10	9	9	**53**
	HSK2	2	16	16	8	12	13	**67**
	HSK5	4	14	14	9	11	12	**64**
	Total	**49**	**91**	**84**	**68**	**75**	**78**	**445**
100	HQ	12	16	12	14	9	9	**72**
	HQ1	12	14	11	15	13	14	**79**
	HH3	13	14	13	12	10	11	**73**
	HQ2	16	15	11	12	10	9	**73**
	HH1	16	14	10	11	9	7	**67**
	HSK2	14	11	1	11	12	11	**60**
	HSK5	13	11	1	12	13	12	**62**
	Total	**96**	**95**	**59**	**87**	**76**	**73**	**486**
	GT	**313**	**385**	**332**	**342**	**324**	**325**	**2021**

Table 7. Total Number of Top Three Rankings for Each Test

HQ	HQ1	HH3	HQ2	HH1	HSK2	HSK5
269	332	274	287	253	301	305

three tables. (This is expected given the findings we previously enumerated.) One additional point important to mention is that the HSK2 and HSK5 hinge estimators methods as well as the A&K method resulted in well controlled coverage probabilities for the conditions where the affected procedures did not; that is, their coverage probabilities were not affected even though sample size was small ($n_1 = n_2 = 20$) and data were $g = .76 \text{ and } h = -.098$ distributed, for any percentage of total trimming.

Based on the preceding descriptions of our results, it would be difficult to try to pick out the 'best' one, two, or three methods for CIs around the robust PES. Indeed, Table 1 summary results indicate that all empirical values for all procedures were contained in the .925-.975 interval and accordingly, based on these results and the generally robust findings reported in Tables 2-5 (and those not tabled), specific recommendations would be challenging, and perhaps somewhat arbitrary, to make. Nonetheless, applied researchers usually like guidance from quantitative researchers regarding our recommendation of 'best' choice of procedure for their analyses. Accordingly, an even finer examination of our data was made.

In our second phase of analyses, the three hinge estimator methods for setting intervals having coverage probabilities closest to .95 were located; this was done for each combination of sample size, population distribution, total percentage of trimming and PES. Hinge estimator methods having identical empirical coverage probabilities received the same rank (either 1-closest, 2-next closest, or 3-third closest). Preferred ranks were given to deviations that were above .95 as opposed to below .95. Thus, if procedure 'A' resulted in a .951 coverage probability while procedure 'B' had coverage probability of .949, procedure A received the better rank -- the preference was for conservative rather than liberal values. Finally, any value that did not fall into a stringent criterion [($\pm 2\sigma_{1-\alpha}$ for $1-\alpha = .95$) i.e., .945-.955] was excluded from ranking.

Accordingly, in Table 6 the total number of top three rankings as a function of sample size and PES for the seven hinge estimator ES intervals are presented. What one can also see from Table 6 is that: (a) the total number of top three rankings, not surprisingly, increased with the size of sample; for $n_1 = n_2 = 20, 40, 60, 80, \text{ and } 100$, the total number of top three rankings was 260, 394, 436, 445, and 486, respectively; (b) the procedures were most disparate (range=48) from one another in terms of accuracy (i.e., number of top three rankings) when $n_1 = n_2 = 20 \text{ and } 40$ and were much more similar to one another when $n_1 = n_2 = 60, 80, \text{ and } 100$; and (c) the number of top three rankings increased with PES up until $PES = .2$ and then remained almost the same for $PES = .5\text{-}1.4$ Finally, the numbers presented in Table 6 and summarized in Table 7 indicate that HQ1 had the greatest number (332) of top three rankings while HSK2 and HSK5 had the second and third most top three rankings (301 and 305, respectively).

Discussion

Algina and Keselman (2003) and Algina et al. (2005) compared two estimates of ES and associated CIs in an independent two-groups design, in which either least squares or robust estimators were used and where the critical values used in computing the interval were based on either a theoretical or bootstrap distribution. The procedures were compared under different conditions of nonnormality and for various sample sizes and magnitudes of PES. It was found that probability coverage for the CI was only controlled when the interval used robust estimators (i.e., trimmed means and Winsorized variances) and the critical values of the interval were obtained via a bootstrap empirical distribution. The authors used *a priori* $2 \times 100\alpha$ % symmetric trimming to remove the biasing effects of skewed data and/or outlying values and only investigated $\alpha = .20$.

In an unrelated study, Keselman et al. (in press) found that tests for treatment group equality based on asymmetrically obtained trimmed means and Winsorized variances, resulted in exceptionally good Type I error control and power to detect effects in nonnormal heterogeneous one-way models. Consequently, it is believed that it would be possible to obtain more accurate probability coverage for intervals of ES in nonnormal models if the ES statistic was based on asymmetrically trimmed data. Accordingly, a Monte Carlo investigation was conducted to probe this hypothesis, varying population shape, magnitude of PES, sample size, and total percentage of trimming.

The results from the investigation clearly suggest that coverage probabilities for robust ES intervals were very well controlled under the conditions of nonnormality that were investigated. That is, only 56 of the 3840 empirical coverage probabilities (less than 1.5% of the values) did not fall within Bradley's (1978) criterion of .925-.975. And, these liberal values (i.e., intervals were too narrow), almost exclusively occurred when sample size was at the minimum value ($n_1 = n_2 = 20$) investigated. However, coverage probabilities, with the exception of two cases, were always within the Bradley interval once sample size reached our medium sample size condition ($n_1 = n_2 = 60$). Thus, based on these findings, any of the hinge estimators for setting a CI around a robust parameter of ES are recommended.

Nonetheless, in the interest of trying to separate the procedures in order to provide a more specific recommendation for researchers intending to set an interval around an ES statistic in a two-groups paradigm, a comparison of the hinge estimator ES intervals with a more stringent criterion was made, a criterion where a procedure would be judged robust if the empirical estimate did not fall outside a .944-.956 interval ($\pm 2\sigma_{1-\alpha}$ for $1-\alpha = .95$). Based on this more stringent criterion, the three hinge estimator methods were located having empirical coverage probabilities closest to .95. Specifically, it was found that HQ1, HSK2, and HSK5 had, respectively, the highest number of top three rankings: 332, 301, and 305. Accordingly, from the set of seven hinge estimator ES interval estimation procedures, any one of these three methods are recommended. Keselman et al. (in press) also recommended these three procedures for comparing treatment group trimmed means. Furthermore, the results suggest that, in general, one needs to have group sizes larger than 20 and that one can obtain good coverage with as little as 15% total trimming. The reader should remember however, that the differences between the empirical probabilities among these methods generally occurred in the third decimal place, and therefore, as stated, any of the seven hinge estimator approaches to setting an interval around the PES would be satisfactory, and in particular, much better than the usual approach of setting an interval around the nonrobust PES.

It was also found that *a priori* symmetric trimming provided very accurate probability coverage. All empirical coverage probabilities were within the Bradley (1978) liberal interval. Based on the summary values presented in Table 1, one can also note that the average probabilities are very tightly bunched around the target value of .95. Additionally, it is worth noting that, on average, researchers can obtain a very precise interval when adopting 5% symmetric trimming. Accordingly, the choice between *a priori* fixed trimming and asymmetric trimming methods might rest on ones comfort quotient for fixing the trimming rate prior to an examination of the data versus letting the data determine whether data should be trimmed in each tail of the data distribution and by what amount.

The comments provided by Keselman et al. (in press) regarding the choice of a best method of analysis are echoed. First, it needs to be repeated that no one method will be universally best. It could be that, at times, probability coverage for the classical method (i.e., Cohen's ES statistic) could provide a reasonable CI for ES. And as Wilcox and Keselman (2003) had noted, there is no way of knowing *a priori* which approach will be best. As they recommend, one could compute both approaches, that is, the classical approach and one of the robust methods enumerated in this paper. When the conclusions are the same, one can be comfortable with this common finding, otherwise, a robust approach to setting a CI for ES is recommended.

Keselman et al. noted that researchers should always carefully examine graphs of their data before proceeding with a particular method of analysis. Indeed, as many others have previously noted, a careful examination of outlying values can provide researchers with insights into the phenomenon under investigation.

It is reiterated that the parameter δ has a serious shortcoming because it is defined by using the usual population mean and standard deviation. These least squares parameters are not robust. While there are several criteria for assessing robustness of a parameter: qualitative robustness, quantitative robustness, and infinitesimal robustness (see Wilcox, 2005, Section 2.1 for a description of these criteria), the general notion is that *a parameter is not robust if a small change in the population distribution can strongly affect the parameter*. It can be shown that the least squares mean and variance are not robust (see, for example, Staudte and Sheather, 1990) when judged by any one of these three criteria. Accordingly, many authors, including us, subscribe to the position that inferences pertaining to robust parameters are more valid than inferences pertaining to the usual least squares parameters when dealing with populations that are nonnormal (e.g., Hampel, Ronchetti, Rousseeuw & Stahel, 1986; Huber, 1981; Staudte & Sheather, 1990; Wilcox & Keselman, 2003).

By itself, Cohen's δ, or any other ES (i.e., δ_R) for that matter, has little value in assessing whether or not a mean difference is large or small. What is required is experience in applying the ES. For example, as part of a review of the power of studies in abnormal and social psychology, Cohen (1962) suggested 0.25, 0.50, and 1.00 as small, medium, and large δs, respectively. In defense of these values, Cohen argued that the values "were chosen to seem reasonable." (p. 146) and cited three research studies on group differences in IQ research as justification for these guidelines. Cohen was clearly aware of the provisional nature of these guidelines and subsequently (Cohen, 1969) modified the guidelines to 0.20, 0.5, and 0.80, as small, medium, and large δs, respectively, and again emphasized that he regarded these to be reasonable based on his experience with research in the behavioral sciences. Cohen's guidelines, and his justification for them, illustrate an important point: Understanding of an ES measure will increase through experience with that measure.

References

Algina, J., & Keselman, H. J. (2003, May). *Confidence intervals for Cohen's effect size.* Paper presented at a conference in honor of H. Swaminathan, University of Massachusetts, Amherst.

Algina, J. Keselman, H. J., & Penfield, R. D. (2005). 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.). Washington, DC: Author.

Babu, J. G., Padmanabhan A. R., & Puri, M. P. (1999). Robust one-way ANOVA under possibly non-regular conditions. *Biometrical Journal, 41*(1978), 321-339.

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

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, *65*, 145-153.

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

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

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

Fidler, F., Thomason, N., Cumming, G., Finch, S., & Leeman, J. (2004). Editors can lead researchers to confidence intervals, but can't make them think. *Psychological Science, 15*, 119-126.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). *Robust statistics*. New York: Wiley.

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

Hill, M., & Dixon, W. J. (1982). Robustness in real life: A study of clinical laboratory data. *Biometrics, 38*, 377-396.

Hoaglin, D. C. (1985). Summarizing shape numerically: The g- and h distributions. In D. Hoaglin, F. Mosteller, & J. Tukey, (Eds.), *Exploring data tables, trends, and shapes* (pp. 461-513). 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.

Hogg, R. V. (1982). On adaptive statistical inferences. *Communications in Statistics: Theory and Methods, 11*, 2531-2542.

Huber, P. J. (1970). Studentizing robust estimates. In M. L. Puri (Ed.), *Nonparametric techniques in statistical inference*. London: Cambridge University Press.

Huber, P. J. (1972). Robust statistics: A review. *Annals of Mathematical Statistics, 43*, 1041-1067.

Huber, P. J. (1977). Discussion. *The Annals of Statistics*, *5*, 1090-1091.

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

Keselman, H. J., Wilcox, R. R., Lix, L. M., Algina, J., & Fradette, K. H. (in press). Adaptive robust estimation and testing. *British Journal of Mathematical and Statistical Psychology*.

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

Mudholkar, A., Mudholkar, G. S., & Srivastava, D. K. (1991). A construction and appraisal of pooled trimmed-t statistics. *Communications in Statistics: Theory and Methods, 20*, 1345-135.

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

Reed III, J. F. (1998). Contributions to adaptive estimation. *Journal of Applied Statistics, 25*, 651-669.

Reed III, J. F., & Stark, D. B. (1996). Hinge estimators of location: Robust to asymmetry. *Computer Methods and Programs in Biomedicine, 49*, 11-17.

Rosenberger, J. L., & Gasko, M. (1983). Comparing location estimators: Trimmed means, medians and trimean. In D. Hoaglin, F. Mosteller, & J. Tukey (Eds.), *Understanding robust and exploratory data analysis* (pp. 297-336). New York:.Wiley.

Rothman, K. J. (1975). Computation of exact confidence intervals for the odds ratio. *International Journal of Bio-Medical Computing, 6*, 33-39.

Rothman, K. J. (1978a). A show of confidence. *New England Journal of Medicine, 299*, 1362-1363.

Rothman, K. J. (1978b). Estimation of the confidence limits for the cumulative probability of survival in life table analysis. *Journal of Chronic Disease, 31*, 557-560.

SAS Institute Inc. (1999). *SAS/STAT user's guide, Version 7*, 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 and J. H. Steiger (Eds.), *What if there were no significance tests?* Hillsdale, NJ: Lawrence Erlbaum.

Stigler, S. M. (1977). Do robust estimators work with real data? *The Annals of Statistics*, *5*, 1055-1098.

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

Tukey, J. W., & McLaughlin, D. H. (1963). Less vulnerable confidence and significance procedures for location based on a single sample: Trimming/Winsorization. *Sankhya, Series A*, *25*, 331-352.

Wilcox, R. R. (1997). *Introduction to robust estimation and hypothesis testing.* San Diego: Academic Press.

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

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

Appendix 1

One question that might be asked about δ_R is whether it is necessary to multiply

$$\delta_R = \frac{\mu_{t2} - \mu_{t1}}{\sigma_W}$$

by .643 to obtain a robust parameter. The answer is, of course, no. When the multiplier is not used, the difference between the trimmed means is divided by the Winsorized standard deviation. By contrast, when using the multiplier, the difference between the trimmed means is divided by a rescaled Winsorized standard deviation $\left(\text{i.e.,}\ \sigma_W/.643\right)$.

The same multiplier would be applied to the sample ES and, as a result, *regardless of whether the multiplier is used, coverage probability is the same.* Therefore, our results have relevance to researchers who prefer to include the multiplier and researchers who prefer to exclude the multiplier. Incorporating the multiplier requires a different value for different levels of trimming. The multipliers for 10%, 15%, and 25% trimming would be $1/\sqrt{.824}$, $1/\sqrt{.734}$, $1/\sqrt{.537}$, respectively.

Appendix 2

Huber (1972) and Hogg (1974) noted that the best way of conceptualizing the unknown parameter $\theta\left(\alpha_1,\alpha_1\right)$ is that it is the population counterpart of $m\left(\alpha_1,\alpha_1\right)$. Hogg (1974, p. 920) indicated that in the one-sample case the statistic $[m(\alpha_1,\alpha_2)-\theta(\alpha_1,\alpha_2)]/s_{m(\alpha_1,\alpha_2)}$ has an approximate t-distribution with $h-1$ degrees of freedom if trimming is reasonably symmetric about the mode of a unimodal skewed distribution. Moreover, he noted that, even for fairly skewed situations, the distribution of this statistic will "probably be closer to this approximating distribution than the ratio $[m(\alpha)-\theta]/s_{m(\alpha)}$, which is the statistic based on a symmetrically trimmed mean. (p. 920)".