Psychological Methods
2008, Vol. 13, No. 2, 110–129

1082-989X/08/$12.00 DOI: 10.1037/1082-989X.13.2.110

A Generally Robust Approach for Testing Hypotheses and Setting Confidence Intervals for Effect Sizes

H. J. Keselman
University of Manitoba

James Algina
University of Florida

Lisa M. Lix
University of Manitoba

Rand R. Wilcox
University of Southern California

Kathleen N. Deering
University of British Columbia

Standard least squares analysis of variance methods suffer from poor power under arbitrarily small departures from normality and fail to control the probability of a Type I error when standard assumptions are violated. This article describes a framework for robust estimation and testing that uses trimmed means with an approximate degrees of freedom heteroscedastic statistic for independent and correlated groups designs in order to achieve robustness to the biasing effects of nonnormality and variance heterogeneity. The authors describe a nonparametric bootstrap methodology that can provide improved Type I error control. In addition, the authors indicate how researchers can set robust confidence intervals around a robust effect size parameter estimate. In an online supplement, the authors use several examples to illustrate the application of an SAS program to implement these statistical methods.

Keywords: classical/robust test statistics, classical/robust effect size statistics, classical/robust confidence intervals, online numerical examples, online computer program

Supplemental materials: http://dx.doi.org/10.1037/1082-989X.13.2.110.supp

Within the context of analysis of variance (ANOVA), it has been well established that heteroscedasticity, skewness, and outliers (i.e., a point, or points, far from the central point in the distribution) can have a devastating effect on both Type I errors and power. Indeed, even with large sample sizes, classic homoscedastic techniques can be unsatisfactory. Regarding unequal variances, Keselman, Huberty, et al. (1998) noted that ratios of largest to smallest variances of 8:1 are not uncommon in educational and psychological studies and can have deleterious effects on the performance of many classical test statistics, especially when paired with unequal sample sizes. Research has shown that the deleterious effects of (co)variance heterogeneity on the usual omnibus ANOVA F and linear contrast tests (Student's t), as well as setting confidence intervals (CIs) around effect size (ES) statistics, generally can be overcome by adopting Welch (1938, 1951)-type, or other nonpooled, statistics (see Brown & Forsythe, 1974; Keselman, Kowalchuk, & Lix, 1998; Kohr & Games, 1974; Lix & Keselman, 1998), that is, statistics that do not pool across heterogeneous sources of variability and where error degrees of freedom (*df*) are estimated from the sample data.

The biasing effects of nonnormality can also generally be overcome by adopting robust measures of central tendency and variability, that is, by using trimmed means and Winsorized (co)variances rather than the usual least squares estimators (see Algina, Keselman, & Penfield, 2005a; Lix & Keselman, 1998; Wilcox, 1997). Adopting a nonrobust

H. J. Keselman, Department of Psychology, University of Manitoba, Winnipeg, Canada; James Algina, Department of Educational Psychology, University of Florida; Lisa M. Lix, Department of Community Health Sciences, University of Manitoba; Rand R. Wilcox, Department of Psychology, University of Southern California; Kathleen N. Deering, Department of Health Care and Epidemiology, University of British Columbia, Vancouver, Canada.

This work was supported by the Social Sciences and Humanities Research Council of Canada, the Manitoba Health Research Council, and the Canadian Institutes of Health Research.

Correspondence concerning this article should be addressed to H. J. Keselman, Department of Psychology, University of Manitoba, 190 Dysart Road, Duff Roblin Building, Winnipeg, Manitoba R3T 2N2 Canada. E-mail: kesel@cc.umanitoba.ca

measure “can give a distorted view of how the typical individual in one group compares to the typical individual in another, and about accurate probability coverage, controlling the probability of a Type I error, and achieving relatively high power” (Wilcox, 1995, p. 66; see also Wilcox & Keselman, 2003). By substituting robust measures of location (e.g., trimmed mean) and scale (e.g., Winsorized variance) for the usual mean and variance, it should be possible to obtain test statistics which are relatively insensitive to the combined effects of variance heterogeneity and nonnormality. Many researchers subscribe to the position that inferences pertaining to robust parameters are more valid than inferences pertaining to the usual least squares parameters when they are dealing with populations that are nonnormal in form (e.g., Hample, Ronchetti, Rousseeuw, & Stahel, 1986; Huber, 1981; Staudte & Sheather, 1990; Wilcox, 2005). Indeed, as Marazzi and Ruffieux (1999) noted, “the (usual) mean is a difficult parameter to estimate well: the sample mean, which is the natural estimate, is very nonrobust” (p. 79). Tukey (1960) suggested that outliers are a common occurrence in distributions, and others have indicated that skewed distributions frequently depict psychological data (e.g., reaction time data).

When trimmed means are being compared, the null hypothesis pertains to the equality of population trimmed means, that is, the μ_ts, not the usual population mean μs. This is an important point for the reader to remember. Some readers may not want to compare the μ_ts; however, as just noted, strong arguments can be made for abandoning tests comparing the usual means in favor of methods that compare population trimmed means.

Indeed, a number of articles have demonstrated that one can generally achieve robustness to nonnormality and (co)-variance heterogeneity in unbalanced independent and correlated groups designs by using robust estimators with heteroscedastic test statistics and for setting CIs around ES parameters (Algina et al., 2005a; Keselman, Algina, Wilcox, & Kowalchuk, 2000; Keselman, Kowalchuk, & Lix, 1998). Further improvement in Type I error control is often possible by obtaining critical values for test statistics and CIs through bootstrap methods. Such improvement has been demonstrated with statistics and CIs for independent as well as correlated groups designs (Algina, Keselman, & Penfield, 2005b, 2006; Wilcox, Keselman, & Kowalchuk, 1998).

We use a benchmark of $.025 \leq \hat{\alpha} \leq .075$ ($\hat{\alpha}$ is the empirical rate of Type I error) to define a robust test, when the criterion of significance is set at $\alpha = .05$. That is, for a particular case of nonnormality and/or variance heterogeneity, if the empirical rate of Type I error is contained in this interval, we, as well as many others, consider the procedure to be insensitive (i.e., not substantially affected) to the assumption violation(s). However, this criterion (i.e., the length of the interval) is not universally accepted, and other researchers/writers use other criteria to assess robustness, for example, $\pm 2\sigma_\alpha$. That is, the issue of robustness, invariably, involves subjective decisions (e.g., How disparate do variances have to be before a distortion will occur in the probability of committing a Type I error? How much power should be sacrificed in order to ensure the rate of Type I error is maintained below $\alpha = .05$?, etc.). Nonetheless, the robust methods that we present often do provide better Type I error protection and/or power to detect effects than do the classical methods and therefore, we maintain, should seriously be considered by applied researchers. As well, we note that we are not discounting other methods of analysis—nonparametric, rank transformation, quantile distribution comparisons, and so on.

At the outset we want to be quite specific about our claim that by adopting nonpooled test statistics with trimmed means and Winsorized variances and a bootstrapping methodology researchers will “generally” obtain better Type I error protection, increased power to detect effects, and more precise CIs. In particular, we are not saying that the methods we are about to describe always result in these benefits. What we are saying though, is that, based on the empirical literature (as cited in this article), by adopting the procedures enumerated it will be the case that typically these outcomes will be observed compared with the outcomes (i.e., inflated or depressed rates of Type I error, decreases in power to detect effects, and inaccurate probability coverage for CIs) when pooled statistics and the usual least squares estimators are used and critical values are obtained from theoretical sampling distributions. For example, the Type I error rate for the methods we describe will not always be equal to .05 (for a 5% significance level), but typically will be closer to .05 than will the Type I error rate for classical methods. Furthermore, the empirical literature does not, and cannot, report results for every imaginable case of variance heterogeneity and nonnormality that researchers may encounter. Nonetheless, the cases that have been examined (i.e., variances that are quite disparate and cases of nonnormality including symmetric and asymmetric distributions) have included a diverse set of conditions that may be encountered in practice.

Other methods of analysis may, for a particular data set, provide better Type I error control and power to detect effects. That is, applied researchers have available to them a plethora of data analytic methods to apply in any one research investigation: the classical general linear model methods, nonparametric methods, rank transformation methods, as well as other robust methods of estimation and testing, and so on. In the ideal application of data analytic methods, a careful examination of the data regarding metric of measurement(s), shape of distribution(s), dispersion of the data, outlying values, and so on, should be undertaken by the data analyst—an ideal that we strongly endorse. However, our experience suggests that, more often than not, researchers apply a method of analysis that they are familiar with, such as omnibus and specific tests from one perspective of data analysis, for example, omnibus and contrast tests based on the classical general linear model perspective. We believe such a strategy is understandable since most data analysts would certainly not be fluent in the myriad of analysis strategies that could be applied to their data and thus would rely on systems that they are most familiar with—those taught in most introductory graduate-level programs in psychology and education departments. What we attempt to accomplish in this article is to present another framework for data analysis, methods of analysis (i.e., robust estimation and testing) that, though discussed for many decades, still are not very familiar to most applied researchers. Thus, we present methods that rely on another method of estimation (i.e., trimmed means) and general test statistic (i.e., one that does not pool variances). Lastly, we integrate into this framework a unified approach for setting robust CIs around ES statistics.

In this article, we elaborate on the benefits of adopting robust estimators and test statistics instead of the classical methods for testing hypotheses and for setting intervals around robust ES statistics, and we indicate when bootstrap methods are superior to classical methods for determining critical values for test statistics and/or for setting CI limits. These topics are discussed both for independent and correlated groups designs. We illustrate through a series of online resources (see the supplemental materials and http://home.cc.umanitoba.ca/~lixlm/home.cc.umanitoba.ca/~lixlm/) how researchers can obtain numeric results by using a new SAS program. As a further resource, Wilcox (2005) presents a nontechnical exposition of robust estimation and testing, by using R and S^+ software.

Comparing Two Independent Groups

Hypothesis Testing

To introduce the ideas in a simple context, we consider an example in which subjects are randomly assigned to two experimental treatments designed to affect beliefs about constructivist approaches to teaching mathematics. The data (the Y_is; $i = 1, \ldots, n_j$) are scores on a measure of subjects' engagement in the reading materials that constitute the treatments.

A commonly used approach for comparing the means for two different groups is the independent samples t test. The test is based on the following assumptions:

1a. The data used to estimate the mean for one group are statistically independent of the data used to estimate the mean for the other group.

1b. Within each of the groups, the data contributed by different sampling units are statistically independent of one another.

2. The data in each group are drawn from a normal distribution.

3. The two populations from which the samples are drawn have equal variances.

Histograms of the data for the two groups are presented in Figure 1, and various descriptive statistics are presented in Table 1. These results suggest several important features of the data. First, both the plots and the skew and kurtosis statistics indicate the data are not drawn form normal distributions. Second, both plots suggest the presence of outlying data points. (Users can also use box plots to identify outliers. See also Wilcox, 2005, pp. 99–101, for other outlier detection procedures.) Third, the standard deviations suggest the treatments affected the variability of the data. Thus, it appears that both the second and third assumptions have been violated.

Violation of the equal variance assumption can negatively impact the Type I error rate of the test (Ramsey, 1980). That is, when the sample sizes are equal but small, perhaps seven or fewer, and the null hypothesis is true, violating the equal

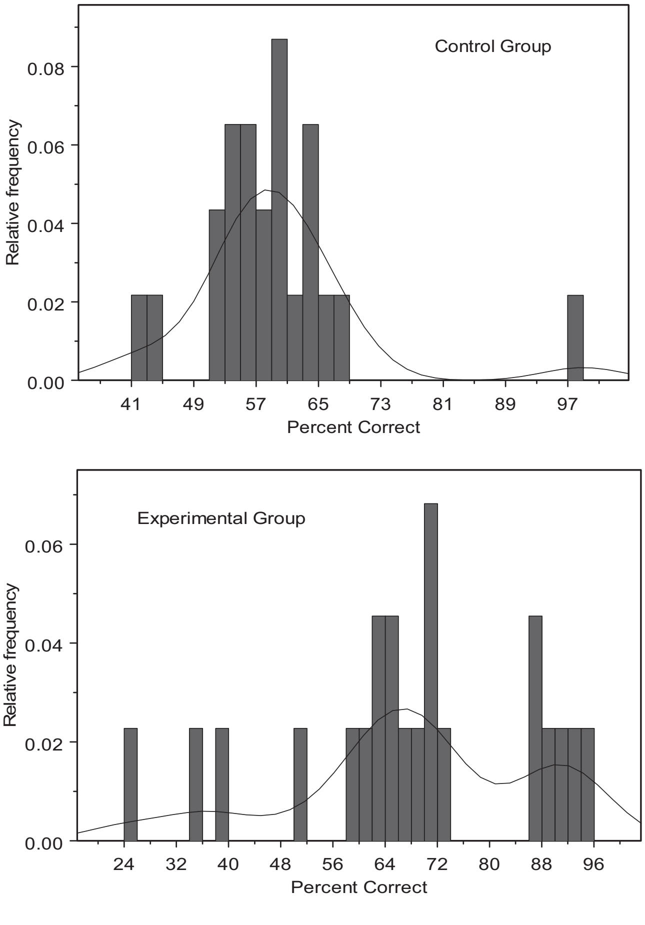

Figure 1. Histograms of percent correct scores for the control and experimental groups.

Table 1
Descriptive Statistics by Group

	Group	
Group	Control	Experimental
N	23	22
M	59.7	68.0
SD	10.8	19.1
Skew	2.0	–0.6
Kurtosis	8.0	0.1

variance assumption can result in a spuriously large *t* statistic and, consequently, the Type I error rate of the test can be larger than the alpha level used to assess the test for statistical significance. If the sample sizes are not equal, the null hypothesis is true, and the equal variance assumption is violated, then *t* tends to be spuriously large in magnitude when the larger sample size is drawn from the less variable population and spuriously small when the smaller sample size is drawn from the more variable population. These effects occur regardless of whether the sample sizes are small or large.

Violation of the normality assumption can also have a deleterious effect on the Type I error rate of the independent samples *t* test. Although the Type I error rate is widely viewed as being relatively unaffected by nonnormality, Bradley (1980) has pointed out conditions in which this is not true. Perhaps more important, when the data are drawn from nonnormal distributions, power can be increased by using an appropriate alternative to the independent samples *t* test. It seems shortsighted to use a familiar test (i.e., the *t* test) when another will provide more power.

Although the formula for the *t* statistic is well known, we report it here for purposes of comparison with another approach we will consider. The formula is

$$t = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\hat{\sigma}^2_{\mathrm{P}}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \tag{1}$$

where

$$\hat{\sigma}^2_{\mathrm{P}} = \frac{(n_1 - 1)\hat{\sigma}^2_1 + (n_2 - 1)\hat{\sigma}^2_2}{N - 2} \tag{2}$$

is the pooled estimate of error variance, n_j, $\hat{\mu}_j$, and $\hat{\sigma}^2_j$ are the sample size ($N = \Sigma_j n_j$), sample mean, and sample variance, respectively, for the *j*th treatment ($j = 1, 2$). Given the characteristics of the data in the present example, it is unlikely that the *t* statistic is a correct approach for analyzing the data. Nevertheless, for illustrative purposes, we conducted the test and found $t(43) = 1.79$, $p = .08$, indicating that the evidence is not sufficient to conclude there is a treatment effect.

An alternative to the independent sample *t* test that might be considered for these data is the Welch (1938)–James (1951) test, which is designed to address inequality of variance. For the two-group situation, the formula for the test statistic is

$$t_{\mathrm{WJ}} = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\frac{\hat{\sigma}^2_1}{n_1} + \frac{\hat{\sigma}^2_2}{n_2}}}. \tag{3}$$

The denominator for t_{WJ} allows for the possibility that the samples may have been drawn from populations with unequal variances. Comparison of the formulas for *t* and t_{WJ} indicates the difference between the two statistics is in the way the sample variances enter the statistics.

Several authors have presented methods for obtaining the *df* for t_{WJ} (Aspin, 1947; Welch, 1938, 1947). None of the formulas provide absolutely correct *df*s, and, thus, they are referred to as approximate *df* (ADF). The formula that is most widely used is due to Welch (1938):

$$\nu = \frac{\left(\frac{\sigma^2_1}{n_1} + \frac{\sigma^2_2}{n_2}\right)^2}{\frac{\sigma^4_1}{n^2_1(n_1 - 1)} + \frac{\sigma^4_2}{n^2_2(n_2 - 1)}}. \tag{4}$$

In practice, sample variances are substituted for the population variances. For the current example, $t_{\mathrm{WJ}}(32.9) = 1.79$, $p = .08$. Both the test statistic and the *p* value are very similar to those for the independent samples *t* test. This occurs because the sample sizes for the two groups are nearly equal.

If the equal variance assumption is violated but the data are sampled from normal distributions, the Welch (1938)–James (1951) statistic should provide a more adequate analysis than does the independent samples *t* test. However, which analysis is likely to be better when both assumptions are met? Although the independent samples *t* test has a slight advantage in terms of both the Type I error rate and power, the performance of the two tests is very similar. This is important because it has been shown (Wilcox, Charlin, & Thompson, 1986) that tests for variance homogeneity often do not have enough power to detect situations in which there is an advantage to using the Welch–James test (see also Wilcox, 2003, p. 298). Thus, if nonnormality were not a concern, we would argue that the Welch–James procedure should always be used in place of the independent samples *t* test.

Table 2 summarizes several approaches, including the Welch (1938)–James (1951) test that could be used to analyze the data in our example. As indicated by the table, the four alternatives differ in terms of whether least squares means and variances or trimmed means and Winsorized variances are used, and in terms of whether the *t* distribution

Table 2
Approaches to Data Analysis Based on Least Squares and Robust Estimators and Theoretical and Bootstrap Critical Values

Means and variances	p value	
	t distribution	Bootstrap
Least squares	t_{WJ}	t^*_{WJ}
Trimmed means and Winsorized variances	t_{WJ_t}	$t^*_{\mathrm{WJ}_t}$

or bootstrap methodology is used to compute the p value. In Table 2, a least squares mean refers to the common method for calculating a sample mean,

$$\hat{\mu}_j = \frac{\sum_{i=1}^{n_j} Y_{ij}}{n_j} \tag{5}$$

and a least squares variance refers to the common method for computing a variance

$$\hat{\sigma}_j^2 = \frac{\sum_{i=1}^{n_j} (Y_{ij} - \hat{\mu}_j)^2}{n_j - 1}. \tag{6}$$

On the other hand, to compute a trimmed mean, one removes an a priori determined percentage of the largest and smallest observations and computes the mean from the remaining observations.[1] If the target percentage to be removed is $2c$, the number of observations removed from each tail of the distribution is the integer that is just smaller than $c \times n_j$. Denote this integer by g_j. The smallest g_j observations are removed, as are the largest g_j observations. We denote a sample trimmed mean by $\hat{\mu}_{tj}$ and refer to it as the c% trimmed mean. A common trimming percentage is 20%, meaning that, in total, 40% of the data is trimmed. Based on various considerations, we recommend, and use as the default option in our SAS program, 20% symmetric trimming; see Wilcox (2005, p. 57) for a justification of 20% trimming.

When trimmed means are introduced to researchers, a common reaction is that one throws away data when the trimmed means are used and therefore the trimmed mean cannot be a better choice than the usual least squares mean. However, the fact of the matter is that the trimmed mean has advantages that the least squares mean does not. Consider a situation in which the sampled distribution is symmetric. Then both the trimmed mean and least squares mean estimate the same parameter, the population mean, and an important criterion for selecting between the two estimators is the standard error. It can be shown that when the data are sampled from a long-tailed distribution, the trimmed mean will have a smaller standard error than does the mean. Granted, if the data are sampled from a short-tailed or normal distribution, the mean will have a smaller standard error. However, the advantage for the mean under these conditions is often not large, whereas the advantage for the trimmed mean can be substantial when the data are sampled from a long-tailed distribution. Thus, one argument for the trimmed mean is that it can have a substantial advantage in terms of accuracy of estimation when sampling from long-tailed symmetric distributions, although it has a slight disadvantage when sampling from normal and short-tailed distributions.

Additionally, the trimmed mean is preferable when the data are subject to outliers. In contrast, the mean is not resistant to outliers and a single outlying observation can destroy its effectiveness as an estimator. A 20% trimmed mean will work well provided that fewer than 40% of the observations are outliers. If the distribution from which the data are sampled is skewed, the argument against the mean is that it is not a good indicator of the central point of the distribution because the extreme scores drag it away from the center. Because the trimmed distribution will be more nearly symmetric than is the nontrimmed distribution, the trimmed mean will be a better indicator of the center of the distribution. We find these arguments for using the trimmed mean, rather than the mean, persuasive. Nevertheless, some may not. However, researchers who use the mean should be aware that they are using an estimator that has poor accuracy (in some situations), is not resistant to outliers, and may not be a good indicator of the center of the distribution.

To compute a Winsorized variance, the smallest nontrimmed score replaces the scores trimmed from the lower tail of the distribution, and the largest nontrimmed score replaces the scores removed from the upper tail. The nontrimmed and replaced scores are called Winsorized scores. A Winsorized mean is calculated by applying the usual formula for the mean to the Winsorized scores, and a Winsorized variance is calculated as the sum of squared deviations of Winsorized scores from the Winsorized mean divided by $n_j - 1$. The Winsorized variance is used because it can be shown that the standard error of a trimmed mean is a function of the Winsorized variance (Wilcox, 2005). We denote a sample Winsorized variance by $\hat{\sigma}^2_{\mathrm{W}_j}$. See Wilcox (2003) for formulas corresponding to our verbal descriptions of the trimmed mean and Winsorized variance.

[1] Note that the class of trimmed means contains the median as a special case. However, a concern about the median is that it trims too much, possibly resulting in relatively low power. This occurs under normality and has the potential of occurring when outliers tend to be rare. By trimming 20%, good power is obtained under normality, and power can remain relatively high when sampling from heavy-tailed distributions.

Let

$$\tilde{\sigma}^2_{\mathrm{W}_j} = \frac{(n_j - 1)\hat{\sigma}^2_{\mathrm{W}_j}}{h_j - 1} \tag{7}$$

be the scaled Winsorized variance, where h_j stands for the effective sample size, that is, the size after trimming the data. In computing the robust Welch–James statistic with robust estimators, the quantity $\tilde{\sigma}^2_{\mathrm{W}_j}$ plays the role that $\hat{\sigma}^2_j$ plays in computing the Welch (1938)–James (1951) statistic based on least squares estimators. Specifically, the robust Welch–James statistic is

$$t_{\mathrm{WJ}_\mathrm{t}} = \frac{\hat{\mu}_{\mathrm{t}1} - \hat{\mu}_{\mathrm{t}2}}{\sqrt{\dfrac{\tilde{\sigma}^2_{\mathrm{W}_1}}{h_1} + \dfrac{\tilde{\sigma}^2_{\mathrm{W}_2}}{h_2}}}. \tag{8}$$

If the t distribution is used to compute the p value, then the estimated dfs are

$$\hat{v}_\mathrm{t} = \frac{\left(\dfrac{\tilde{\sigma}^2_{\mathrm{W}_1}}{h_1} + \dfrac{\tilde{\sigma}^2_{\mathrm{W}_2}}{h_2}\right)^2}{\dfrac{\tilde{\sigma}^4_{\mathrm{W}_1}}{h_1^2(h_1 - 1)} + \dfrac{\tilde{\sigma}^4_{\mathrm{W}_2}}{h_2^2(h_2 - 1)}}. \tag{9}$$

We emphasize that the approach adopted here is based on an a priori trimming percentage. Inspecting the data for outliers, deleting the data points identified as outliers, and then using a procedure such as the independent samples t test or the Welch–James test is likely to result in misleading inferences because the procedure fails to take into account the post hoc nature of the trimming. In particular, the standard errors of the mean differences (i.e., the denominator of t or t_{WJ}) will be incorrect because these standard errors were developed for situations in which trimming does not take place.

The trimmed means ($\hat{\mu}_{\mathrm{t}j}$),Winsorized standard deviations ($\hat{\sigma}_{\mathrm{W}_j}$), and scaled Winsorized standard deviations ($\tilde{\sigma}_{\mathrm{W}_j}$), for the data shown in Figure 1, are reported in Table 3. The mean difference is 68.0 – 59.7 = 8.3 and the trimmed mean difference is 69.8 – 59.8 = 10.0. Thus, trimming results in a larger estimate of the treatment effect (expressed on the scale for the data). In addition, the Winsorized variances are smaller than the least squares variances, primarily because trimming removed the influence of the outliers in the data. When $t_{\mathrm{WJ}_\mathrm{t}}$ is used and the p value is computed by using the t distribution, $t_{\mathrm{WJ}_\mathrm{t}}(15.52) = 2.68$ and $p = .02$. Thus, using trimmed means and scaled Winsorized variances (i.e., $t_{\mathrm{WJ}_\mathrm{t}}$) results in sufficient evidence to conclude there is a treatment effect, whereas using t_{WJ} did not. This result is consistent with results in the literature demonstrating superior power when trimmed means and Winsorized variances are used.[2]

Table 3
Descriptive Statistics by Group

Group	N	Trimmed M	Winsorized SD	Scaled Winsorized SD^a
Control	22	58.8	3.77	4.73
Experimental	23	69.8	11.51	14.63

[a] $\tilde{\sigma}_{\mathrm{W}_j}$.

Thus far we have illustrated inferential procedures that differ in terms of the means and variances used to compute the t statistic but which use the theoretical t distribution to obtain the p value. The other two alternatives enumerated in Table 2 also differ in terms of the means and variances used to compute the t statistic, but in each a nonparametric bootstrap is used to compute the p value. The strategy behind our use of the bootstrap in hypothesis testing is to shift the sample distributions of the scores for each group and variable by subtracting the group mean (least squares or trimmed) from each score and using the shifted empirical distributions to estimate an appropriate critical value. In particular, for each j, obtain a bootstrap sample by randomly sampling with replacement n_j observations from the shifted values, yielding $Y_1^*, \ldots, Y_{n_j}^*$. Let t^* be the value of a test statistic (t_{WJ} or $t_{\mathrm{WJ}_\mathrm{t}}$) based on the bootstrap sample. For a two-tailed test, the B values of $|t^*|$, where B represents the number of bootstrap simulations, are put in ascending order, that is, $|t^*_{(1)}| \leq \cdots \leq |t^*_{(B)}|$, and an estimate of an appropriate critical value is $|t^*_{\mathrm{WJ}_{\mathrm{t}(q)}}|$, $q = (1 - \alpha)B$, rounded to the nearest integer. One will reject H_0: $\mu_1 = \mu_2$ or H_0: $\mu_{\mathrm{t}1} = \mu_{\mathrm{t}2}$ when $|t| \geq |t^*_{(q)}|$, where $|t|$ is the value of the heteroscedastic statistic (t_{WJ} or $t_{\mathrm{WJ}_\mathrm{t}}$) based on the original nonbootstrapped data. Recall that $t_{\mathrm{WJ}} = 1.79$ and $t_{\mathrm{WJ}_\mathrm{t}} = 2.68$. Applying the bootstrap procedure to compute the p value for $t_{\mathrm{WJ}} = 1.79$ yields $p = .08$, the same as was obtained by using the t distribution. Applying the bootstrap procedure to $t_{\mathrm{WJ}_\mathrm{t}}$ yields $p = .01$. Thus, for $t_{\mathrm{WJ}_\mathrm{t}}$, the p value is smaller when computed by the bootstrap than when obtained by using the theoretical t distribution (see Footnote 2).

To sum up, applying the independent samples t test or the Welch (1938)–James (1951) test resulted in nonsignificant evidence for a treatment effect. For the Welch–James test, this result was obtained both when the p value was obtained by using the t distribution and by using the bootstrap. Applying the Welch–James test to trimmed means and Winsorized variances resulted in significant evidence of a

[2] We cannot emphasize enough that the results of our numerical example should not be taken to mean that with all data sets the relative outcomes between classical methods and robust methods will always lead to a similar conclusion, that is, results favoring the robust method of analysis. Moreover when we say our numerical results are consistent with the statistical literature in regard to how classical and robust methods perform, we do not mean that our numerical results prove that the statistical literature is correct.

treatment effect, both when the p value was obtained by using the t distribution and by using the bootstrap. The p value computed by using the bootstrap was slightly smaller. This difference, obtained by using trimmed means and Winsorized variances, most likely occurred because of the outliers in the data. As well, the difference resulting from using the bootstrap most likely occurred because the data appear to be drawn from nonnormal distributions, particularly for the experimental treatment group.

CIs for ESs

Since the 1960s, recommendations to report an ES in addition to, or in place of, a hypothesis test have been proposed (e.g., Cohen, 1965; Hays, 1963). In perhaps the past 15 years or so, there has been renewed emphasis on reporting ESs because of editorial policies (e.g., Murphy, 1997; Thompson, 1994) and official support for the practice. The practice of reporting ESs has also received support from the American Psychological Association Task Force on Statistical Inference (Wilkinson and the Task Force on Statistical Inference, 1999). 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). 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 (see also Smithson, 2003; Steiger, 2004). Bird (2002) presented software for calculating approximate CIs for a wide variety of ANOVA designs.

A commonly reported ES for measuring the size of effect between two treatment groups is Cohen’s (1965) d:

$$d = \frac{\hat{\mu}_2 - \hat{\mu}_1}{\hat{\sigma}_P}. \tag{10}$$

In this context we refer to $\hat{\sigma}_P$ as the *standardizer*.[3,4] Cohen’s d estimates

$$\delta = \frac{\mu_2 - \mu_1}{\sigma}, \tag{11}$$

where σ, the population standard deviation, is assumed to be equal for both groups.[5] (Table 4 summarizes the ES estimators and parameters defined in this article.)

Researchers might consider $\hat{\sigma}_P$ and σ to be the “natural” standardizers for ES. However, there is no one universally appropriate standardizer. Indeed, as Kline (2004, pp. 97–98) noted, “There is more than one possible population standard deviation for a comparative study.” For example, the standardizer “could be the standard deviation in just one of the populations [*the view taken by* Glass et al. (1965)]” (italicized words are our own), and with regard to the sample estimator, Kline noted that an estimate of the standardizer “is not the same in all kinds of d statistics.” This is important to keep in mind as we define our robust ES parameters and sample estimators. It further highlights that reporting ESs can be as complicated, and potentially as fraught with problems, as is the use of test statistics for assessing the presence of a treatment. Indeed, as we are about to illustrate, defining and estimating an ES is particularly problematic when variances are not homogeneous.

In addition to other measures of ES that have received attention in the literature, the American Psychological Association Task Force on Statistical Inference (Wilkinson and the Task Force on Statistical Inference, 1999) indicated that ES may be expressed in dependent variable units when such units are interpretable. Thus, our accompanying SAS program also allows users to obtain ES estimators expressed in dependent variable units. Our view, however, is that scales of measurement in psychological research are, with few exceptions, arbitrary, and thus standardized measures of ES are appropriate. However, when the scale is not arbitrary (e.g., reaction time data recorded in milliseconds), a nonstandardized ES may be computed (i.e., the value of a contrast among means). Indeed, when it is appropriate to report an ES in dependent variable units, one no longer faces the issue of what is the correct standardizer.

It is known (see, e.g., Cumming & Finch, 2001; Steiger & Fouladi, 1997) that when the sample data are drawn from normal distributions, the variances of the populations are equal, and the scores are independently distributed, then an exact CI for the population ES (i.e., δ) can be constructed by using the noncentral t distribution. For the current example, $d = 0.54$; and a 95% CI for $\delta = [-0.06, 1.13]$. The value for d suggests the effect is of moderate size using Cohen’s guidelines. Furthermore, the CI indicates the ES is imprecisely estimated and is consistent with an ES ranging from a trivial value to a very large value.

Cohen’s (1965) ES is based on the homogeneity of variance assumption. When this assumption is likely to be

[3] Cohen used the Latin letter d to refer to the population ES. Following more typical practice, we use d to refer to the sample ES and the Greek letter δ to refer to the population ES.

[4] Other, less popular, measures of ES have been proposed by Cliff (1993, 1996); Hedges and Olkin (1985); Kraemer and Andrews (1982); McGraw and Wong (1992); Vargha and Delaney (2000); and Wilcox and Muska (1999); see Hogarty and Kromrey (2001) for the definitions of these procedures.

[5] The reader should remember that d, as given in Equation 10, and δ, as given in Equation 11, assume variance homogeneity and, thus, given equal group sizes, Equation 11 can be written as

$$\delta = \frac{\mu_2 - \mu_1}{\sqrt{(\sigma_1^2 + \sigma_2^2)/2}},$$

while Equation 10 could be written as

$$d = \frac{\hat{\mu}_2 - \hat{\mu}_1}{\sqrt{(\hat{\sigma}_1^2 + \hat{\sigma}_2^2)/2}}.$$

Table 4
Effect Size Statistics and Parameters

Estimator	Eq	Parameter	Eq	Comments
Independent Groups Designs				
$\hat{\mu}_2 - \hat{\mu}_1$	*	$\mu_2 - \mu_1$	*	A nonstandardized measure of ES expressed in the original measurement units. When the measurement scale is not arbitrary (e.g., reaction time), this would be an appropriate measure of ES and the issue of a standardizer is no longer relevant.
$d = \frac{\hat{\mu}_2 - \hat{\mu}_1}{\hat{\sigma}_{\mathrm{P}}}$	10	$\delta = \frac{\mu_2 - \mu_1}{\sigma}$	11	Cohen's (1965) ES estimator and parameter that assumes equality of variances. When homogeneity is violated, and the sample sizes are unequal, Cohen's estimator will be biased.
$\frac{\hat{\mu}_2 - \hat{\mu}_1}{\sqrt{\frac{n_1\hat{\sigma}_1^2 + n_2\hat{\sigma}_2^2}{N}}}$	14	$\frac{\mu_2 - \mu_1}{\sqrt{\frac{n_1\sigma_1^2 + n_2\sigma_2^2}{N}}}$	15	For the case of unequal variances, Kulinska and Staudte (2006) suggested a weighted sum of the group variances (i.e., variances are pooled). The estimate and parameter is sample size (design) dependent.
$\hat{\delta}^* = \frac{\hat{\mu}_2 - \hat{\mu}_1}{\sqrt{\frac{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}{2}}}$	16	$\delta^* = \frac{\mu_2 - \mu_1}{\sqrt{\frac{\sigma_1^2 + \sigma_2^2}{2}}}$	*	This alternative uses an unweighted average of the variances as the standardizer. This estimator and parameter have been popular choices for data analysts. Note, however, that variances are pooled.
$\hat{\delta}_j = \frac{\hat{\mu}_2 - \hat{\mu}_1}{\hat{\sigma}_j}$	17	$\delta_j = \frac{\mu_2 - \mu_1}{\sigma_j}$	18	The Glass et al. (1981) estimator and parameter. A nonpooled standardizer, e.g., $\hat{\sigma}_1$ or $\hat{\sigma}_2$, or both, is used. It is this approach that is preferred by the authors of this article.
$\hat{\delta}_{R_j} = .642\,\frac{\hat{\mu}_{\mathrm{t}2} - \hat{\mu}_{\mathrm{t}1}}{\hat{\sigma}_{\mathrm{W}_j}}$	**19**	$\delta_{R_j} = .642\,\frac{\mu_{\mathrm{t}2} - \mu_{\mathrm{t}1}}{\sigma_{\mathrm{W}_j}}$	**20**	Robust versions of (17) and (18); that is, the estimator and statistic are based on trimmed means and a Winsorized variance. A bootstrap critical value (CV) is recommended. This is our recommended approach.
$\hat{\delta}_{R_j} = .642\,\frac{\hat{\psi}}{\hat{\sigma}_{\mathrm{W}_j}}$	*	$\delta_{R_j} = .642\,\frac{\Psi}{\sigma_{\mathrm{W}_j}}$	*	Generalizations of (19) and (20), allowing for a nonpairwise contrast — $\Psi = c_1\mu_1 = c_2\mu_2 + \ldots + c_J\mu_J$ — where $\Sigma_j c_j = 0$ in the numerator.
$\hat{\delta}_R^* = \frac{\hat{\mu}_{\mathrm{t}2} - \hat{\mu}_{\mathrm{t}1}}{\sqrt{\frac{\hat{\sigma}_{\mathrm{W}_1}^2 + \hat{\sigma}_{\mathrm{W}_2}^2}{2}}}$	22	$\delta_R^* = \frac{\mu_{\mathrm{t}2} - \mu_{\mathrm{t}1}}{\sqrt{\frac{\sigma_{\mathrm{W}_1}^2 + \sigma_{\mathrm{W}_2}^2}{2}}}$	21	Robust analogues of (16) and δ^*. A bootstrap CV is recommended. However, (Winsorized) variances are pooled.
Correlated Groups Designs				
$\hat{\mu}_{\mathrm{t}2} - \hat{\mu}_{\mathrm{t}1}$	*	$\mu_{\mathrm{t}2} - \mu_{\mathrm{t}1}$	*	A nonstandardized measure of ES expressed in the original measurement units.
$\hat{\delta}_R^{\dagger} = .642\,\frac{\hat{\mu}_{\mathrm{t}2} - \hat{\mu}_{\mathrm{t}1}}{\sqrt{\frac{\hat{\sigma}_{\mathrm{W}_1}^2 + \hat{\sigma}_{\mathrm{W}_2}^2}{2}}}$	28	$\delta_R^{\dagger} = .642\,\frac{\mu_{\mathrm{t}2} - \mu_{\mathrm{t}1}}{\sqrt{\frac{\sigma_{\mathrm{W}_1}^2 + \sigma_{\mathrm{W}_2}^2}{2}}}$	27	Analogues of (21) and (22). A bootstrap CV is recommended. However, (Winsorized) variances are pooled.
$\hat{\delta}_k = \frac{\hat{\mu}_2 - \hat{\mu}_1}{\hat{\sigma}_k}$	29	$\delta_k = \frac{\mu_2 - \mu_1}{\sigma_k}$	30	Estimator and parameter based on least squares values and a nonpooled standardizer. A bootstrap CV is recommended.
$\hat{\delta}_{R_k} = .642\,\frac{\hat{\mu}_{\mathrm{t}2} - \hat{\mu}_{\mathrm{t}2}}{\hat{\sigma}_{\mathrm{W}_k}}$	**32**	$\delta_{R_k} = .642\,\frac{\mu_{\mathrm{t}2} - \mu_{\mathrm{t}2}}{\sigma_{\mathrm{W}_k}}$	**31**	Analogues of (19) and (20). That is, the standardizer uses a single Winsorized variance. A bootstrap CV is recommended. Our analogue of the Glass et al. (1981) approach.
$\hat{\delta}_{R_k} = .642\,\frac{\hat{\psi}}{\hat{\sigma}_{\mathrm{W}_k}}$	*	$\delta_{R_k} = .642\,\frac{\Psi}{\sigma_{\mathrm{W}_k}}$	*	Generalizations of (31) and (32), allowing for a nonpairwise contrast — $\Psi = c_1\mu_1 = c_2\mu_2 + \ldots + c_K\mu_K$ — where $\Sigma_k c_k = 0$ in the numerator.

Note. * denotes an equation that is implied, not defined, in the article. Estimators and parameters whose equation numbers are bolded are those procedures that we recommend.

violated and the sample sizes are unequal, Cohen's ES is not appropriate because the parameter one estimates changes depending on the sample sizes. This problem occurs because the standardizer is the pooled standard deviation and estimates

$$\sqrt{\frac{(n_1 - 1)\sigma_1^2 + (n_2 - 1)\sigma_2^2}{N - 2}}. \tag{12}$$

For the case of unequal variances, Kulinska and Staudte (2006) suggested a squared ES, standardizing by a weighted sum of the group variances:

$$\hat{\theta} = \frac{(\hat{\mu}_2 - \hat{\mu}_1)^2}{\dfrac{n_1\hat{\sigma}_1^2 + n_2\hat{\sigma}_2^2}{N}}. \tag{13}$$

The Kulinska and Staudte ES is a squared quantity but can be converted to a nonsquared quantity:

$$\sqrt{\hat{\theta}} = \frac{\hat{\mu}_2 - \hat{\mu}_1}{\sqrt{\dfrac{n_1\hat{\sigma}_1^2 + n_2\hat{\sigma}_2^2}{N}}}, \tag{14}$$

which estimates

$$\frac{\mu_2 - \mu_1}{\sqrt{\dfrac{n_1\sigma_1^2 + n_2\sigma_2^2}{N}}}. \tag{15}$$

We believe that general use of the Kulinska and Staudte ES is unwise, for the same reason we do not recommend the use of Cohen's ES statistic when population variances are unequal. An obvious alternative to d uses the unweighted average of the variances as the standardizer:

$$\hat{\delta}^* = \frac{\hat{\mu}_2 - \hat{\mu}_1}{\sqrt{\dfrac{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}{2}}}. \tag{16}$$

Of the two, $\hat{\delta}^*$ seems the most useful in that it does not result in an ES whose population value depends on the design.

An alternative definition of the ES, based on the standard deviation for the jth group, is

$$\hat{\delta}_j = \frac{\hat{\mu}_2 - \hat{\mu}_1}{\hat{\sigma}_j}, \tag{17}$$

which estimates

$$\delta_j = \frac{\mu_2 - \mu_1}{\sigma_j}. \tag{18}$$

One could use the standard deviation (standardizer) for either group. Referring to a case in which $\hat{\delta}_1$ and $\hat{\delta}_2$ are not equal, 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). In writing about the choice between $\hat{\delta}_j$ and an ES defined as the mean difference divided by the average standard deviation, Glass et al. wrote "*the average standard deviation should probably be eliminated as a mindless statistical reaction to a perplexing choice*" (p. 106; italicized words indicate our own emphasis). Although we recognize that the selection of one of the two standard deviations could be, in many cases, a difficult substantive decision, our preference, like that of Glass et al., is for $\hat{\delta}_j$. Regardless of whether $\hat{\delta}_j$ or $\hat{\delta}^*$ is used, the same issues arise for construction of a CI: Should the CI be based on a distributional assumption or should the bootstrap be used? Thus far the relative performance of these alternative methods for constructing CIs has not been investigated in connection with $\hat{\delta}^*$.[6]

When $\hat{\delta}_j$ is used, the CI for d is no longer correct. As with hypothesis testing, there are two alternatives for constructing a CI based on $\hat{\delta}_j$. Under the assumptions that the data in each group are normally distributed, and all data are distributed independently, an approximate CI for δ_j based on a noncentral t distribution can be derived. However, the interval is based on normal distribution theory. This normality assumption is likely to be problematic because $\hat{\mu}_2 - \hat{\mu}_1$ and $\hat{\sigma}_j$ are not independent when the distribution is skewed for the jth treatment. For example, if the distribution is positively skewed for the first treatment, the correlation between $\hat{\mu}_2 - \hat{\mu}_1$ and $\hat{\sigma}_1$ will be negative. Therefore, large values for $\hat{\mu}_2 - \hat{\mu}_1$ will tend to be associated with small values for $\hat{\sigma}_1$, 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 (Algina et al., 2006). (Kelly, 2005, presented similar results for the CI for Cohen's δ based on a noncentral t distribution.)

An alternative to the noncentral t-based approximate CI for δ_j is to use the nonparametric percentile bootstrap

[6] As we indicate, the choice of standardizer is not a statistical issue but rather a conceptual one. Also, though our preference is to compute a robust version of $\hat{\delta}_j$, we, nonetheless, recognize that others might want to compute $\hat{\delta}^*$. Thus, our SAS program (SAS Institute, 1999) allows users to compute this statistic for each design illustrated in the article. In particular, the program will compute the standardizer as the average of the variances over the cells involved in the contrast. Furthermore, the program uses our bootstrap methodology to set a CI around the unknown parameter $\hat{\delta}^*$. However, we once again remind the reader that, though we suspect this CI would be generally robust to heterogeneity, there is no proof that this is indeed the case.

method to construct a CI for δ_j.[7] To apply the percentile bootstrapping method for setting CIs around a parameter of ES, first, a random sample with replacement of size n_1 is selected from the scores for the first group. Second, a random sample with replacement of size n_2 is selected from the scores for the second group. These two samples are combined to form a bootstrap sample. The ES (i.e., $\hat{\delta}_j$) is then calculated from the bootstrap sample. The ES estimates for all bootstrap samples are ranked from low to high. The lower limit of the $100(1 - \alpha)\%$ CI is determined by finding the $[\langle B(\alpha/2)\rangle + 1]$th estimate in the rank order, where $\langle B(\alpha/2)\rangle$ indicates rounding $B(\alpha/2)$ to the nearest whole number; the upper limit would be determined by finding the $[B - \langle B(\alpha/2)\rangle]$th estimate in the rank order.[8]

Applying $\hat{\delta}_1$ in the present example, the ES is 0.77 and indicates the mean difference is about eight tenths of a standard deviation for the first treatment. The ES $\hat{\delta}_2$ is 0.44 and indicates the mean difference is a bit larger than four tenths of a standard deviation for the second treatment. These results reflect the fact that data are much less variable for the first treatment group. The ES $\hat{\delta}^* = 0.54$ and is approximately midway between $\hat{\delta}_1$ and $\hat{\delta}_2$. Ninety-five percent CIs, constructed by using the noncentral *t* distribution and the percentile bootstrap, are [–0.11, 1.64]; and [–0.08, 2.71]; respectively for δ_1. The wider interval constructed by using the bootstrap is consistent with results in Algina et al. (2006) indicating that when the second group was more variable than the first, and data were sampled from nonnormal distributions, coverage probability was too small for both the noncentral *t* distribution based CI and the percentile bootstrap CI, but coverage probability was worse for the noncentral *t* distribution based CI. The 95% CIs for δ_2 are [–0.06, 0.92]; and [–0.04, 1.11]. In addition, a 95% percentile bootstrap CI for δ^* is [–0.04, 1.43]. It should be noted that despite the differences in ESs and the CIs, the results are consistent with hypothesis testing using least square means and variances: There is not sufficient evidence to claim a treatment effect (see Footnote 2).

Wilcox and Keselman (2003) argued that the common population definition and sample estimate of ES (i.e., δ and *d* or δ_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 population ES value and its estimate to be grossly misleading (Wilcox, 2003, Section 8.11). One reasonable alternative is to replace the least squares estimates in $\hat{\delta}_j$ with robust estimates. Accordingly, in place of $\hat{\delta}_j$, we suggest using

$$\hat{\delta}_{R_j} = .642\,\frac{\hat{\mu}_{t2} - \hat{\mu}_{t1}}{\hat{\sigma}_{W_j}}, \tag{19}$$

where .642 is the Winsorized standard deviation for a 20% trimmed standard normal distribution. The ES $\hat{\delta}_{R_j}$ estimates

$$\delta_{R_j} = .642\,\frac{\mu_{t2} - \mu_{t1}}{\sigma_{W_j}}. \tag{20}$$

For a normal distribution, both $\hat{\delta}_{R_j}$ and $\hat{\delta}_j$ converge to δ_j as the sample sizes increase.[9] Again there are two alternatives for constructing a CI, a CI constructed by using the noncentral *t* distribution and a CI constructed by using the percentile bootstrap. Algina et al. (2006) found, however,

[7] Other bootstrapping methods are available such as the bias-corrected and accelerated bootstrap. Because all of the research that we are familiar with uses the percentile bootstrap, that is the method that is employed in our SAS program. Future work could result in our adopting another bootstrap method.

[8] To elaborate suppose $B = 1{,}000$ ES estimates are ranked from low to high. The lower limit of the CI is determined by finding the 26th estimate in the rank order (i.e., the estimate just larger than the $0.025 \times 1{,}000$th estimate); the upper limit is determined by finding the 975th estimate (i.e., the $0.975 \times 1{,}000$th estimate). Use of the $[B(\alpha/2) + 1]$th and $B[(1 - \alpha)/2]$th values as the percentiles is based on Wilcox (2003, p. 211) and defines the 95% CI as the range of the bootstrap distribution containing the middle 950 bootstrap estimates.

[9] One question that might be asked about δ_{R_j} is whether it is necessary to multiply

$$\frac{\mu_{t2} - \mu_{t1}}{\sigma_{W_j}}$$

by .642 to obtain a robust parameter. The answer is no. In

$$\frac{\mu_{t2} - \mu_{t1}}{\sigma_{W_j}},$$

the difference between the trimmed means is divided by the Winsorized standard deviation. By contrast, for δ_{R_j} the difference between the trimmed means is divided by a rescaled Winsorized standard deviation. The rescaled Winsorized standard deviation is a consistent estimator of the usual standard deviation

$$\sigma = \sqrt{E(X - \mu)^2}$$

when the data are normal. This kind of rescaling is not unusual in robust statistics. For example, the median absolute deviation is often divided by .6745 so that it estimates the usual standard deviation when the data are sampled from a normal distribution (see Wilcox, 2005). Deleting .642 would make the interpretation of $\mu_{t2} - \mu_{t1}/\sigma_{W_j}$ analogous to that of δ_{R_j}: $\mu_{t2} - \mu_{t1}/\sigma_{W_j}$ is the number of Winsorized standard deviations that separate the trimmed means. However, deleting .642 would also mean that the ES would not equal δ_{R_j} when the data are normal. Moreover, when the 20% trimmed distribution is similar in shape to a 20% trimmed normal distribution (just what trimming is intended to accomplish), using .642 will put the ES on a scale that is similar to the scale of Cohen's (1965) ES. Not using .642 means that the ES will not be on a familiar scale in any situation. In addition, as was true in regard to the interpretation of δ, the meaning of δ_{R_j} will emerge from repeated use of the ES.

that the first approach (i.e., using robust estimators with the noncentral t limits) can result in inaccurate confidence coefficients when data are nonnormal. They also found that a robust CI can be obtained by using a percentile bootstrap method for empirically determining the limits of the interval.

For the current example $\hat{\delta}_{R_1} = 1.87$. This ES is much larger than $\hat{\delta}_1 = 0.77$ because the Winsorized standard deviation is much smaller than the least squares standard deviation. From the point of view of robust estimation, the least squares standard deviation is too large because of the influence of outliers on the estimate, and defining the ES by using a robust measure of spread is preferable. CIs for δ_{R_1}, constructed by using the noncentral t distribution and the percentile bootstrap are [0.31, 3.37]; and [0.26, 3.87]; respectively. As with the CIs for δ_1, the bootstrap results in a wider interval. This result is again consistent with results in Algina et al. (2006) indicating that when the second group was more variable than the first and the data were sampled from nonnormal distributions, coverage probability was too small for the CI based on the noncentral t distribution. The ES $\hat{\delta}_{R_2}$ is 0.61 and the CIs are [0.10, 1.11]; and [0.07, 1.35]. We can define a robust version of δ^*:

$$\delta^*_R = \frac{\mu_{t2} - \mu_{t1}}{\sqrt{\dfrac{\sigma^2_{W_1} + \sigma^2_{W_2}}{2}}}, \tag{21}$$

which is estimated by

$$\hat{\delta}^*_R = \frac{\hat{\mu}_{t2} - \hat{\mu}_{t1}}{\sqrt{\dfrac{\hat{\sigma}^2_{W_1} + \hat{\sigma}^2_{W_2}}{2}}}. \tag{22}$$

For the example data, we find $\hat{\delta}^*_R = 0.82$, and a 95% bootstrap CI for $\hat{\delta}^*_R$ is [0.09, 1.57]. Despite the differences in ESs and the CIs, the CIs are consistent with hypothesis testing using trimmed means and Winsorized variances: There is sufficient evidence to claim a treatment effect (see Footnote 2).

A General ADF Test Statistic

Methods that generally give improved power and better control over the probability of a Type I error can be formulated within the framework of the general linear model (GLM) ADF perspective.[10] Lix and Keselman (1995) showed how the various Welch (1938, 1951) statistics that appear in the literature for testing omnibus main and interaction effects, as well as focused hypotheses using contrasts in univariate and multivariate independent and correlated groups designs, can be formulated from a GLM ADF perspective, thus allowing researchers to apply one statistical procedure to any testable model effect. We adopt their approach in this article and begin by presenting, in abbreviated form, its mathematical underpinnings.

Consider the linear model:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\xi}, \tag{23}$$

where $\mathbf{Y}$ is an $N \times p$ matrix of scores on p dependent variables or p repeated measurements, N is the total sample size, $\mathbf{X}$ is an $N \times r$ design matrix consisting entirely of zeros and ones, $\boldsymbol{\beta}$ is an $r \times p$ matrix of population means, and $\boldsymbol{\xi}$ is an $N \times p$ matrix of random error components. Let $\mathbf{Y}_j$ ($j = 1, \ldots, r$) denote the $n_j \times p$ submatrix of $\mathbf{Y}$ containing the scores associated with the n_j subjects in the jth group (cell). It is typically assumed that the rows of $\mathbf{Y}_j$ are independently and normally distributed, with mean vector $\boldsymbol{\beta}_j$ and variance–covariance matrix $\boldsymbol{\Sigma}_j$, that is, $N(\boldsymbol{\beta}_j, \boldsymbol{\Sigma}_j)$, where $\boldsymbol{\beta}_j = (\boldsymbol{\mu}_{j1} \ldots \boldsymbol{\mu}_{jp})$, the jth row of $\boldsymbol{\beta}$, and $\Sigma_j \neq \boldsymbol{\Sigma}_{j'}$ $(j \neq j')$. Specific formulas for estimating $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}_j$, as well as an elaboration of $\mathbf{Y}_j$ are provided in Lix and Keselman (1995, Appendix A).

The general linear hypothesis is

$$H_0\colon \mathbf{R}\boldsymbol{\mu} = \mathbf{0}, \tag{24}$$

where $\mathbf{R} = \mathbf{C} \otimes \mathbf{U}^{\mathrm{T}}$, $\mathbf{C}$ is a $df_{\mathrm{C}} \times r$ matrix which controls contrasts on the independent groups effect(s), $\mathbf{U}$ is a $p \times df_{\mathrm{U}}$ matrix which controls contrasts on the within-subjects effect(s), $\otimes$ is the Kronecker or direct product function (Timm, 2002, describes the properties of the Kronecker product function), and T is the transpose operator.[11,12] For multivariate independent groups designs, $\mathbf{U}$ is frequently an identity matrix of dimension p (i.e., $\mathbf{I}_{\mathrm{p}}$). (An example of when the dimension of $\mathbf{U}$ is not p would be when one were interested in testing a single dependent measure, in which case the dimension of $\mathbf{U}$ is one.) The $\mathbf{R}$ contrast matrix has $(df_{\mathrm{C}})(df_{\mathrm{U}})$ rows and $(r)(p)$ columns. In Equation 24, $\boldsymbol{\mu} = \mathrm{vec}(\boldsymbol{\beta}^{\mathrm{T}}) = [\boldsymbol{\beta}_1 \ldots \boldsymbol{\beta}_r]^{\mathrm{T}}$. In other words, $\boldsymbol{\mu}$ is the column vector with $r \times p$ elements obtained by stacking the col-

[10] The following material provides the mathematical underpinnings to the methods, in their most general form, we present throughout the remainder of the article. A complete understanding of the matrix algebra approach will not prevent the reader from utilizing the methods we recommend, particularly as we clarify this material with the numerical examples provided in the accompanying online supplement.

[11] Though the null hypothesis is almost universally stated as a vector of zeros, the reader should remember that a vector of any real numbers can be hypothesized.

[12] If $\mathbf{C}_J$ is a matrix of dimension $(J - 1) \times J$ and $\mathbf{C}_K$ is a matrix of dimension $(K - 1) \times K$, then the Kronecker product is defined as

$$\mathbf{C}_J \otimes \mathbf{C}_K = \begin{bmatrix} c_{11}\mathbf{C}_K & \cdots & c_{1J}\mathbf{C}_K \\ \vdots & \ddots & \vdots \\ c_{(J-1)1}\mathbf{C}_K & \cdots & c_{(J-1)J}\mathbf{C}_K \end{bmatrix}.$$

umns of $\boldsymbol{\beta}^{\mathrm{T}}$. The $\mathbf{0}$ column vector is of order $df_{\mathrm{C}} \times df_{\mathrm{U}}$ (see Lix & Keselman, 1995, for illustrative examples).

The generalized test statistic given by Johansen (1980) is

$$T_{\mathrm{WJ}}=(\mathbf{R}\hat{\boldsymbol{\mu}})^{\mathrm{T}}(\mathbf{R}\hat{\boldsymbol{\Sigma}}\mathbf{R}^{\mathrm{T}})^{-1}(\mathbf{R}\hat{\boldsymbol{\mu}}), \qquad (25)$$

where $\hat{\boldsymbol{\mu}}$ estimates $\boldsymbol{\mu}$, and $\hat{\boldsymbol{\Sigma}} = \mathrm{diag}[\hat{\boldsymbol{\Sigma}}_1/n_1 \ldots \hat{\boldsymbol{\Sigma}}_r/n_r]$, a block matrix with diagonal elements $\hat{\boldsymbol{\Sigma}}_j/n_j$.[13] This statistic, divided by a constant, c (i.e., T_{WJ}/c), approximately follows an F distribution with $v_1 = df_{\mathrm{C}} \times df_{\mathrm{U}}$ and $v_2 = v_1(v_1 + 2)/(3A)$, where $c = v_1 + 2A - (6A)/(v_1 + 2)$. The formula for the statistic A is provided in the online supplement.

When $p = 1$, that is, for a univariate model, the elements of $\mathbf{Y}$ are typically assumed to be independently and normally distributed with mean μ_j and variance σ_j^2, that is, $N(\mu_j, \sigma_j^2)$. To test the general linear hypothesis, $\mathbf{C}$ has the same form and function as for the multivariate case, but now $\mathbf{U} = 1$, $\hat{\boldsymbol{\mu}} = [\hat{\mu}_1 \ldots \hat{\mu}_r]^{\mathrm{T}}$ and $\hat{\boldsymbol{\Sigma}} = \mathrm{diag}[\hat{\sigma}_1^2/n_1 \ldots \hat{\sigma}_r^2/n_r]$.

Using an ADF Solution With Robust Estimators and/or Bootstrapping

One-Way Independent Groups Design

For an independent groups experiment with n_j subjects ($\Sigma_j n_j = N$) in each of J groups, and using the notation of Equation 23, $\mathbf{Y} = (Y_{ij})$, where Y_{ij} is the score associated with the ith subject in the jth group ($j = 1, \ldots, J$; $i = 1, \ldots, n_j$), $\mathrm{E}(\bar{Y}_j) = \mu_j$, the jth population mean, $\boldsymbol{\beta}^{\mathrm{T}} = [\mu_1 \ldots \mu_J]$ and $\boldsymbol{\xi} = (\varepsilon_{ij})$ defines the random error term. The Y_{ij}s are assumed to be $N(\mu_j, \sigma_j^2)$ variates, with $\hat{\mu}_j$ and $\hat{\sigma}_j^2$, respectively, representing the jth sample mean and unbiased variance.

To test the general linear hypothesis of Equation 24, $\mathbf{R} = \mathbf{C} = \mathbf{C}_j$, because $\mathbf{U} = 1$. That is, $\mathbf{C}_j$ is a $(J - 1) \times J$ matrix for which the rows represent a set of linearly independent contrasts among the levels of the independent groups factor. For example, if there are four independent groups ($J = 4$) then $\mathbf{C}_4(3 \times 4)$ can be defined as

$$\mathbf{C}_4 = \begin{bmatrix} 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{bmatrix}.$$

It should be noted, however, that selecting $\mathbf{C}_4$ as any (3×4) matrix in which the elements in each row sum to zero and the rows are linearly independent will result in the same value of T_{WJ}. With respect to Equation 25, $\hat{\boldsymbol{\mu}} = [\hat{\mu}_1 \ldots \hat{\mu}_J]^{\mathrm{T}}$ and $\hat{\boldsymbol{\Sigma}} = \mathrm{diag}\ [\hat{\sigma}_1^2/n_1 \ldots \hat{\sigma}_J^2/n_J]$.

Pairwise contrasts on the group means are frequently of great interest (see Keselman, Cribbie, & Holland, 2004). Using Equation 24, $\mathbf{R} = \mathbf{C} = \mathbf{c}_{jj'}, = (c_1 \ldots c_J)$, the $1 \times J$ vector of coefficients which contrasts the jth and j'th means ($\Sigma_j c_j = 0$). In other words, we test the null hypothesis $H_{jj'}$: $\mu_j = \mu_{j'}$ ($j \neq j'$). For example, if again $J = 4$ and we want to compare the first and fourth means, $\mathbf{c_{14}} = [\ 1 \quad 0 \quad 0 \quad -1\ ]$.

Robust estimation, testing, and CIs for ES statistics. As indicated in the introduction, a great deal of evidence indicates that the traditional tests for mean equality are adversely affected by nonnormality, particularly when variances are heterogeneous and group sizes are unequal (see Lix & Keselman, 1998; Wilcox, 1995, 2005). As before, we apply robust estimates of central tendency and variability to the ADF statistic and to CIs for ES statistics.

Omnibus and specific effects testing. To test the general linear hypothesis in a one-way independent groups design we specify H_0: $\mathbf{R}\boldsymbol{\mu}_{\mathrm{t}} = \mathbf{0}$. Thus, with robust estimation, the trimmed group means replace the least squares group means, the Winsorized group variance estimators replace the least squares variances, and h_j ($h_j = n_j - 2g_j$; the number of observations remaining after trimming) replaces n_j, and accordingly one computes the robust version of T_{WJ}, defined as T_{WJt}. As noted, we refer the value of T_{WJt}/c to the sampling distribution of F (see Yuen, 1974). The program described in the accompanying document computes the appropriate value of c based on robust estimators.

Bootstrapping/omnibus and specific effects tests. As we previously indicated, the strategy behind the bootstrap is to shift the sample distributions of the scores for each group and variable by subtracting the group mean (least squares or trimmed, depending on whether a nonrobust or robust mean is used) from each score and using the shifted empirical distributions to estimate an appropriate critical value. For each j, obtain a bootstrap sample by randomly sampling with replacement n_j observations from the shifted values, yielding $Y_1^*, \ldots, Y_{n_j}^*$. Let F_{t}^* be the value of a test statistic (T_{WJ}/c or T_{WJt}/c) based on the bootstrap sample. The B values of F_{t}^*, where B represents the number of bootstrap simulations, are sorted in ascending order, that is, $F_{\mathrm{t}(1)}^* \leq \ldots \leq F_{\mathrm{t}(B)}^*$, and an estimate of an appropriate critical value is $F_{\mathrm{t}(a)}^*$, where $a = (1 - \alpha)B$, rounded to the nearest integer. One will reject H_0: $\mathbf{R}\boldsymbol{\mu} = \mathbf{0}$ or H_0: $\mathbf{R}\boldsymbol{\mu}_{\mathrm{t}} = \mathbf{0}$ when $F_{\mathrm{t}} \geq F_{\mathrm{t}(a)}^*$, where F_{t} is the value of the heteroscedastic statistic (T_{WJ}/c or T_{WJt}/c) based on the raw data. Thus, bootstrapping can be applied to two statistics (i.e., T_{WJ}/c or T_{WJt}/c).

Focused contrast tests such as pairwise contrasts are computed either by using the usual least squares estimators with T_{WJ}/c or by substituting robust estimators for least squares values, resulting in the T_{WJt}/c statistic. To control the FWER for a set of contrasts, the following approach is used. Let F_{t}^* be the value of the statistic based on the bootstrap sample. Set $t^* = \max F_{\mathrm{t}}^*$, the maximum being taken over all contrasts in the set, for example, $J(J - 1)/2$

[13] Lix and Keselman (1995) used the notation T_{WJ} to refer to this statistic, relating it to a Welch (1951)–James (1951) statistic. For consistency, we continue to use this designation.

tests for all possible pairwise tests on the J means. Repeat this process B times yielding $t_1^*, \ldots, t_B^*$. Let $t_{(1)}^* \leq \ldots \leq t_{(B)}^*$ be the t_b^* values written in ascending order, and $q = (1 - \alpha)B$, rounded to the nearest integer. After repeating the process B times, T_{WJ}/c or T_{WJt}/c is compared with $t_{(q)}^*$ (i.e., $T_{\mathrm{WJ}}/c \geq t_{(q)}^*$ or $T_{\mathrm{WJt}}/c \geq t_{(q)}^*$), where q is determined so that the FWER is approximately α.[14]

CIs for ES statistics. The approach we have taken (in the two-group example) was to arrive at a CI around a population ES that is robust to variance heterogeneity and nonnormality. This can be accomplished by using the nonparametric percentile bootstrap method described previously.[15] Further, the user can either set the confidence limits per interval or over the entire set of ES intervals with a Bonferroni adjustment (e.g., if five CIs are of interest, the confidence coefficient per interval can be set at 99% if one desires that the overall confidence coefficient is not less than 95%).

Factorial Independent Groups Design

Application of the general ADF solution for hypothesis testing in factorial independent groups designs will be discussed only from the perspective of a two-way design. However, the same concepts may be readily extended to higher order designs.

Let $\mathbf{Y} = (Y_{ijk})$, where Y_{ijk} represents the score associated with the ith subject in the (j,k)th treatment combination cell ($j = 1, \ldots, J$; $k = 1, \ldots, K$; $i = 1, \ldots, n_{jk}$; $\Sigma_j\Sigma_k n_{jk} = N$). Then $\mathrm{E}(Y_{ijk}) = \mu_{ijk}$ is the (j,k)th population mean, $\boldsymbol{\beta}^{\mathrm{T}} = (\mu_{11}\ \mu_{12} \ldots \mu_{JK})$, and $\boldsymbol{\xi} = (\varepsilon_{ijk})$ defines the random error. The Y_{ijk}s are assumed to be $N(\mu_{jk}, \sigma_{jk}^2)$ variates, with $\hat{\mu}_{jk}$ and $\hat{\sigma}_{jk}^2$ respectively representing the (j,k)th sample mean and unbiased variance estimates.

The sensitivity of the ANOVA F test to violations of its derivational assumptions for tests of main and interaction hypotheses in factorial designs has been studied in less detail than for one-way designs (Harwell, Rubenstein, Hayes, & Olds, 1992). Nevertheless, the evidence available supports the conclusion that the test may become seriously biased when equality of the σ_{jk}^2s is not a tenable assumption, particularly when the n_{jk}s are unequal, that is, for nonorthogonal designs, when hypotheses involving unweighted means are tested (Milligan, Wong, & Thompson, 1987). The deleterious effects of nonnormality are described by Wilcox (2003, 2005). Keselman, Carriere, and Lix (1995, 1996) identified that an ADF solution is largely robust in such situations.

To test the general linear hypothesis of Equation 24, $\mathbf{R} = \mathbf{C} = \mathbf{C}_{JK}$, $\mathbf{C}_J$, and $\mathbf{C}_K$, respectively for tests of the interaction, row, and column hypotheses, since $\mathbf{U} = 1$ in all cases. Here, $\mathbf{C}_{JK} = \mathbf{C}_j \otimes \mathrm{C}_k$, where $\mathbf{C}_j$ and $\mathbf{C}_k$ are matrices of order $(J - 1) \times J$ and $(K - 1) \times K$, respectively, for which the rows represent sets of linearly independent contrasts among the levels of the independent groups factors. Thus, $\mathbf{C}_{JK}$ is a contrast matrix of order $(J - 1)(K - 1) \times JK$.[16] For example if $J = 3$ and $K = 4$, $\mathbf{C}_j$ and $\mathbf{C}_k$ can be selected as

$$\mathbf{C}_3 = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{bmatrix}$$

and

$$\mathbf{C}_4 = \begin{bmatrix} 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{bmatrix}.$$

However, consistent with the definition of $\mathbf{C}_j$ in the design with one between-subjects factor, $\mathbf{C}_3$ can be any (2×3) matrix in which the elements in each row sum to zero and the rows are linearly independent. Similarly, $\mathbf{C}_4$ can be any (3×4) matrix in which the elements in each row sum to zero and the rows are linearly independent.

For the main effect tests, $\mathbf{C}_J = \mathbf{C}_j \otimes \mathbf{1}_K^{\mathrm{T}}$, and $\mathbf{C}_K = \mathbf{1}_J^{\mathrm{T}} \otimes \mathrm{C}_k$, where $\mathbf{1}_K$ and $\mathbf{1}_J$ are column vectors of ones, of order K and J, respectively, which serve to sum the means over the appropriate factor. Consequently, $\mathbf{C}_J$, a matrix of order (J –

[14] Though researchers can use the ADF statistic with trimmed means to compute a priori complex contrasts applying our bootstrapping method to achieve Type I error familywise (FWER) control (we do not claim FWER control if the tests are computed post hoc), we caution researchers that we are not aware of any study confirming that this approach provides adequate Type I error control for heteroscedastic and/or nonnormal data. Thus, until such empirical evidence is forthcoming, we recommend that our method be used only with pairwise and/or tetrad comparisons.

[15] It should be noted that the approach we present is not limited to setting a CI around a pairwise ES parameter. In its most general form, the approach we present would have Ψ and $\hat{\psi}$ in the numerator of the ES parameter and sample statistic, where Ψ and $\hat{\psi}$ define a population and sample contrast (pairwise or complex), respectively. As well, bootstrapping would involve as many groups that are involved in computing the value of $\hat{\psi}$. The hypothesis H_0: $\Psi = 0$ can be written as H_0: $\mathbf{r}\boldsymbol{\mu} = 0$, where $\mathbf{r} = \mathbf{c} \otimes \mathbf{u}^{\mathrm{T}}$, and the test of this hypothesis will not be affected if $\mathbf{r}$ is replaced by $k\mathbf{r}$ where k is a constant. For example, suppose an independent samples design has three means and we define $\mathbf{c} = [\ 1/2 \quad 1/2 \quad -1\ ]$ and $u = 1$ to test the hypothesis that the average of the first two means is equal to the third mean. Changing the contrast vector to $\mathbf{c} = [\ 1 \quad 1 \quad -2\ ]$ will not affect the hypothesis test. However, the change will affect the ES and the CI. The vector $\mathbf{c} = [1/2\ 1/2\ -1]$ is more appropriate for estimating an ES because then Ψ will be a difference between one mean and the average of two means. Thus, for the purpose of estimating ESs and forming CIs it is important to make sure that contrast weights have appropriate magnitudes and not merely an appropriate pattern. Examples are provided in the online supplement.

[16] For higher order designs, Algina and Olejnik (1984) have developed a set of general rules which can be used to form $\mathbf{C}$.

1) × JK, has (J – 1) contrast rows which sum across the levels of factor K, and $\mathbf{C}_K$, a matrix of order (K – 1) × JK, has (K – 1) contrast rows which sum across the levels of factor J.[17]

For pairwise comparisons on the row marginal means, $\mathbf{R} = \mathbf{C} = \mathbf{c}_{jj'} \otimes \mathbf{1}_K^{\mathrm{T}}$, a 1 × JK vector, where $\mathbf{c}_{jj'}$ contains the coefficients which contrast the jth and j'th row means. Similarly, when $\mathbf{R} = \mathbf{C} = \mathbf{1}_J^{\mathrm{T}} \otimes \mathbf{c}_{kk'}$, also a 1 × JK vector, where $\mathbf{c}_{kk'}$ contains the coefficients which contrast the kth and k'th column means, a pairwise contrast on the column marginal means is formed.

A significant interaction effect could be probed by using a variety of procedures, including tetrad contrasts.[18] Tetrad contrasts are used to test for the presence of an interaction in a 2 × 2 submatrix of the J × K data matrix. Tetrad contrasts are defined as $\mathbf{R} = \mathbf{c}_{jj'} \otimes \mathbf{c}_{kk'}$. For such contrasts, $\mathbf{R}$ is of order 1 × JK.

Robust estimation. With robust estimation, the trimmed cell means and Winsorized cell variances are once again substituted for their least squares counterparts into the ADF statistic. In addition, error *df* (υ_2s) are based on the effective sample sizes.

The results reported by Keselman et al. (1995, 1996) and Keselman, Kowalchuk, and Lix (1998) indicated that for moderate degrees of skewness (e.g., χ_3^2-type data) and variance heterogeneity (σ_{jk}^2 ratio of 1:1:1:9), the T_{WJ}/c test, with the usual least squares estimators for central tendency and variability, typically is robust in nonorthogonal designs. However, for more disparate assumption violations, the ADF test using trimmed means and Winsorized variances provides better Type I error control. Thus, the T_{WJt}/c statistic appears to us to be the more versatile procedure in that it controls rates of Type I error when conditions are substantially as well as moderately unfavorable.

Bootstrapping. The bootstrap can be generalized to factorial designs from the one-way methodology. That is, empirical sampling distributions can be created for each effect by using resampled data. Additional research is required to determine how much researchers have to gain by determining statistical significance through bootstrap methods. Results presented by Keselman, Kowalchuk, and Lix (1998) indicated that robust and powerful tests can be obtained by using trimmed means, Winsorized variances, and the t distribution in nonorthogonal heterogeneous designs when data are nonnormal. However, their findings were for 2 × 2 designs and did not include results on the performance of the bootstrap. Given results for one-way designs showing that use of the bootstrap can improve control of the Type I error rate and power, we would expect that similar results will emerge for factorial designs. Despite the need for additional research, if researchers choose to adopt the bootstrap for factorial designs, the methodology we presented for one-way independent groups designs is applicable to higher order factorial designs when estimating an appropriate critical value.

For each cell of the design, obtain a bootstrap sample by randomly sampling with replacement n_{jk} observations from the shifted values, yielding $Y_1^*, \ldots, Y_{n_{jk}}^*$. For omnibus effects ($J$, K, JK) let F_{t}^* be the value of the test statistic [T_{WJt}/c] based on the bootstrap sample. As previously indicated, one will reject the appropriate null hypothesis when $F_{\mathrm{t}} \geq F_{\mathrm{t}(\alpha)}^*$, where F_{t} is the value of the heteroscedastic statistic (for J, K, and/or JK) based on the original data.

Marginal mean or interaction contrasts can be obtained in a manner that is analogous to contrast testing in the one-way design. That is, let F_{t}^* be the value of the statistic based on the bootstrap sample. Set $t^* = \max F_{\mathrm{t}}^*$, the maximum being taken over all contrasts in the set, for example, $J(J - 1)/2$ tests for all possible pairwise tests on the J marginal means or $[J(J - 1)/2][K(K - 1)/2]$ tests for all possible tetrad contrasts on the cell means. Repeat this process B times yielding $t_1^*, \ldots, t_B^*$. Let $t_{(1)}^* \leq \ldots \leq t_{(B)}^*$ be the t_b^* values written in ascending order, and $q = (1 - \alpha)B$, rounded to the nearest integer. After repeating the process B times, T_{WJt}/c is compared with $t_{(q)}^*$.

CIs for ES statistics. Because the methods in this article do not assume homogeneity of variances or that data are normally distributed for the phenomena that psychologists investigate, our modus operandi has been to present methods which will be robust to the combined effects of variance heterogeneity and nonnormality. Thus, to control for variance heterogeneity one would not want to adopt a pooled estimator of the standard deviation (e.g., $\sqrt{\mathrm{MSW}}$), and to deal with the effects of nonnormality one would not want to rely on least squares estimators.

Accordingly, the question arises as to what standardizer should be used to set a CI around a contrast effect in order to circumvent the biasing effects of variance heterogeneity. In some contexts, this question appears to have a seemingly straightforward solution. That is, consider a 2 × 2 design where each factor involves a comparison of a control group with a group receiving a treatment. Thus, in this four-cell design, the cell which combines the control subjects from each factor seems to be a reasonable choice for the standardizer. Naturally, if one subscribes to this concept, the choice of standardizer is apparent in designs containing control groups, no matter the number of factor levels.

If factors do not contain a control level, one may simply

[17] In nonorthogonal designs, the researcher may test main effect hypotheses involving either weighted or unweighted means, depending on the values assigned to the elements of **C** (see Keselman et al., 1995, 1996). For the sake of simplicity, we have assumed that the researcher is interested in testing hypotheses of unweighted means (see Maxwell & Delaney, 2004, pp. 320–342).

[18] Tetrad contrasts are a special case of interaction contrasts (see Hochberg & Tamhane, 1987, pp. 295–303).

select the standardizer from a particular cell. This approach is a generalization of the method recommended by Glass et al. (1981), and the one we adopted in the one-way completely randomized design. Having selected a standardizer, one can use the nonparametric percentile bootstrap to obtain a CI; the interval, of course, would be set around the robust parameter of ES, that is, the parameter and CI are based on trimmed means and Winsorized variances.

The idea of using the standard deviation from just one cell of the design applies to setting CIs around tetrad contrasts (i.e., interaction contrasts) as well. That is, we do not recommend using a pooled standard deviation or other pooled values from subset cells of the design, approaches that are cited in the literature, because, in the presence of variance heterogeneity, these intervals for ES statistics will not be robust (see, e.g., Kline, 2004, pp. 227–228).[19] Therefore, we recommend that users adopt our generally robust ES statistic and the percentile bootstrap method, remembering that the numerator of the statistic is $\hat{\psi}$, which can represent a pairwise and/or tetrad contrast. As was pointed out in the one-way design, one can also use the unweighted average of the cell variances as the standardizer. Again, the confidence coefficient can be set per interval or over the set of intervals (with a Bonferroni adjustment).

Correlated Groups Design

Omnibus and specific effects testing. Keselman, Carriere, and Lix (1993) have shown how T_{WJ}/c can be used to test for treatment effects in between- by within-subjects correlated groups designs (see also Keselman, 1998). Furthermore, they have demonstrated, through Monte Carlo methods, that this statistic is generally robust to nonnormality and covariance heterogeneity in nonspherical unbalanced repeated measures designs.[20]

Even though it has been demonstrated that the ADF procedure is generally robust to the combined effects of nonnormality and covariance heterogeneity, under some conditions of departure from multisample sphericity and multivariate normality, its rate of Type I error has been found to be inflated (see Algina & Keselman, 1997; Keselman et al., 1993). Further improvement in Type I error is possible by applying the procedure with robust estimators, that is, with trimmed means and Winsorized variances and covariances and/or by obtaining critical values through bootstrap methods (see Keselman, Algina, et al., 2000; Keselman, Kowalchuk, Algina, Lix, & Wilcox, 2000).

Furthermore, the sample sizes necessary to achieve robustness with these estimators and/or bootstrapping can be substantially smaller than the sizes required to achieve robustness with the ADF procedure based on least squares estimators. Thus, though we subscribe to the analysis procedures advocated by Keselman (1998) for the analysis of repeated measures effects, procedures based on the usual least squares estimators, his analyses should, when appropriate, be adopted with robust estimators.[21]

Consider the design in which n_j subjects ($\Sigma_j n_j = N$) in each of J groups are measured on a single dependent variable at K points in time, or under each of K treatments. Using the notation of Equation 23, the observations $\mathbf{Y} = (\mathrm{Y}_{ij})$, where $\mathbf{Y}_{ij} = [Y_{ij1} \ldots Y_{ijK}]$ ($j = 1, \ldots, J$; $i = 1, \ldots n_j$) and $\boldsymbol{\beta}^{\mathrm{T}} = (\boldsymbol{\mu}_j) = (\mu_{j1} \ldots \mu_{jK})$ and $\boldsymbol{\xi} = (\boldsymbol{\varepsilon}_{ij}) = (\varepsilon_{ij1} \ldots \varepsilon_{ijK})$. The $\mathbf{Y}_{ij}$s are assumed to be $N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ $\mathbf{K}$-vector variates, with $\hat{\boldsymbol{\mu}}_j$ and $\hat{\boldsymbol{\Sigma}}_j$ denoting the jth sample mean vector and variance–covariance matrix, respectively.

To test the general linear hypothesis of Equation 24, both $\mathbf{C}$ and $\mathbf{U}$ are defined in terms of the effect to be tested, to create the appropriate $\mathbf{R}$ contrast matrix. For the within-subjects interaction effect, $\mathbf{R} = \mathbf{C}_j \otimes \mathbf{U}_K^{\mathrm{T}}$, because $\mathbf{C} = \mathbf{C}_j$, where $\mathbf{C}_j$ has the same form and function as for the one-way univariate independent groups design, and $\mathbf{U} = \mathbf{U}_k$, where $\mathbf{U}_k$ is a $K \times (K - 1)$ matrix whose columns form a set of linearly independent contrasts among the levels of the within-subjects factor. Thus, $\mathbf{R}$ is of order $(J - 1)(K - 1) \times JK$. For example, if the between-subjects factor has $J = 3$ levels and the within-subjects factor has $K = 4$ levels,

$$\mathbf{C}_3 = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{bmatrix}$$

and

$$\mathbf{U}_4 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ -1 & -1 & -1 \end{bmatrix}.$$

As indicated in the presentation of between-subjects designs, other forms for $\mathbf{C}_j$ and $\mathbf{U}_k$ will result in the same test statistic.

For tests of the within-subjects main effect, $\mathbf{C} = \mathbf{1}_J^{\mathrm{T}}$ and

[19] Because our accompanying SAS program allows the user to select the standard deviation from any cell of a factorial design, users can when computing ESs for contrasts (i.e., pairwise, complex, tetrad) select different cell standard deviations for each contrast; however, this is not what we recommend.

[20] Neither single factor nor multifactor within-subjects designs are considered since covariance heterogeneity is not an issue when the design does not contain an independent groups variable. However, users should always use a procedure and critical value which can either contend with violations of the sphericity assumption, such as an adjusted-*df* test (see Keselman & Rogan, 1980), or bypass it altogether, such as a multivariate test. As well, users should be cognizant of the normality assumption.

[21] Wilcox, Keselman, Muska, and Cribbie (2000) found that applying robust estimators with a multivariate statistic did not result in good Type I error control under conditions of nonnormality.

$\mathbf{U} = \mathbf{U}_k$; for the independent groups main effect, $\mathbf{C} = \mathbf{C}_j$ and $\mathbf{U} = \mathbf{1}_K$. Consequently, these $\mathbf{R}$ matrices are of order $(K - 1) \times JK$ and $(J - 1) \times JK$, respectively.

As we have indicated, significant interaction effects can be probed with a variety of procedures, including tetrad contrasts. FWER control can be obtained with a procedure described in Lix and Keselman (1996). Main effects may be probed with pairwise comparisons of the marginal means (see Keselman & Lix, 1995). To test within-subjects pairwise comparison hypotheses with an ADF approach, $R = \mathbf{1}_J^{\mathrm{T}} \otimes \mathbf{u}_{kk'}^{\mathrm{T}}$, of order $1 \times JK$, where $\mathbf{C} = \mathbf{1}_J^{\mathrm{T}}$ and $\mathbf{U} = \mathbf{u}_{kk'}$.

Robust estimation/omnibus and specific effects testing. Keselman, Kowalchuk, Algina, et al. (2000) indicated how one Winsorizes the observations in order to compute the Winsorized covariance matrices and, as well, indicate how to compute trimmed means in a $J \times K$ design. Robust estimators can then be applied to T_{WJ}/c.

Keselman, Algina, et al. (2000) and Keselman, Kowalchuk, Algina, et al. (2000) found that in the context of $J \times K$ repeated measures designs, bootstrapping did not result in better control of Type I errors than did test statistics that just adopted trimmed means and Winsorized variances and covariances. However, their findings are applicable to a limited number of designs, and therefore may not generalize to other repeated measures designs. Accordingly, we present bootstrap methodology for those who believe it could be beneficial for the designs they utilize.

For a fixed value of j, randomly sample with replacement n_j rows of observations from the matrix

$$\begin{bmatrix} Y_{1j1} & ,\ldots, & Y_{1jK} \\ & \vdots & \\ Y_{n_jj1} & ,\ldots, & Y_{n_jjK} \end{bmatrix}.$$

Label the results

$$\begin{bmatrix} Y^*_{1j1} & ,\ldots, & Y^*_{1jK} \\ & \vdots & \\ Y^*_{n_jj1} & ,\ldots, & Y^*_{n_jjK} \end{bmatrix}.$$

Shift the bootstrap samples.[22] Next compute F^*_{t}, the value of the statistic which is based on the shifted bootstrapped values. Repeat this process B times yielding $F^*_{b\mathrm{t}}$, $b = 1, \ldots, B$. Once again, effects are significant if $T_{\mathrm{WJt}}/c \geq F^*_{(a)}$. We again recommend that B be set at 1,000. Focused hypothesis tests using contrasts are accomplished in the same manner previously enumerated.

Robust CIs for ES statistics. In the context of a single repeated measures factor with two levels ($i = 1, \ldots, n$; $k = 1, 2$), Algina and Keselman (2003) developed and evaluated a noncentral t-based approximate CI for an ES that was suggested by Bird (2002):

$$\delta^\dagger = \frac{\mu_2 - \mu_1}{\sqrt{(\sigma_1^2 + \sigma_2^2)/2}}. \tag{26}$$

Subsequently Algina et al. (2005b) showed that the approximate CI had poor coverage probability when the data were drawn from a nonnormal distribution. As an alternative to the approximate CI for the ES in Equation 26, they developed an approximate CI for the robust ES based on the noncentral t distribution.

$$\delta_R^\dagger = .642\,\frac{\mu_{\mathrm{t}2} - \mu_{\mathrm{t}1}}{\sqrt{(\sigma^2_{\mathrm{W}_1} + \sigma^2_{\mathrm{W}_2})/2}}, \tag{27}$$

where $\mu_{\mathrm{t}k}$ and $\sigma^2_{\mathrm{W}_k}$ are the population trimmed mean and Winsorized variance, respectively, for the kth level. The robust ES can be estimated by

$$\hat{\delta}_R^\dagger = .642\,\frac{\hat{\mu}_{\mathrm{t}2} - \hat{\mu}_{\mathrm{t}1}}{\sqrt{(\hat{\sigma}^2_{\mathrm{W}_1} + \hat{\sigma}^2_{\mathrm{W}_2})/2}}. \tag{28}$$

Algina et al. (2005b) presented evidence that coverage probability was not always adequate when data were drawn from a nonnormal distribution. Alternatively, the nonparametric percentile bootstrap can be used to calculate a CI for $\delta_R^\dagger$ and was shown by Algina et al. (2005b) to provide better coverage probability under the nonnormal distributions included in the study, whereas a percentile bootstrap CI for $\delta^\dagger$ had poor coverage probability for many of these distributions.[23]

Recall that when a researcher suspects that the population variances are not equal, Glass et al. (1981) proposed, in a completely randomized design, using the ES

$$\hat{\delta}_k = \frac{\hat{\mu}_2 - \hat{\mu}_1}{\hat{\sigma}_k}. \tag{29}$$

Similarly, in a within-subjects design in which a researcher suspects the variances for the repeated measures are not equal, $\hat{\delta}_k$ can also be used. The ES $\hat{\delta}_k$ estimates

$$\delta_k = \frac{\mu_2 - \mu_1}{\sigma_k}, \tag{30}$$

which seems more appropriate than does $\delta^\dagger$ when variances are not equal. A robust alternative to δ_k is

[22] Following Westfall and Young (1993), and as enumerated by Wilcox (1997), the shifted values are the empirical distribution centered so that the sample trimmed mean is zero. That is, the empirical distributions are shifted so that the null hypothesis of equal trimmed means is true in the sample. The strategy behind the bootstrap is to use the shifted empirical distributions to estimate an appropriate critical value.

[23] Algina et al. (2005b) examined both symmetric and asymmetric distributions. The nonnormal distributions had skewness values (γ_1) and kurtosis values (γ_2) of ($\gamma_1 = 2$, $\gamma_2 = 6$), ($\gamma_1 = 0$, $\gamma_2 = 154.84$), and ($\gamma_1 = 4.90$, $\gamma_2 = 4{,}673.80$), where for a normal distribution $\gamma_1 = \gamma_2 = 0$.

$$\delta_{R_k} = .642 \frac{\mu_{t2} - \mu_{t1}}{\sigma_{W_k}}. \quad (31)$$

The corresponding sample estimator is

$$\hat{\delta}_{R_k} = .642 \frac{\hat{\mu}_{t2} - \hat{\mu}_{t1}}{\hat{\sigma}_{W_k}}. \quad (32)$$

Algina et al. (2005b) developed noncentral *t*-based CIs for δ_k and δ_{R_k} and evaluated coverage probability for these CIs as well as for percentile bootstrap CIs for δ_k and δ_{R_k}. Results indicated that when data were drawn from nonnormal distributions coverage probability for the noncentral *t*-based CIs for δ_k and δ_{R_k} could be inadequate, though coverage probability was better for the CI for δ_{R_k}. Under the nonnormal distributions included in the study, coverage probability of the percentile bootstrap CI for δ_k was also poor in some conditions. Coverage probability of the percentile bootstrap CI for δ_{R_k} was much better, with only 1 of 800 estimated values outside the interval [0.925, 0.975] and the vast majority inside the interval [0.94, 0.96].

As was the case in completely randomized factorial designs, defining the standardizer is a conceptual issue, not a statistical issue, in factorial repeated measures designs. The choice will sometimes be straightforward and at other times will be more arbitrary. For example, for factorial repeated measures designs that contain a pretest condition, a most reasonable choice is to select the standard deviation from the pretest cell (see Kline, 2004, pp. 104–114).[24] However, for repeated measures designs not containing a pretest condition, the choice can be from among any cell of the between- by within-subjects repeated measures design (see also Footnote 6).

Thus, once again our recommended approach for setting a robust CI around a robust parameter of ES is to adopt the statistic presented in Equation 32, where, in its most general form, the numerator is $\hat{\psi}$, the sample value of a pairwise and/or interaction contrast. The percentile bootstrap is then used to find the upper and lower confidence limits.

Conclusions

When psychological researchers encounter populations that are nonnormal in form and subscribe to the position that inferences pertaining to robust parameters are more valid than inferences pertaining to the usual least squares parameters, then procedures based on robust estimators should be adopted. In our article, we presented an approximate *df* test statistic for one-way and factorial completely randomized and correlated groups designs based on robust estimators (trimmed means and Winsorized variances and covariances) in order to circumvent the biasing effects of variance heterogeneity and nonnormality. As well, we indicated when testing could be improved by determining statistical significance through a bootstrap method. In addition, we indicated how researchers can set robust CIs around robust ES parameters.

We presented this methodology in order to encourage researchers to adopt a procedure that has been shown to be generally robust to variance heterogeneity and nonnormality. That is, the empirical literature has indicated that distortion in rates of Type I error can generally be eliminated by applying robust estimators with heteroscedastic test statistics. Moreover, the power to detect treatment effects is also improved through the use of robust estimators in the presence of nonnormal data. As well, the accuracy of confidence coefficients based on robust estimators for robust parameters was noted.

Within the context of independent groups designs, we indicated that for one-way designs, the use of a bootstrap methodology does indeed result in better Type I error control. For factorial designs, the current literature suggests that the adoption of robust estimators may be sufficient to eliminate the biasing effects of variance heterogeneity and nonnormality, though researchers can apply the bootstrap to assess statistical significance.

With respect to the analysis of effects in correlated groups designs, we strongly support the recommendations presented by Keselman (1998). Keselman (1998) recommended the use of the ADF statistic in repeated measures designs based on least squares estimators. However, as was pointed out, sample sizes must meet the prescriptions enumerated by Keselman et al. (1993) and Algina and Keselman (1997), in order to obtain a robust test with the ADF solution. When sample sizes do not meet these prescriptions, researchers can still obtain a robust test of treatment effects by applying trimmed means and Winsorized variances and covariances with the ADF statistic. Indeed, the results reported by Keselman, Algina, et al. (2000) indicated that robustness can be achieved with very modest sample sizes (e.g., $n_j = 22$; see also Luh & Guo, 2007).

The ADF solution can be applied to a wide range of designs by using a GLM framework to define the hypothesis of interest and the computer program demonstrated in our online supplement. Lix and Keselman (1995) showed how to specify multivariate designs and the associated tests of model effects. The ADF solution has been explored in a limited manner in multivariate repeated measures designs (Keselman & Lix, 1997), but not with robust estimators and/or bootstrapping. The software that we describe in our online resources can also be used to set robust CIs around robust ES estimates.

[24] Some might take the view that a pretest should be treated as a covariate, not a level of the repeated measures variable, and, therefore, the standardizer should be obtained from a "real" level of the repeated measures variable.

As a final note, we want to once again generally recommend that applied researchers adopt robust estimation and testing methods. In particular, we recommend that trimmed means be used in the ADF statistic and in CIs for ES statistics. In addition, generally, critical values for test statistics and CIs should be obtained through the bootstrap methodology described in this article. Moreover, with respect to defining and estimating ES, we strongly believe in the view originally voiced by Glass et al. (1981), namely, "the average standard deviation should probably be eliminated as a mindless statistical reaction to a perplexing choice" (p. 106).

However, as we previously acknowledged, there will always be instances (i.e., data sets) where other methods of analysis would provide better Type I error control, better power to detect effects, and better CIs of ESs. Moreover, and most importantly, we could not a priori enumerate all the scenarios (shapes of distributions, magnitudes of variance heterogeneity, extent of sample size imbalance, etc.) where one method would dominate others. Thus, what we attempted to do in this article was provide researchers with a general methodology that has four options to consider: (a) least squares estimators with a robust test statistic, (b) robust estimators with a robust test statistic, and (c) using bootstrapping methodology with options (a) or (b). This substantially simplifies selecting a method of analysis for the applied researcher. With regard to setting CIs for an ES statistic, we recommend adopting the method enumerated in this article: a standardized mean difference statistic based on robust estimators, where the limits of the CI are based on a bootstrap method. Lastly, because we acknowledge that it is not a simple matter to discover how groups may differ from one another, researchers should not only consider the methods of analysis we enumerated in our article but other analysis strategies as well, for example, rank transformation methods (see Conover & Iman, 1981) and nonparametric methods (see Brunner, Dette, & Munk, 1997; Brunner & Munzel, 2000; Zimmerman & Zumbo, 1993, etc.).

References

Algina, J., & Keselman, H. J. (1997). Testing repeated measures hypotheses when covariance matrices are heterogeneous: Revisiting the robustness of the Welch–James test. *Multivariate Behavioral Research, 32,* 255–274.

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). An alternative to Cohen's standardized mean difference effect size: A robust parameter and confidence interval in the two independent group case. *Psychological Methods, 10,* 317–328.

Algina, J., Keselman, H. J., & Penfield, R. D. (2005b). 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. (2006). Confidence intervals for an effect size when variances are not equal. *Journal of Modern Applied Statistical Methods, 5,* 2–13.

Algina, J., & Olejnik, S. F. (1984). Implementing the Welch–James procedure with factorial designs. *Educational and Psychological Measurement, 44,* 39–48.

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

Aspin, A. A. (1947). An examination and further development of a formula arising in the problem of comparing two mean values. *Biometrika, 35,* 88–96.

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

Bradley, J. V. (1980). Nonrobustness in *Z, t,* and *F* tests at large sample sizes. *Bulletin of the Psychonomic Society, 16,* 333–336.

Brown, M. B., & Forsythe, A. B. (1974). The small sample behavior of some statistics which test the equality of several means. *Technometrics, 16,* 129–132.

Brunner, E., Dette, H., & Munk, A. (1997). Box type approximations in nonparametric factorial designs. *Journal of the American Statistical Association, 92,* 1494–1502.

Brunner, E., & Munzel, U. (2000). The nonparametric Behrens–Fisher problem: Asymptotic theory and a small sample approximation. *Biometrical Journal, 42,* 17–25.

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

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

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

Conover, W. J., & Iman, R. L. (1981). Rank transformation as a bridge between parametric and nonparametric statistics. *The American Statistician, 35,* 124–129.

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.

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

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

Harwell, M. R., Rubinstein, E. N., Hayes, W. S., & Olds, C. C. (1992). Summarizing Monte Carlo results in methodological research: The one- and two-factor fixed effects ANOVA cases. *Journal of Educational Statistics, 17,* 315–339.

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

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

Hochberg, Y., & Tamhane, A. C. (1987). *Multiple comparison procedures.* New York: Wiley.

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

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

James, G. S. (1951). The comparison of several groups of observations when the ratios of the population variances are unknown. *Biometrika, 38,* 324–329.

Johansen, S. (1980). The Welch–James approximation to the distribution of the residual sum of squares in a weighted linear regression. *Biometrika, 67,* 85–92.

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

Keselman, H. J. (1998). Testing treatment effects in repeated measures designs: An update for psychophysiological researchers. *Psychophysiology, 35,* 470–478.

Keselman, H. J., Algina, J., Wilcox, R. R., & Kowalchuk, R. K. (2000). Testing repeated measures hypotheses when covariance matrices are heterogeneous: Revisiting the robustness of the Welch–James test again. *Educational and Psychological Measurement, 60,* 925–938.

Keselman, H. J., Carriere, K. C., & Lix, L. M. (1993). Testing repeated measures hypotheses when covariance matrices are heterogeneous. *Journal of Educational Statistics, 18,* 305–319.

Keselman, H. J., Carriere, K. C., & Lix, L. M. (1995). Robust and powerful nonorthogonal analyses. *Psychometrika, 60,* 395–418.

Keselman, H. J., Carriere, K. C., & Lix, L. M. (1996). Errata to "Robust and powerful nonorthogonal analyses." *Psychometrika, 61,* 191.

Keselman, H. J., Cribbie, R. A., & Holland, B. (2004). Pairwise multiple comparison test procedures: An update for clinical child and adolescent psychologists. *Journal of Clinical Child and Adolescent Psychology, 33,* 623–645.

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R., Donahue, B., et al. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Review of Educational Research, 68,* 350–386.

Keselman, H. J., Kowalchuk, R. K., Algina, J., Lix, L. M., & Wilcox, R. R. (2000). Testing treatment effects in repeated measures designs: Trimmed means and bootstrapping. *British Journal of Mathematical and Statistical Psychology, 53,* 175–191.

Keselman, H. J., Kowalchuk, R. K., & Lix, L. M. (1998). Robust nonorthogonal analyses revisited: An update based on trimmed means. *Psychometrika, 63,* 145–163.

Keselman, H. J., & Lix, L. M. (1995). Improved repeated measures stepwise multiple comparison procedures. *Journal of Educational and Behavioral Statistics, 20,* 83–99.

Keselman, H. J., & Lix, L. M. (1997). Analyzing multivariate repeated measures designs when covariance matrices are heterogeneous. *British Journal of Mathematical and Statistical Psychology, 50,* 319–338.

Keselman, H. J., & Rogan, J. C. (1980). Repeated measures *F* tests and psychophysiological research: Controlling the number of false positives. *Psychophysiology, 17,* 499–503.

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research.* Washington, DC: American Psychological Association.

Kohr, R. L., & Games, P. A. (1974). Robustness of the analysis of variance, the Welch procedure and a Box procedure to heterogeneous variances. *Journal of Experimental Education, 43,* 61–69.

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

Kulinska, E., & Staudte, R. G. (2006). Interval estimates of weighted effect sizes in the one-way heteroscedastic ANOVA. *British Journal of Mathematical and Statistical Psychology, 59,* 97–111.

Lix, L. M., & Keselman, H. J. (1995). Approximate degrees of freedom tests: A unified perspective on testing for mean equality. *Psychological Bulletin, 117,* 547–560.

Lix, L. M., & Keselman, H. J. (1996). Interaction contrasts in repeated measures designs. *British Journal of Mathematical and Statistical Psychology, 49,* 147–162.

Lix, L. M., & Keselman, H. J. (1998). To trim or not to trim: Tests of mean equality under heteroscedasticity and nonnormality. *Educational and Psychological Measurement, 58,* pp. 409–429, 853.

Luh, W., & Guo, J. (2007). Approximate sample size formulas for the two-sample trimmed mean test with unequal variances. *British Journal of Mathematical and Statistical Psychology, 60,* 137–146.

Marazzi, A., & Ruffieux, C. (1999). The truncated mean of an asymmetric distribution. *Computational Statistics & Data Analysis, 32,* 79–100.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective.* (2nd ed.) Mahwah, NJ: Erlbaum.

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

Milligan, G. W., Wong, D. S., & Thompson, P. A. (1987). Robustness properties of nonorthogonal analysis of variance. *Psychological Bulletin, 101,* 464–470.

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

Ramsey, P. H. (1980). Exact Type 1 error rates for robustness of Student's *t* test with unequal variances. *Journal of Educational Statistics, 5,* 337–349.

SAS Institute. (1999). *SAS/IML: User's guide* (Version 8). Cary, NC: Author.

Smithson, M. (2003). *Confidence intervals.* Thousand Oaks, CA: Sage.

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

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

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

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

Timm, N. H. (2002). *Applied multivariate analysis.* New York: Springer.

Tukey, J. W. (1960). A survey of sampling from contaminated normal distributions. In I. Olkin et al. (Eds.), *Contributions to probability and statistics.* (pp. 448–485) Stanford, CA: Stanford University Press.

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

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika, 29,* 350–362.

Welch, B. L. (1947). The generalization of Students' problem when several different population variances are involved. *Biometrika, 34,* 23–35.

Welch, B. L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika, 38,* 330–336.

Westfall, P. H., & Young, S. S. (1993). *Resampling-based multiple testing.* New York: Wiley.

Wilcox, R. R. (1995). ANOVA: A paradigm for low power and misleading measures of effect size? *Review of Educational Research, 65*(1), 51–77.

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

Wilcox, R. R. (2003). *Applying contemporary statistical techniques.* New York: Academic Press.

Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing* (2nd ed.). New York: Elsevier.

Wilcox, R. R., Charlin, V. L., & Thompson, K. (1986). New Monte Carlo results on the robustness of the ANOVA *F,* W, and F^* statistics. *Communications in Statistics. Simulation and Computation, 15,* 933–944.

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

Wilcox, R. R., Keselman, H. J., & Kowalchuk, R. K. (1998). Can tests for treatment group equality be improved?: The bootstrap and trimmed means conjecture. *British Journal of Mathematical and Statistical Psychology, 51,* 123–134.

Wilcox, R. R., Keselman, H. J., Muska, J., & Cribbie, R. (2000). Repeated measures ANOVA: Some new results on comparing trimmed means and means. *British Journal of Mathematical and Statistical Psychology, 53,* 69–82.

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

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

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

Zimmerman, D. W., & Zumbo, B. D. (1993). Relative power of parametric and nonparametric statistical methods. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Vol. 1. Methodological issues* (pp. 481–517). Hillsdale, NJ: Erlbaum.

Received November 14, 2006
Revision received March 5, 2008
Accepted March 11, 2008 ■