Effect of Nonnormality on Test Statistics for One-Way Independent Groups Designs

Robert A. Cribbie[1], Lisa Fiksenbaum[1], H. J. Keselman[2] & Rand R. Wilcox[3]

[1] Department of Psychology, York University

[2] Department of Psychology, University of Manitoba

[3] Department of Psychology, University of Southern California

Correspondence concerning this manuscript should be sent to: Robert A. Cribbie, Department of Psychology, York University, Toronto, ON, M3J 1P3; cribbie@yorku.ca.

Abstract

The data obtained from one-way independent groups designs is typically nonnormal in form and rarely is equally variable across treatment populations (i.e., population variances are heterogeneous). Consequently, the classical test statistic that is used to assess statistical significance [i.e., the analysis of variance (ANOVA) *F*-test] typically provides invalid results (e.g., too many Type I errors, reduced power). For this reason, there has been considerable interest in finding a test statistic that is appropriate under conditions of nonnormality and variance heterogeneity. Previously recommended procedures for analyzing such data include the James (1951) test, the Welch (1951) test applied either to the usual least squares estimators of central tendency and variability, or the Welch test with robust estimators, i.e., trimmed means and Winsorized variances. A new statistic proposed by Krishnamoorthy, Lu and Mathew (2007), intended to deal with heterogeneous variances, though not nonnormality, uses a parametric bootstrap procedure. In their investigation of the parametric bootstrap test, the authors examined its operating characteristics under limited conditions and did not compare it to the Welch test based on robust estimators. Thus, we investigated how the parametric bootstrap procedure, and a modified parametric bootstrap procedure based on trimmed means, perform relative to previously recommended procedures when data are nonnormal and heterogeneous. The results indicated that the tests based on trimmed means offer the best Type I error control and power when variances are unequal and at least some of the distribution shapes are nonnormal.

Effects of Nonnormality on Test Statistics for One-Way Independent Groups Designs

A common question in the behavioural sciences is whether treatment groups differ on an outcome variable. For example, a researcher may be interested in determining if eating disorder symptomatology (e.g., obsession with weight) vary across different cultural backgrounds. The procedure that is most popular for analyzing data from one-way independent groups designs is the analysis of variance (ANOVA) *F*-test. The ANOVA can be a valid and powerful test for identifying treatment effects; but, when the validity assumptions underlying the test are violated, the results from the test are typically unreliable and invalid.

One mathematical validity assumption of the test (i.e., a condition that was stipulated in order to derive the test statistic) is that the distribution of each population is normal in form. Although this is assumed by most researchers, it is very often not the case (Micceri, 1989). Nonnormality can have deleterious effects on the *F*-test, where predominantly there is a lack of sensitivity to detect treatment effects (Wilcox, 1997). As well, there is an increased risk that null effects will be falsely declared statistically significant (i.e., an elevated probability of committing a Type I error), especially when sample sizes are small.

A second mathematical restriction that was adopted when deriving the test statistic was that the population variances be equal. It is well known that unequal variances are the norm, rather than the exception, with behavioral science data (Erceg-Hurn & Mirosevich, 2008; Golinski & Cribbie, 2009; Grissom, 2000; Keselman et al., 1998), with largest to smallest group ratios greater than ten not uncommon (Grissom, 2000; Wilcox, 1987). Moreover, unequal variances can have drastic effects on the reliability and validity of the *F*-test, especially when

group sample sizes are also unequal (Glass, Peckham & Sanders, 1972; Harwell, Rubenstein, Hayes & Olds, 1992; Kohr & Games, 1974; Scheffé, 1959). When distributions are nonnormal and variances are unequal, the empirical probability of a Type I or Type II error for the *F*-test can deviate even more substantially from the nominal levels than when either assumption is independently violated (Glass, Peckham & Sanders, 1972; Luh & Guo, 2001).

Several procedures have been recommended for analyzing the data from one-way independent groups designs when distributions are nonnormal and variances are unequal (e.g., Brunner, Dette, & Munk, 1997; Cribbie, Wilcox, Bewell & Keselman, 2007; Wilcox & Keselman, 2003). Currently, the most recommended approaches involve utilizing the James (1951) or Welch (1951) heteroscedastic *F*-tests (based on the usual least squares estimators), or the Welch heteroscedastic *F*-test with trimmed means and Winsorized variances. Several studies have demonstrated that the original James and Welch procedures are generally robust (with respect to Type I errors and power) when group variances and sample sizes are extremely unequal (e.g., Kohr & Games, 1974; Krisnamoorty, Lu & Mathew, 2007), and further that the test is robust to unequal variances and nonnormal data, as long as the nonnormality is mild to moderate (Algina, Oshima, & Lin, 1994). The Welch test with trimmed means and Winsorized variances has also been shown to provide excellent Type I error control and power even under extreme violations of the normality and variance equality assumptions (Keselman, Wilcox, Othman & Fradette, 2002).

An important condition of nonnormality that has received very little attention in the methodological literature is the case of dissimilar distribution shapes across treatment groups. For example, it is not uncommon for behavioral science researchers to encounter one group with

an approximately normal distribution and another group with a skewed distribution. For example, Leentjens, Wielaert, van Harskamp and Wilmink (1998) found that scores on many measures of nonverbal aspects of language (i.e., prosody) were normally distributed in control groups, but were extremely skewed in schizophrenic patients. Wilcox (2005) notes that skewed distributions in general are not as problematic as when groups have different amounts of skewness. Indeed, Tiku (1964) explored situations where skew differed between groups and found that Type I and Type II errors were adversely affected when groups are skewed in opposite directions, especially with smaller sample sizes. It is important to point out that when distribution shapes are dissimilar, isolating the specific nature of the differences in the distributions is an important part of the data analysis (and comparisons of central tendencies may be less informative). For example, when distribution shapes are dissimilar, alternative descriptive statistics, such as the specific quantiles (e.g., 10th, 25th, 75th, 90th) for each distribution, can be useful in understanding differences between the distributions. Further, if one suspects that distribution shapes might be dissimilar, it might be fruitful to explicitly test for differences in the distributions using a runs test, such as the Wald-Wolfowitz, or a test of a common distribution, such as the Kolmogorov-Smirnov or Cramer-von Mises tests (see Sprent & Smeeton, 2001, pages 185-188). For example, in the Leentjens et al. (1998) study described above, the goal of the researchers was to compare the central tendencies of the groups, although specific tests used to isolate differences in the shapes of the distributions may have also been informative. Thus, when distribution shapes differ, researchers may be interested in exploring differences in the central tendencies, exploring the nature of the distributional differences, or both. Since the underlying goal of most studies in psychology that involve comparing groups is to compare the

central tendencies, this study addresses the important question of how available test statistics perform under these conditions.

The parametric bootstrap procedure proposed by Krishnamoorthy et al. (2007) is a relatively new statistic for comparing the means of independent groups when the variances of the groups are unequal. This test involves generating sample statistics from parametric models, where the parameters in the model are replaced by their estimates (see below for details regarding the parametric bootstrap procedure). This procedure was found by the authors to provide a better balance of Type I error control and power than the original Welch (1951) procedure, especially when sample sizes were small and the number of groups was large.

There are, however, important questions that were not explored by Krishnamoorthy et al. (2007). For example, how well will the Krishnamoorthy et al. procedure perform (with respect to controlling Type I and II error rates) when distribution shapes are nonnormal? This question is important because, as discussed earlier, distributions in the behavioural sciences are rarely normal. An important point related to this issue is how to distinguish between a normally distributed variable and nonnormally distributed variable. Although numerous test statistics have been proposed for detecting deviations from normality (e.g., Chen & Shapiro, 1995; D'Agostino, 1971; Shapiro & Wilk, 1965), it is also important to consider that: 1) the performance of tests of normality are greatly affected by sample size, the form of nonnormality, etc. (Seier, 2002); 2) graphical methods (e.g, histograms, boxplots, normal quantile plots) can sometimes be as informative as tests of normality for detecting deviations from normality (Holgersson, 2006); and most importantly, 3) the power of many traditional parametric tests can be severely affected by even slight deviations from normality (Wilcox, 2005). Therefore, even though there is

subjectivity in deciding whether or not a distribution is normal, it is important that we are aware of how various test statistics perform under different degrees of nonnormality in order to be able to make informed recommendations regarding the appropriate test statistics to use with nonnormally distributed variables.

A second important question is how each of the previously recommended procedures performs, with respect to Type I errors and power, when distribution shapes are dissimilar across treatment groups? For example, how would the available test statistics perform if one distribution is normal in shape and one distribution is positively skewed? This question has yet to be investigated in one-way independent groups designs.

Test Statistics

Welch's (1951) Heteroscedastic F Test. Welch derived a heteroscedastic alternative to the ANOVA *F*-test that would be robust to violations of the variance homogeneity assumption. The hypothesis that is tested is Ho: $\mu_1 = ... = \mu_J$ (j = 1, ..., J), and is rejected if $F_W \geq F_{\alpha, J-1, \nu_W}$, where:

$$F_w = \frac{\dfrac{\sum_j w_j\left(\overline{X}_j - \overline{X}_.'\right)^2}{J-1}}{1+\dfrac{2(J-2)}{J^2-1}\sum_j\left(\dfrac{1}{n_j-1}\right)\left(1-\dfrac{w_j}{\sum_j w_j}\right)^2},$$

$$v_w = \frac{J^2-1}{3\sum_j\left(\dfrac{1}{n_j-1}\right)\left(1-\dfrac{w_j}{\sum_j w_j}\right)^2},$$

$$w_j = \frac{n_j}{s_j^2},$$

$$\overline{X}_.' = \frac{\sum_j w_j \overline{X}_j}{\sum_j w_j},$$

n_j represents the number of subjects in group j, s_j represents the standard deviation of group j, and $F_{\alpha, J-1, \nu_W}$ represents the α-level *F* critical value with J-1 and ν_W degrees of freedom.

Welch's (1951) Heteroscedastic F Test with Trimmed Means and Winsorized Variances. 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) note, “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).

Let $\lambda = [\kappa\, n]$, where $[\kappa\, n]$ is the largest integer $\leq \kappa\, n$ and κ represents the proportion of observations trimmed from each tail of the distribution. Then, $h = n - 2\lambda$ represents the effective sample size (i.e., the sample size after trimming). The sample trimmed mean is:

$$\overline{X}_t = \frac{1}{h}\sum_{i=\lambda+1}^{n-\lambda} X_i .$$

The sample Winsorized mean, which represents the sample mean after replacing the trimmed observations in the lower and upper tails with the lowest and highest untrimmed observations, respectively, is:

$$\overline{X}_W = \frac{1}{n}\sum_i Y_i ,$$

where:

$$Y_i = \begin{cases} X_{(\lambda+1)} \; if \;\; X_i \leq X_{(\lambda+1)} \\ X_i \; if \;\; X_{\lambda+1} < X_i < X_{(n-\lambda)} \\ X_{(n-\lambda)} \; if \;\; X_i \geq X_{(n-\lambda)} \end{cases}.$$

The sample Winsorized variance is:

$$s_W^2 = \frac{\sum_i \left(Y_i - \overline{X}_W\right)^2}{n-1}.$$

Let n_j, h_j, s_{Wj}, and 0_{tj} represent the values of n, h, s_W, and 0_t for the jth group, and let:

$$q_j = \frac{\left(n_j - 1\right)S_{Wj}^2}{h_j\left(h_j - 1\right)},$$

$$w_j = \frac{1}{q_j},$$

$$U = \sum_j w_j,$$

$$\tilde{X} = \frac{1}{U}\sum_j w_j \overline{X}_{tj},$$

$$A = \frac{1}{J-1}\sum_j w_j\left(\overline{X}_{tj} - \tilde{X}\right)^2,$$

$$B = \frac{2(J-2)}{J^2-1}\sum_j \frac{\left(1-\frac{w_j}{U}\right)^2}{h_j - 1}, and$$

$$F_t = \frac{A}{B+1}.$$

The null hypothesis H_o: $\mu_{t1} = \ldots = \mu_{tJ}$ is rejected if $F_t \geq F_{\alpha, J-1, \nu_{Wt}}$, where:

$$\nu_{Wt} = \frac{1}{\frac{3}{J^2-1}\sum_j \frac{\left(1-\frac{w_j}{U}\right)^2}{h_j - 1}}.$$

Wilcox (1996, 1998a, 1998b) and Rosenberger and Gasko (2000) have found through simulation that 20% symmetric trimming provides an excellent balance between Type I error

control and power for many nonnormal distributions; in other words, too much trimming would reduce power by substantially reducing the effective sample size, whereas not enough trimming would not provide adequate control over the probability of Type I errors (and for some distributions may not provide sufficient power).

This issue of symmetric versus asymmetric trimming (i.e., trimming the same proportion of observations from each tail, or trimming a greater proportion of observations from the longer tail, respectively) is another important topic of interest. Although some authors have found that asymmetric trimming, and related M-estimators that remove empirical outliers, can provide better Type I error control and/or power than symmetric trimming for certain asymmetric distributions (e.g., Hogg, Fisher, & Randles, 1975; Keselman et al., 2002; Wilcox, 2003), asymmetric strategies become much more conceptually and computationally intense because they require the calculation of the optimal proportion of trimming from each tail. In this study we focus on symmetric trimming because of its conceptual simplicity and availability to applied researchers (i.e., the test statistics for symmetric trimming can be adopted with popular software packages such as R and SPSS).

When trimmed means are being compared the null hypothesis pertains to the equality of population trimmed means, i.e., the μ_ts, not the usual population means (μs). This is an important point for the reader to remember. Some readers may prefer to give equal weight to all of the observations, which results in using the usual mean, rather than giving the more extreme values zero weight, as is done when using a trimmed mean. 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 papers have demonstrated that one can generally

achieve robustness to nonnormality and variance heterogeneity in unbalanced independent (and correlated groups) designs by using robust estimators with heteroscedastic test statistics (Algina, Keselman & Penfiled, 2005; Keselman, Algina, Wilcox & Kowalchuk, 2000; Keselman, Kowalchuk & Lix, 1998). Further, by comparing the μ_ts, the researcher has the advantage of comparing the bulk of the data from each distribution, thus minimizing the drastic effects that extreme cases can have on the usual means and variances.

James Second Order Test. James (1951) developed a heteroscedastic one-way independent groups test. Like the Welch test above, this test was designed to be insensitive to the effects of variance inequality, even when paired with unequal sample sizes. The first step in computing the James procedure is to compute the standard error, S_j, for each of the J groups. A weight, a_j, for each group is computed as:

.

$$a_j = \frac{1/S_j^2}{\sum_{j=1}^{J} 1/S_j^2}$$

A variance weighted common mean is calculated as:

$$\overline{Y} = \sum_j a_j \overline{X}_j$$

and a *t* statistic can be calculated for each group as:

.

H_o: $\mu_1 = ... = \mu_J$ is rejected if $t_j = \frac{\overline{X}_j - \overline{Y}}{S_j}$ $J > CV_\alpha$ where:

$$J = \sum_j t_j^2$$

and

$$CV_\alpha = C + \frac{1}{2}\left(3\chi_4 + \chi_2\right)T + \left\{\begin{array}{l} \frac{1}{16}\left(3\chi_4 + \chi_2\right)^2\left(1 - \frac{J-3}{C}\right)T^2 \\ + \frac{1}{2}\left(3\chi_4 + \chi_2\right)\left[\begin{array}{l}\left(8R_{23} - 10R_{22} + 4R_{21} - 6R_{12}^2 + 8R\right) \\ + \left(2R_{23} - 4R_{22} + 2R_{21} - 2R_{12}^2 + 4R_{12}R_{11} - 2R_{11}^2\right)\left(\chi_2 - 1\right) \\ + \frac{1}{4}\left(-R_{12}^2 + 4R_{12}R_{11} - 2R_{12}R_{10} - 4R_{11}^2 + 4R_{11}R_{10} - R_{10}^2\right) \\ \left(3\chi_4 - 2\chi_2 - 1\right)\end{array}\right] + \\ \left(R_{23} - 3R_{22} + 3R_{21} - R_{20}\right)\left(5\chi_6 + 2\chi_4 + \chi_2\right) \\ + \frac{3}{16}\left(R_{12}^2 - 4R_{23} + 6R_{22} - 4R_{21} + R_{20}\right)\left(35\chi_8 + 15\chi_6 + 9\chi_4 + 5\chi_2\right) \\ + \frac{1}{16}\left(-2R_{22} + 4R_{21} - R_{20} + 2R_{12}R_{10} - 4R_{11}R_{10} - 4R_{11}R_{10} + R_{10}^2\right) \\ \left(9\chi_8 - 3\chi_6 - 5\chi_4 - \chi_2\right) + \frac{1}{4}\left(-R_{22} + R_{11}^2\right)\left(27\chi_8 + 3\chi_6 + \chi_4 + \chi_2\right) \\ + \frac{1}{4}\left(R_{23} - R_{12}R_{11}\right)\left(45\chi_8 + 9\chi_6 + 7\chi_4 + 3\chi_2\right) \end{array}\right\}$$

Parametric Bootstrap Procedure. The parametric bootstrap procedure is an alternative heteroscedastic test statistic that was shown to provide a better balance of Type I error control and power than the original Welch test when sample sizes are small and there are many groups (Krisnamoorty et al., 2007). The first step in the parametric bootstrap procedure is to compute the sample test statistic T_{N0}, where:

$$T_{N0} = \sum_j \frac{n_j}{s_j^2} \overline{X}_j^2 - \frac{\left(\sum_j \frac{n_j \overline{X}_j}{s_j^2} \right)^2}{\sum_j \frac{n_j}{s_j^2}}.$$

Then, after completing i = 1, ..., I bootstrap samples, reject Ho: $\mu_1 = ... = \mu_J$ if $\Sigma_i(T_{NBi} > T_{N0})/I \leq \alpha$, where T_{NBi} represents the test statistic for the ith bootstrap sample, and:

$$T_{NB} = \sum_j \frac{z_j^2 \left(n_j - 1 \right)}{\chi^2_{n_j - 1}} - \frac{\left\{ \sum_j \frac{\sqrt{n_j} z_j \left(n_j - 1 \right)}{s_j \chi^2_{n_j - 1}} \right\}^2}{\sum_j \frac{n_j \left(n_j - 1 \right)}{s_j^2 \chi^2_{n_i - 1}}}.$$

z_j is a standard normal random variable and $\chi^2_{n_i-1}$ is a random chi-square variable with $n_j - 1$ degrees of freedom.

Trimmed Parametric Bootstrap Procedure. A trimmed version of the parametric bootstrap procedure was also utilized in order to be able to determine how removing the outlying cases affects the parametric bootstrap method. This procedure is identical to the original bootstrap procedure, except that the usual means and sample sizes were replaced with the trimmed means and effective sample sizes, and the usual variances were replaced by:

$$s_{t_j}^2 = \frac{s_{w_j}^2 \left(n_j \right)}{h_j - 1}.$$

Method

A Monte Carlo study was conducted to determine the empirical Type I error rates and power for the original Welch (1951) *F*-test, the Welch *F*-test with trimmed means and Winsorized variances, the James (1951) second-order test, the parametric bootstrap procedure, and the trimmed parametric bootstrap procedure. A Monte Carlo study is an effective way to assess and compare the performance of test statistics when the assumptions underlying these test statistics are violated (Serlin, 2000). For the parametric bootstrap procedure, 2500 bootstrap samples were generated for each simulation. Although we, and many others, would recommend running many more bootstrap samples (e.g., 10000) for a single analysis, the number of bootstrap samples was set at 2500 in the investigation given how computationally intensive it is to run these bootstrap samples (for multiple procedures) over thousands of simulations. Several variables were manipulated: 1) number of groups; 2) group sample sizes; 3) population variances; 4) population means; and 5) distribution shapes. A summary of these conditions is presented in Table 1.

The conditions investigated represent data characteristics that are common in applied studies, including extreme cases of nonnormality and variance heterogeneity. The number of groups was set at 3 and 8, which are used to represent studies with a small and large number of groups, respectively. Group sample sizes were set to be equal (all n = 30) or unequal (n = 20, 30, 40 for 3 groups; n = 19, 22, 25, 28, 31, 34, 37, 40 for 8 groups). A sample size of n = 30 was thought to represent many studies in psychology, and, as discussed below, is used in conjunction with the population means and variances to establish representative power conditions. Population variances were set to be equal (all $\sigma^2_j = 1$), moderately unequal (largest to smallest variance ratio of 4:1) or extremely unequal (largest to smallest variance ratio of 9:1). Both positive (largest n

paired with largest population variance, and smallest n paired with smallest population variance) and negative (largest n paired with largest population variance, and smallest n paired with smallest population variance) pairings of unequal n and variances were investigated. It is important in Monte Carlo studies to investigate both equal and unequal variance as numerous previous studies have found that group variances are often extremely disparate (e.g., Keselman et al., 1998; Golinski & Cribbie, 2009). Population means for the Type I error conditions were all set equal to 0, and population means for the power conditions (for J =3, μ_j = 0, .4, .8; for J = 8, μ_j = 0, .11, .22, .33, .44, .55, .66, .77) were selected such that the power for the ANOVA *F* test with equal sample sizes and variances and normal distributions was .8. The resulting population η^2 values for J =3 and J = 8 were .097 and .059, respectively. An initial check of the program with equal sample sizes and variances and normal distributions indicated that the Type I error rates for the ANOVA *F* for both J = 3 and J = 8 were equal to α, and the power values for J = 3 and J = 8 were .79 and .80, respectively. It is important to note that the nominal power value of .80 is only expected when distribution shapes are normal and variances are equal. In other cases the population variances are not all 1 (i.e., some or all are greater than 1) and therefore power will be decreased.

Distribution shapes were either all normal, all moderately skewed, all extremely skewed, or a mixture of normal and nonnormal shapes. The skewed distributions were generated using the *g*- and *h*-distribution (Hoaglin, 1985). The *g*- and *h*- distributions used in this study were *g*=.5, *h*= 0 (moderately skewed, skewness = 1.75, kurtosis = 8.90) and *g*=1, *h*=0 (extremely skewed, skewness = 6.18, kurtosis = 113.94). To give meaning to these values it should be noted that for the standard normal distribution *g* = 0 and *h* = 0. When *g* = 0, a distribution is symmetric

and the tails of a distribution will become heavier as h increases in value. As g increases, the distribution becomes increasingly positively skewed. According to Wilcox (1994, 1995), the distributions used in this study are representative of the levels of skewness for dependent variables in many scientific inquiries.

To generate pseudo-random normal variates, we used the R generator 'rnorm' (R Development Core Team, 1995). If Z_{ij} is a standard normal variate, then $X_{ij} = \mu_j + \sigma_j Z_{ij}$ is a normal variate with mean equal to μ_j and standard deviation equal to σ_j. To generate data from a *g*- and *h*-distribution, standard unit normal variables (Z_{ij}) were converted to the random variable:

$$X_{ij} = \frac{e^{gZ_{ij}} - 1}{g} e^{\frac{hZ_{ij}^2}{2}},$$

according to the values of *g* and *h* selected for investigation. To obtain a distribution with standard deviation σ_j, each X_{ij} was multiplied by a value of σ_j (from Table 1). It is important to note that this does not affect the value of the null hypothesis when $g = 0$ (see Wilcox, 1994). However, when $g > 0$, the population mean for a *g*- and *h*- variable is:

$$\mu_{gh} = \frac{1}{\sqrt{g(1-h)}}\left(e^{\frac{g^2}{2(h-1)}} - 1 \right)$$

Thus, for those conditions where $g > 0$, μ_{gh} was first subtracted from X_{ij} before multiplying by σ_j. When working with trimmed means, the proportion of observations trimmed from each tail of the distribution (κ) was set at .2, and the population trimmed mean for the jth group was also subtracted from the variate before multiplying by σ_j. Lastly, it should be noted that the standard deviation of a *g*- and *h*-distribution is not equal to one, and thus the values enumerated in Table 1 reflect only the amount that each random variable is multiplied by and not the actual

values of the standard deviations (see Wilcox, 1994).

Empirical Type I error and power rates were recorded for all tests. The robustness of a procedure, with respect to Type I error control, was determined using Bradley's (1978) liberal criterion. That is, a procedure is deemed robust with respect to Type I errors if the empirical rate of Type I error falls within the range +/- .5 α. We use a benchmark of .025 < □ < .075 (□ is the empirical rate of Type I error) to define a robust test, when the criterion of significance is set at α = .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 to (i.e., not substantially affected by) 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. 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 at α = .05?). However, in our view, Bradley's liberal criterion is acceptable in this study because we are investigating extreme conditions of sample size and variance inequality, nonnormality and distribution shape heterogeneity, and therefore we are not expecting empirical Type I error rates to precisely equal the nominal significance level.

The simulation program was written in R (R Development Core Team, 2005). Five thousand replications of each condition were performed, resulting in a standard error of approximately .0015 for the mean empirical power and Type I error rates. A nominal significance level of .05 was adopted in all analyses.

Results

The pattern of results was similar for the moderately and extremely unequal variance conditions and therefore the results were averaged over these conditions. As expected, more extreme variance ratios had a larger impact on the Type I error rates and power of the procedures, but the overall recommendations, when averaging over these conditions, do not change. Further, we do not present the results for equal sample sizes and unequal variances, with the rationale being that if a procedure performs satisfactorily with unequal sample sizes and unequal variances, then it is also expected to perform satisfactorily with equal sample sizes and unequal variances.

Type I error Rates

Empirical Type I error rates for J = 3 and J = 8 are presented in Tables 2 and 3, respectively. The results indicate that although the empirical Type I error rates for the Welch (1951), James (1951) and parametric bootstrap procedures were acceptable when all distributions were normal or when all distributions were moderately skewed, the Type I error rates deviated considerably from the nominal level when all distribution shapes were extremely nonnormal, or when the distribution shapes were dissimilar. For example, when there were three groups and the first had extremely skewed data and the next two had normally distributed data, the empirical Type I error rates for the Welch, James and parametric bootstrap procedures, even with equal population variances, was approximately .115 (i.e., more than double the nominal .05 rate). When distribution shapes were dissimilar and population variances were unequal, Type I error rates for the Welch, James and parametric bootstrap procedures with J = 8 exceeded .20. On the

other hand, the Welch and parametric bootstrap procedures with trimmed means generally provided excellent Type I error control across the conditions, with the Type I error rates straying above .075 in only a few instances, specifically when there were 8 groups and the distribution shapes were dissimilar.

In order to determine how accurate the Type I error rates would be at smaller sample sizes, we also ran conditions with an average sample size of 10 (for J = 3, n_j = 10, 10, 10 and n_j = 7, 10, 13; for J = 8, n_j = 10, 10, 10, 10, 10, 10, 10, 10 and n_j = 6, 7, 8, 9, 11, 12, 13, 14). We ran these conditions under the most extreme pattern of nonnormality (i.e., the pattern that resulted in the largest deviations of the empirical Type I error rates from α for n = 30), namely the pattern where some of the distribution shapes were normal and some were extremely positively skewed (for J =3, two of the distributions were normal and one was positively skewed, and for J = 8 four of the distributions were normal and four were positively skewed). The pattern of results for J = 3 was identical to that for n = 30; the Type I error rates for the Welch and parametric bootstrap tests with trimmed means never strayed outside of Bradley's liberal bounds, whereas the Type I error rates for the Welch, James and Parametric Bootstrap procedures commonly exceeded .075. For J = 8, the Type I error rates were more liberal for n = 10 than for n = 30. For the Welch, James and Parametric Bootstrap procedures the Type I error rates were much more liberal, with rates ranging from .12 - .27, whereas the Type I error rates for the Welch and parametric bootstrap tests with trimmed means were moderately inflated, with rates reaching as large as .13.

Power

Empirical Power rates for J = 3 and J = 8 are presented in Tables 4 and 5, respectively. Given the unsatisfactory Type I error rates for the Welch, James, and parametric bootstrap

procedures under many conditions, an investigation of power may be unnecessary due to the fact that generally only the tests with trimmed means could be recommended. However, the power results do help to highlight that the Welch, James and parametric bootstrap procedures are generally less powerful than the tests based on trimmed means. Therefore, not only did we find that the Welch and parametric bootstrap procedures based on trimmed means had substantially improved Type I error control relative to the usual Welch, James, and parametric bootstrap procedures, but also that they were generally more powerful than the other procedures, even in conditions where the empirical Type I error rates for the other procedures were well above .075.

Discussion

Given the popularity of independent groups designs in experimental research, there is still considerable interest in finding a test statistic that is appropriate under cases of distribution nonnormality and variance heterogeneity across the groups. This study addressed two important questions related to reliably and validly assessing treatment effects in this design : 1) How will the recently proposed parametric bootstrap test proposed by Krishnamoorthy et al. (2007), or a modified procedure based on trimmed means, perform when distributions are nonnormal? and 2) How well will previously recommended test statistics for analyzing nonnormal data with unequal variances (i.e., the original Welch, 1951, heteroscedastic test and the Welch heteroscedastic test with trimmed means and Winsorized variances) perform, relative to the original parametric bootstrap procedure or a modified parametric bootstrap procedure based on trimmed means, when distribution shapes vary across groups? These questions have not previously been addressed. We used Monte Carlo methods to examine these questions, varying both the variances and the shapes of distributions across treatment groups.

The tests based on trimmed means (Welch or parametric bootstrap) were the only tests that provided acceptable Type I error control across the vast majority of conditions examined in our study. However, Type I error rates did on occasion exceed the nominal level when there were eight groups and distribution shapes were dissimilar; however, this only rarely occurred relative to the Type I error inflation that occurred for the procedures not based on trimmed means. An important extension of these results involves the appropriateness of comparing central tendencies (e.g., means, trimmed means) when the distribution shapes differ.

The Welch (1951) and James (1951) *F*-tests with the usual means and variances and the original parametric bootstrap procedure had Type I error rates that deviated substantially from the nominal level when all distribution shapes were nonnormal or when the distribution shapes were dissimilar. Type I error rates were generally more extreme when distribution shapes were dissimilar, with maximum rates for the Welch, James and parametric bootstrap procedure occurring when the distributions had opposite skews. With respect to power, generally the Welch (1951) *F*-test with trimmed means and the parametric bootstrap procedure with trimmed means were more powerful than the original Welch, James or the parametric bootstrap procedure. It is also noteworthy to point out that the results for the Welch, James and parametric bootstrap procedures were very similar across all of the conditions investigated in this study, and therefore there is little to be gained by adopting the bootstrapping procedure.

Thus, based on the results of our investigation, the Welch (1951) test or the parametric bootstrap procedure (Krishnamoorthy et al., 2007) using robust estimators (trimmed means and Winsorized variances) are preferable to the test statistics based on the usual means and variances because they provide much better Type I error control and are more sensitive to the presence of

treatment effects in the presence of nonnormal data where variances are unequal as well.

References

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

Algina, J., Oshima, T. C. & Lin, W. (1994). Type I error rates for Welch's test and James' second-order test under nonnormality and inequality of variance when there are two groups. *Journal of Educational and Behavioral Statistics, 19*, 275-291.

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

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

Chen, L. and Shapiro, S. (1995) An Alternative test for normality based on normalized spacings. Journal of Statistical Computation and Simulation 53, 269-287.

Cribbie, R. A., Wilcox, R. R., Bewell, C., & Keselman, H. J. (2007). Tests for treatment group equality when data are nonnormal and heteroscedastic. *Journal of Modern Applied Statistical Methods, 6*, 117-132.

D'Agostino, R. B. (1971) An omnibus test of normality for moderate and large size samples. Biometrika 58, 341-348.

Erceg-Hurn, D. M. & Mirosevich, V. M. (2008). An easy way to maximize the accuracy and power of your research. *American Psychologist, 63*, 591-601.

Glass, G V, Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed-effects analysis of variance and covariance. *Review of*

Educational Research, 42, 237-288.

Golinski, C. & Cribbie, R. A. (2009). The expanding role of quantitative methodologists in advancing psychology. *Canadian Psychology, 50,* 83-90.

Grissom, R. J. (2000). Heterogeneity of variance in clinical data. *Journal of Consulting and Clinical Psychology, 68,* 155-165.

Hampel, F. R., Ronchetti, E. M., Rousseuw, 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.

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

Hogg, R. V., Fisher, D. M., & Randles, R. H. (1975). A two-sample adaptive distribution free test. *Journal of the American Statistical Association, 70,* 656–661.

Holgersson, H. E. T. (2006). A graphical method for assessing multivariate normality. Computational Statistics and Data Analysis, 21, 141-149.

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

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., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R., Donahue, B., Kowalchuk, R. K., Lowman, L. L., Petoskey, M. D., Keselman, J. C., & Levin, J. R. (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., & Lix, L. M. (1998). Robust nonorthogonal analyses revisted: An update based on trimmed means. *Psychometrika, 63*, 145-163.

Keselman, H. J., Wilcox, R. R., Othman, A. R., & Fradette, K. (2002). Trimming, transforming statistics, and bootstrapping: Circumventing the biasing effects of heteroscedasticity and nonnormality. *Journal of Modern Applied Statistical Methods, 1*, 281–287.

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.

Krishnamoorthy, K., Lu, F., & Mathew, T. (2007). A parametric bootstrap approach for ANOVA with unequal variances: Fixed and random models. *Computational Statistics and Data Analysis, 51*, 5731-5742.

Leentjens, A. F., Wielaert, S. M., van Harskamp, F., & Wilmink, F. W. (1998). Disturbances of affective prosody in patients with schizophrenia: A cross sectional study. *Journal of Neurology, Neurosurgery, and Psychiatry, 64*, 375-378.

Luh, W-M., & Guo, J-H. (2001). Using Johnson's transformation and robust estimators with heteroscedastic test statistics: An examination of the effects of non-normality and heterogeneity in the non-orthogonal two-way ANOVA design. *British Journal of Mathematical and Statistical Psychology, 54*, 79-94.

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

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*, 156-166.

R Development Core Team (2005). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.r-project.org.

Scheffé, H. (1959). The analysis of variance. New York: Wiley.

Seier, E. (2002). Comparison of tests of univariate normality. *Interstat, January 2002*.

Shapiro, S., & Wilk, M. B. (1965) An analysis of variance test for normality. *Biometrika, 52,* 591-611.

Serlin, R. C. (2000). Testing for robustness in Monte Carlo studies. *Psychological Methods, 5,* 230-240.

Sprent, P., & Smeeton, N. C. (2001). *Applied nonparametric statistical methods.* New York: Chapman & Hall/CRC.

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

Tiku, M. L. (1964). Approximating the general non-normal variance-ratio sampling distributions. *Biometrika, 51*, 83-95.

Tukey, J.W. (1960). *A survey of sampling from contaminated normal distributions*. In I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow, & H. B. Mann (Eds.), Contributions to probability and statistics (pp. 448-485). Stanford, CA: Stanford University Press.

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

Wilcox, R. R. (1987). A heteroscedastic ANOVA procedure with specified power. *Journal of Educational Statistics, 12*,271-281.

Wilcox, R. R. (1994). Some results on the Tukey-McLaughlin and Yuen methods for trimmed means when distributions are skewed. *Biometrical Journal, 36*, 259–273.

Wilcox, R. R. (1995). ANOVA: The practical importance of heteroscedastic methods, using trimmed means versus means, and designing simulation studies. *British Journal of Mathematical and Statistical Psychology, 48,* 99–114.

Wilcox, R. R. (1997). A bootstrap modification of the Alexander-Govern ANOVA method, plus comments on comparing trimmed means. *Educational and Psychological Measurement, 57*,655–665.

Wilcox, R. R. (1998a). How many discoveries have been lost by ignoring modern statistical methods? *American Psychologist, 53*, 300-314.

Wilcox, R. R. (1998b). The goals and strategies of robust methods. *Journal of Mathematical and Statistical Psychology, 51*, 1-39.

Wilcox, R. R. (2005). *Introduction to Robust Estimation and Hypothesis Testing* (2nd Ed.). San Diego, CA: Academic Press.

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

Table 1.
Summary of conditions investigated in the Monte Carlo study (all conditions are crossed).

J	Sample Sizes	σ_j	Population Distribution Shapes
3	30 X 3	1, 1, 1	Normal X 3
	20, 30, 40	1, 1.5, 2	g=.5 X 3
		1, 2, 3	g=1 X 3
		2, 1.5, 1	Normal X 2, g=1
		3, 2, 1	Normal, g=.5, g=1
			g=1, Normal X 2
			g=1, g=.5, Normal
			g=1(-skew), Normal, g=1(+skew)
			g=1(+skew), Normal, g=1(-skew)
8	30 X 8	1 X 8	Normal X 8
	19, 22, 25, ..., 40	1, 1.14, 1.28, ..., 1.98	g=.5 X 8
		1, 1.28, 1.56, ..., 2.96	g=1 X 8
		1.98, 1.84, 1.70, ..., 1	Normal X 4, g=1 X 4
		2.96, 2.68, 2.40, ..., 1	Normal X 3, g=.5 X 2, g=1 X 3
			g=1 X 4, Normal X 4
			g=1 X 3, g=.5 X 2, Normal X 3
			g=1 (-skew) X 4, g=1 (+skew) X 4
			g=1 (+skew) X 4, g=1 (-skew) X 4

Note: J represents the number of groups; σ represents the population standard deviation; ‘a X 2' indicates that ‘a’ is replicated 2 times; g=.5 represents a moderately skewed distribution shape; g=1 represents an extremely skewed distribution shape (h=0 for all g- and h- distribution shapes).

Table 2.
Type I error rates for each test statistic for J = 3.

		Test Statistic				
Distribution Shape	Population Variances	Welch	James	$Welch_t$	PB	PB_t
normal X 3	Equal	.054	.054	.054	.054	.050
	PP	.053	.052	.054	.053	.047
	NP	.046	.055	.057	.054	.051
g=.5 X 3	Equal	.049	.049	.054	.049	.051
	PP	.053	.053	.051	.054	.046
	NP	.065	.065	.052	.066	.056
g=1 X 3	Equal	.052	.052	.050	.052	.048
	PP	.063	.063	.053	.063	.045
	NP	.110	.110	.063	.109	.062
normal X 2, g=1	Equal	.081	.081	.052	.081	.054
	PP	.092	.092	.056	.093	.054
	NP	.058	.057	.054	.058	.056
normal, g=.5, g=1	Equal	.075	.075	.050	.074	.052
	PP	.083	.084	.049	.084	.052
	NP	.048	.048	.055	.050	.050
g=1, normal X 2	Equal	.115	.115	.053	.114	.050
	PP	.075	.075	.048	.076	.045
	NP	.123	.123	.058	.122	.055
g=1, g=.5, normal	Equal	.101	.101	.054	.102	.055
	PP	.066	.066	.051	.067	.042
	NP	.127	.126	.065	.126	.059
g=1, normal, g=1(-)	Equal	.123	.123	.062	.123	.058
	PP	.118	.118	.055	.117	.051
	NP	.126	.126	.064	.126	.058
g=1(-), normal, g=1	Equal	.129	.129	.056	.129	.050
	PP	.113	.123	.053	.121	.053
	NP	.121	.121	.067	.121	.053

Note: g=.5 represents a moderately skewed distribution shape; g=1 represents an extremely skewed distribution shape (h=0 for all g- and h- distribution shapes); PP = positively paired sample sizes and variances; NP = negatively paired sample sizes and variances.

Table 3.
Type I error rates for each test statistic for J = 8.

		Test Statistic				
Distribution Shape	Population Variances	Welch	James	$Welch_t$	PB	PB_t
normal X 8	Equal	.050	.049	.051	.049	.046
	PP	.051	.050	.059	.049	.047
	NP	.052	.051	.061	.051	.048
g=.5 X 8	Equal	.064	.063	.060	.063	.050
	PP	.064	.064	.057	.063	.048
	NP	.074	.073	.064	.073	.053
g=1 X 8	Equal	.096	.095	.059	.098	.051
	PP	.100	.099	.061	.099	.053
	NP	.143	.141	.075	.140	.063
normal X 4,	Equal	.131	.130	.063	.130	.058
g=1 X 4	PP	.163	.161	.067	.161	.061
	NP	.105	.105	.057	.104	.046
normal X 3,	Equal	.109	.108	.053	.108	.047
g=.5 X 2,	PP	.139	.139	.066	.138	.058
g=1 X 3	NP	.088	.087	.060	.086	.047
g=1 X 4,	Equal	.196	.194	.067	.195	.058
normal X 4	PP	.142	.141	.063	.141	.057
	NP	.221	.221	.083	.221	.071
g=1 X 3,	Equal	.164	.163	.068	.165	.060
g=.5 X 2,	PP	.127	.127	.061	.126	.047
normal X 3	NP	.199	.197	.083	.198	.070
g=1(+) X 4,	Equal	.286	.283	.080	.283	.075
g=1(-) X 4	PP	.263	.261	.077	.262	.068
	NP	.283	.280	.085	.280	.074
g=1(-) X 4,	Equal	.278	.276	.079	.277	.074
g=1(+) X 4	PP	.264	.263	.081	.263	.071
	NP	.282	.282	.092	.280	.080

Note: g=.5 represents a moderately skewed distribution shape; g=1 represents an extremely skewed distribution shape (h=0 for all g- and h- distribution shapes); PP = positively paired sample sizes and variance; NP = negatively paired sample sizes and variances.

Table 4.
Power rates for each test statistic for J = 3.

		Test Statistic				
Distribution Shape	Population Variances	Welch	James	$Welch_t$	PB	PB_t
normal X 3	Equal	.724	.723	.644	.726	.623
	PP	.360	.361	.311	.361	.299
	NP	.301	.301	.257	.300	.244
g=.5 X 3	Equal	.634	.633	.662	.635	.610
	PP	.213	.214	.263	.213	.265
	NP	.333	.333	.338	.333	.282
g=1 X 3	Equal	.384	.383	.630	.383	.629
	PP	.101	.101	.224	.103	.222
	NP	.299	.299	.337	.300	.325
normal X 2, g=1	Equal	.380	.380	.624	.382	.600
	PP	.144	.145	.250	.142	.235
	NP	.166	.165	.258	.166	.254
normal, g=.5, g=1	Equal	.375	.375	.584	.378	.571
	PP	.108	.108	.210	.107	.205
	NP	.170	.170	.225	.171	.212
g=1, normal X 2	Equal	.582	.576	.623	.582	.614
	PP	.352	.352	.345	.351	.332
	NP	.335	.334	.312	.334	.308
g=1, g=.5, normal	Equal	.581	.581	.685	.580	.576
	PP	.342	.343	.355	.341	.334
	NP	.367	.365	.256	.367	.254
g=1, normal, g=1(-)	Equal	.484	.483	.716	.483	.692
	PP	.334	.334	.418	.335	.403
	NP	.329	.329	.375	.328	.362
g=1(-), normal, g=1	Equal	.105	.104	.493	.104	.470
	PP	.053	.052	.216	.052	.199
	NP	.081	.081	.156	.080	.137

Note: g=.5 represents a moderately skewed distribution shape; g=1 represents an extremely skewed distribution shape (h=0 for all g- and h- distribution shapes); PP = positively paired sample sizes and variance; NP = negatively paired sample sizes and variances; greyed out = Type I error rate >.075.

Table 5.
Power rates for each test statistic for J = 8.

		Test Statistic				
Distribution Shape	Population Variances	Welch	James	$Welch_t$	PB	PB_t
normal X 8	Equal	.774	.770	.687	.770	.649
	PP	.339	.337	.299	.336	.249
	NP	.323	.320	.289	.320	.265
g=.5 X 8	Equal	.654	.653	.661	.651	.649
	PP	.246	.244	.278	.243	.255
	NP	.369	.364	.328	.365	.324
g=1 X 8	Equal	.460	.454	.667	.457	.653
	PP	.155	.154	.228	.154	.221
	NP	.387	.384	.371	.385	.364
normal X 4,	Equal	.298	.295	.572	.297	.558
g=1 X 4	PP	.121	.120	.168	.119	.158
	NP	.159	.157	.243	.158	.227
normal X 3,	Equal	.293	.289	.549	.289	.544
g=.5 X 2,	PP	.105	.104	.172	.103	.159
g=1 X 3	NP	.145	.143	.226	.142	.217
g=1 X 4,	Equal	.721	.720	.759	.720	.746
normal X 4	PP	.473	.472	.389	.472	.371
	NP	.502	.501	.435	.501	.421
g=1 X 3,	Equal	.718	.715	.747	.716	.723
g=.5 X 2,	PP	.442	.440	.365	.440	.335
normal X 3	NP	.485	.484	.402	.484	.397
g=1 X 4,	Equal	.691	.690	.826	.689	.801
g=1(-) X 4	PP	.515	.514	.523	.515	.520
	NP	.515	.514	.489	.513	.487
g=1(-) X 4,	Equal	.127	.125	.416	.124	.394
g=1 X 4	PP	.151	.151	.136	.151	.132
	NP	.176	.175	.139	.174	.135

Note: g=.5 represents a moderately skewed distribution shape; g=1 represents an extremely skewed distribution shape (h=0 for all g- and h- distribution shapes); PP = positively paired sample sizes and variance; NP = negatively paired sample sizes and variances; greyed out = Type I error rate >.075.