Metodološki zvezki, Vol. 3, No. 1, 2006, 49-62

Comparing the "Typical Score" Across Independent Groups Based on Different Criteria for Trimming

S.S. Syed Yahaya[1], A.R. Othman[2], and H.J. Keselman[3]

Abstract

Nonnormality and variance heterogeneity affect the validity of the traditional tests for treatment group equality (e.g. ANOVA *F*-test and *t*-test), particularly when group sizes are unequal. Adopting trimmed means instead of the usual least squares estimator has been shown to be mostly affective in combating the deleterious effects of nonnormality. There are, however, practical concerns regarding trimmed means, such as the predetermined amount of symmetric trimming that is typically used. Wilcox and Keselman proposed the Modified One-Step M-estimator (*MOM*) which empirically determines the amount of trimming. Othman et al. found that when this estimator is used with Schrader and Hettmansperger's *H* statistic, rates of Type I error were well controlled even though data were nonnormal in form. In this paper, we modified the criterion for choosing the sample values for *MOM* by replacing the default scale estimator, MAD_n, with two robust scale estimators, S_n and T_n , suggested by Rousseeuw and Croux (1993). To study the robustness of the modified methods, conditions that are known to negatively affect rates of Type I error were manipulated. As well, a bootstrap method was used to generate a better approximate sampling distribution since the null distribution of *MOM-H* is intractable. These modified methods resulted in better Type I error control especially when data were extremely skewed.

[1] Fakulti Sains Kuantitatif, Universiti Utara Malaysia, 06010 Sintok, Kedah, Malaysia; sharipah@uum.edu.my

[2] Pusat Pengajian Pendidikan Jarak Jauh, Universiti Sains Malaysia, 11800 Minden, Penang, Malaysia; oarahman@usm.my

[3] Dept. of Psychology, University of Manitoba, 190 Dysart Road, Winnipeg, Manitoba R3T 2N2, Canada; kesel@cc.umanitoba.ca

1 Introduction

Parametric procedures for testing the equality of central tendency measures, such as the ANOVA *F*-test and Student's two-sample *t*-test, are adversely affected by nonnormality, variance heterogeneity, particularly when the design is unbalanced (i.e, groups sizes are unequal). Specifically, violations by any of these assumptions can seriously inflate Type I error rates; that is, spurious rejections of null hypotheses of equal means can increase. Nevertheless, the ANOVA *F*-test, for example, is often used in statistical practice even when the data suggest that population variances are unequal (Kulinskaya, Staudte, & Guo, 2003), and even though it is well established that the ANOVA is not robust when the homogeneity assumption does not hold (Wilcox, Charlin, & Thompson, 1986).

In order to overcome the biasing effects of nonnormality and variance heterogeneity, alternative methods have been recommended. Cochran (1937), as noted by Kulinskaya et al. (2003), suggested weighting the terms in the sum of squares explained by the respective inverses of the sample variances, and provided a chi-square test for equal means based on a transformation of the ANOVA *F*-test. However, the design still has to be balanced. Kulinskaya et al. (2003) also noted that for unbalanced designs, the James (1951) and Welch (1951) procedures weight the terms (the sum of squares explained) by estimates of the inverses of the variances of the respective sample means. This weighted sum of squares for explained variance possesses an approximate chi-squared distribution under the null hypotheses of equal population means for large sample sizes.

Nonetheless, even if the problem of unequal variances could be overcome, the assumption of normality must also be satisfied with classical procedures that employ the usual least squares estimates. Furthermore, although ANOVA is known to be robust to small deviations from normality, the extent of these deviations are unknown since there is no exact measurement of these violations or deviations, unless the sample size is large enough to guarantee normality.

Nonparametric counterparts of these procedures, namely the Kruskal-Wallis and the Mann-Whitney tests, were developed to deal with such problems. However, these nonparametric procedures are more appropriate for nonnormal symmetric data. Furthermore, nonparametric procedures are frequently less powerful than parametric procedures, and, accordingly, require larger sample sizes to reject a false hypothesis.

Violations to the homogeneity of variances and/or normality assumptions are common in the behavioural and social sciences (See discussions by Wilcox, 1997, 2003). Thus, researchers should expect distortions of rates of Type I error for classical tests of mean equality. Robust statistical procedures, that is those that use non least squares estimators (e.g., trimmed means), are useful and viable alternatives to traditional methods as they have been shown to (typically) control

rates of Type I error (Keselman, et al., 2002; 2000; Othman et al., 2004; Syed Yahaya et al., 2004; Wilcox et al.1988; 2001). One such test, the *MOM-H* statistic, originally proposed by Wilcox and Keselman (2003), modifies the well-known one-step M-estimator and applies the estimates in a robust test statistic (to be defined).

2 Methods

One of the strategies adopted when dealing with extreme values is trimming. There are many trimming strategies, however, in this paper we consider two, namely, (1) trim a predetermined amount of the data and then compute $\hat{\theta}$, a robust estimator, or (2) empirically determine the amount of trimming, trim that amount, and then computing $\hat{\theta}$. Trimming needs to be done carefully to avoid the loss of information during the process. For instance, when sampling from a light-tailed distribution, it might be desirable to trim very few observations, or if sampling is from a normal distribution, trimming might not be needed at all. For a right-skewed distribution, a natural reaction is to trim more observations from the right versus the left tail of the empirical distribution.

The usual approach to trimming outlying values is to trim symmetrically from each tail of the empirical distribution. By using this method of trimming, even observations from a normal distribution will be trimmed according to a predetermined amount such as 10% or 20% from each tail (in other words, 20% and 40% of the data are removed), when such distribution needs no trimming at all. Furthermore, any trimmed mean has a breakdown point (the number of extreme values that causes the estimator to inflate to an extreme value), which implies that a trimmed mean may not withstand large proportions of extreme values.

To mitigate these drawbacks, Wilcox and Keselman (2003) introduced a robust estimator known as the Modified One-Step *M*-estimator (*MOM*). The *MOM* estimator empirically determines the amount, if any, of data to be trimmed, and results indicate that it competes well with methods based on symmetrically trimmed means with regard to Type I error control. Similar to the sample median, the *MOM* estimator is a robust central tendency estimator that possesses the highest breakdown point. Othman et al. (2004) used *MOM* as the central tendency measure in their work with a robust statistic (*H*) presented by Schrader and Hettmansperger (1980) (*MOM-H*).

2.1 *MOM-H* statistic

The *H* test is defined as

$$H = \frac{1}{N}\sum_{j=1}^{J} n_j(\hat{\theta}_j - \hat{\theta}.)^2, \text{ where} \tag{2.1}$$

$$N = \sum_j n_j \text{ (n_j is the group size), and} \tag{2.2}$$

$$\hat{\theta}. = \sum_j \hat{\theta}_j / J \text{ where } j = 1,\ldots, J. \tag{2.3}$$

This statistic is readily adaptable to any measure of central tendency and appears to give reasonably good results when using the Harrel-Davis estimator of the median. However, its use is not recommended for the comparison of means or even trimmed means (Wilcox, 1997).

Othman et al. (2004) examined the operating characteristics of the *H* test statistic when testing for the equality of the "typical" score across treatment groups. However, they modified this statistic by replacing $\hat{\theta}$ with the *MOM* estimator (denoted as $\hat{\theta}_M$). The modified test statistic is known as *MOM-H*, and, as they indicated, this statistic can be used to test $H_0: \theta_{M1} = \theta_{M2} = \ldots = \theta_{MJ}$ versus $H_1: \theta_{Mi} \neq \theta_{Mj}$ for at least one pair of (*i*, *j*). Othman and his colleagues found that *MOM-H* was quite effective in controlling rates of Type I error even though data were heteroscedastic and nonnormal in shape.

In this paper, we modified the *MOM-H* statistic by substituting the default trimming criterion, incidentally the scale estimator, MAD_n, with two of the robust scale estimators suggested by Rousseuw and Croux (1993), i.e. S_n and T_n. We chose these substitutions because these scale estimators possess higher breakdown points and, accordingly, they may be better for screening the data for extreme values.

2.1.1 *MOM* estimator

MAD_n is the default scale estimator used in the criterion for determining extreme values when computing $\hat{\theta}_M$. Let $Y_j = (Y_{1j}, Y_{2j}, \ldots, Y_{nj})$ be a sample from an unknown skewed distribution F_j and let M_j be the population median of F_j. The estimator as suggested by Wilcox and Keselman (2003) is defined as

$$\hat{\theta}_{Mj} = \sum_{i=i_1+1}^{n_j - i_2} \frac{Y_{(i)j}}{n_j - i_1 - i_2} \tag{2.4}$$

where

$Y_{(i)j}$ = the ith ordered observations in group j ,
i_1 = the number of Y_{ij} observations such that $(Y_{ij} - \hat{M}_j) <$ - 2.24(MAD_{n_j}), and
i_2 = the number of Y_{ij} observations such tha $(Y_{ij} - \hat{M}_j) >$ 2.24(MAD_{n_j}).

2.1.2 Criterion for choosing the sample values

From Equation 2.4 the criterion used to determine the number of extreme observations in each group j, centers around the indices i_1 and i_2, where i_1 and i_2 are the number of extreme observations in the left- and right-tail, respectively. For a sample with no extreme value, wherein $i_1 = i_2 = 0$, $\hat{\theta}_M$ is equal to the mean for the jth group. After eliminating the extreme values, calculate $\hat{\theta}_{M_j}$ and proceed with the calculation of the H statistic.

The next section will briefly outline the scale estimators that were substituted for the default scale estimator, MAD_n.

2.2 Scale estimators

In searching for measures of scale, the breakdown value is of considerable practical importance as it constitutes one of the components in measuring robustness (Wilcox, 1997). The three scale estimators mentioned in this paper have the optimum breakdown value of 0.5. These scale estimators possess explicit formulae guaranteeing the uniqueness of the estimates. Moreover, they also contain bounded influence functions, a vital component of robust estimators. Another advantage of these estimators is their simplicity, making them easy to compute.

For the following sections, let $X = (x_1, x_2, \ldots, x_n)$ be a random sample from any distribution and let the sample median be denoted as $med_i x_i$.

2.2.1 S_n

Rousseeuw and Croux (1993) suggested alternatives to MAD_n that can be used as initial or ancillary scale estimates that are more efficient and as well are not slanted towards symmetric distributions. One such estimator is S_n, defined as

$$S_n = c\ med_i \left\{ med_j \left| x_i - x_j \right| \right\} . \qquad (2.5)$$

This estimator is similar to MAD_n; the only difference between the two is that the med_j operation is transferred to the outside of the absolute value. This makes S_n a location free estimator. Instead of measuring the deviation of observations from a

central value, S_n looks at a typical distance between observations. Another advantage is its explicit formula which means that this estimator is always uniquely defined. A modest simulation study by Rousseeuw and Croux (1993) found that a correction factor, c = 1.1926, succeeded in making S_n unbiased for finite samples. They also proved that S_n has the highest possible breakdown point. In terms of efficiency, S_n was proven to be more efficient (58.23%) than MAD_n (36.74%) with Gaussian distributions.

2.2.2 T_n

Another promising scale estimator proposed by Rousseeuw and Croux (1993) which possesses attractive robust properties is T_{n} , defined as

$$T_n = 1.38\frac{1}{h}\sum\left\{\underset{j\neq i}{med}\left|x_i - x_j\right|\right\}_{(k)} \tag{2.6}$$

It has been proven that T_n has a 50% breakdown point, a continuous influence function, and an efficiency of 52%, thus making it a better scale estimator than MAD_n.

2.3 Bootstrap method

Since the sampling distribution of *MOM-H* is unknown, the *p*-values were obtained by means of the percentile bootstrap method (See, e.g. Efron and Tibshirani, 1993). The bootstrap method is known to give a better approximation than one based on the normal approximation theory and is a suitable method especially when the samples are of moderate size (Babu, Padmanabhan, and Puri, 1999). Keselman, Wilcox, Othman, and Fradette (2002) indicated that Type I error control could be improved by combining bootstrap methods with methods based on certain robust location measures. The basic idea of bootstrapping is that in the absence of any other information about a population, the values in a random sample are the best guide to the distribution, and resampling the sample is the best guide of what can be expected if the population is resampled. To obtain the *p*-value using the percentile bootstrap method, the following steps are followed (See Wilcox, 1997):

(1) Based on the available data, calculate the *MOM-H* statistic.

(2) Randomly sample (with replacement), $\mathrm{b} = 1,\ldots,\mathrm{B}$ bootstrap samples from the data.

(3) Each of the sample points in the bootstrapped groups must be centered at their respective estimated *MOM*s (i.e., $C_{ij}^{*} = Y_{ij}^{*} - \hat{\theta}_{M\,j}$).

(4) Let *MOM-H** (denoted as MH^*) be the value of *MOM-H*, when applied to the C_{ij}^* values.

(5) Repeat Step 2 to Step 4 B times yielding $MH_1^*, MH_2^*, ..., MH_B^*$.

(6) Calculate the *p*-value as (# of $MH_B^* > MOM\text{-}H$)/B.

These calculated *p*-values represent the empirical Type I error rates for the procedures investigated under the *MOM-H* statistic.

3 Procedures and empirical investigatiosns

Three tests for location equality (comparing the typical score across groups) were compared for their sensitivity to the effects of nonnormality and variance heterogeneity in an independent groups design comprising two or four groups. The three procedures that we investigated were:

(1) *MOM-H* with MAD_n

(2) *MOM-H* with S_n

(3) *MOM-H* with T_n.

In the remainder of this paper, each of these methods will be referred to by its respective scale estimator, MAD_n, S_n, and T_n.

In studying the robustness of these procedures, four variables were manipulated, creating conditions which are known to highlight the strengths and weaknesses of tests for the equality of location parameters. The four variables were: (1) number of groups, (2) population distribution, (3) degree of variance heterogeneity, and (4) pairing of unequal variances and group sizes.

Unequal group sizes, when paired with unequal variances, can affect Type I error control for tests that compare the typical score across groups. The total sample sizes and the group sizes for the case of two and four groups were $N = 40$ (15, 25) and $N = 80$ (10, 15, 25, 30), respectively.

Three distributions representing three levels of skewness (zero, mild and extreme) were investigated. The standard normal distribution represents a distribution with zero skewness. For nonnormal distributions, we chose the chi-squared distribution with three degrees of freedom (χ_3^2) to represent mild skewness and the *g*-and-*h* distribution (Hoaglin, 1985) with $g = 0.5$ and $h = 0.5$ to represent extreme skewness. The skewness and kurtosis values for the χ_3^2 distribution are $\gamma_1 = 1.63$ and $\gamma_2 = 4.00$, respectively (Othman et al., 2004). The theoretical values for skewness and kurtosis of the $g = 0.5$ and $h = 0.5$ distribution are $\gamma_1 = \gamma_2 =$ undefined. The purpose of choosing these extreme values is based on the premise that if a method performs well under large departures from normality,

then it offers some reassurance that it will perform well for distributions encountered in practice.

In terms of variance heterogeneity, the largest and smallest variances differed by a 36:1 ratio. Though this ratio may seem extreme, similar and larger ratios have been reported in the literature (Keselman, Wilcox et al., 2004). Keselman et al. (1998), as cited by Keselman, Othman et al. (2004), noted that in a review of articles published in prominent education and psychology journals, ratios as large as 24:1 and 29:1 in one-way and factorial completely randomized designs were observed. Wilcox (2003) cited data sets where the ratio was 17,977:1! Thus although the ratio of 36:1 may appear to be large, it still seems to be a reasonable "potentially" extreme condition under which the efficacy of the tests should be examined.

As indicated, unequal group sizes, when paired with unequal variances, can affect Type I error control for tests that compare the typical score across groups (Keselman et al., 2002; Keselman, et al., 1998; Othman et al., 2004). Therefore, we positively and negatively paired the sample sizes and variances. A positive pairing occurs when the largest group size is associated with the largest group variance, while the smallest group size is associated with the smallest group variance. On the other hand, in a negative pairing, the largest group size is paired with the smallest group variance and the smallest group size is paired with the largest group variance. Positive and negative pairings typically produce conservative and liberal results, respectively, for tests that compare measures of central tendency across groups.

This study was based on simulated data. In terms of data generation, we used the SAS generator RANNOR (SAS Institute, 1999) to obtain pseudo-random standard normal variates. To generate the chi-squared variates with three degrees of freedom, three standard normal variates were generated and then squared and summed.

Observations from a *g*-and-*h* distribution were generated by converting standard normal variables to random variables utilizing the following equation:

$$Y_{ij} \frac{\exp(gZ_{ij}) - 1}{g} \exp(hZ_{ij}^{\ 2} / 2) \qquad (3.1)$$

based on the values of *g*-and-*h* selected for investigation. Specifically, setting $g = 0$, and $h = 0$, yields the standard normal distribution. The case $g = 0$ corresponds to symmetric distributions, and the tails of the *g*-and-*h* distribution get heavier as *h* increases, while the distribution becomes more skewed as *g* increases.

The design specifications are shown in the following tables.

Table 1: Design Specification for Two Groups.

PAIRING	GROUP SIZES		POPULATION VARIANCES	
	1	**2**	**1**	**2**
POSITIVE	10	15	1	36
NEGATIVE	10	15	36	1

Table 2: Design Specification for Four Groups.

PAIRING	GROUP SIZES				POPULATION VARIANCES			
	1	2	3	4	1	2	3	4
POSITIVE	10	15	20	25	1	1	1	36
NEGATIVE	10	15	20	25	36	1	1	1

In examining the Type I error rates the group location measures were set to zero. For each condition examined, 5000 data sets were generated and within each data set, 599 bootstrap samples were obtained. The nominal level of significance was set at α = 0.05.

4 Results

According to Bradley's (1978) liberal criterion of robustness, a test can be considered robust if its empirical rate of Type I error, $\hat{\alpha}$, is within the interval $0.5\alpha \le \hat{\alpha} \le 1.5\alpha$. Thus, if the nominal level is $\alpha = 0.05$, the empirical Type I error rate should be in the interval $.025 \le \hat{\alpha} \le .075$. Based on this criterion of robustness, our preliminary analyses of the empirical values indicated some of the procedures we investigated were remarkably robust in the presence of heterogeneous and nonnormal data. Table 3 contains Type I error rates when we examined two groups, while Table 4 contains the $J = 4$ rates for MAD_n, S_n, and T_n. Most noteworthy is that all of the empirical values in both tables were within Bradley's (1978) interval; therefore, according to this criterion all the methods should be regarded as robust.

However, in order to tease out differences between the procedures, we then adopted a more stringent criterion of robustness. In particular, the more stringent criterion considers a procedure to be robust if its empirical estimate of error is within the interval (.045 -.055). Values outside this interval are in boldface type in the tables.

4.1 Two groups case

The reader can note that by referring to the grand average (last row) of Table 3, which is the overall average for each procedure across distributions, all of the values fell within the stringent criterion of robustness; furthermore, the S_n procedure had a value (.0488) that was closest the nominal value. Setting aside the results from the most extreme distribution (g = 0.5 and h = 0.5), the S_n procedure still produced an averaged value (.0556) closest to the nominal value compared to MAD_n (.0559) and T_n (.0608).

It should also be noted that the average values vary with the distributions investigated. Specifically, (1) for the symmetric distribution, every value fell within the (.045-.055) interval, (2) when data were chi-squared distributed, rates were somewhat larger though all in the robustness interval, and (3) when data were *g*-and-*h* distributed rates ranged from .0324 to .0370.

Table 3: Type I Error Rates.

Distribution	Pairing	*MOM-H* with corresponding scale estimators		
		MAD_n	S_n	T_n
Normal	+ve	.0496	.0510	.0502
	-ve	.0470	.0440	.0482
	Average	*.0483*	*.0475*	*.0492*
$\chi^2_{(3)}$	+ve	**.0626**	**.0624**	**.0718**
	-ve	**.0642**	**.0648**	**.0732**
	Average	***.0634***	***.0636***	***.0725***
g-and-*h*	+ve	**.0328**	**.0370**	**.0354**
	-ve	**.0324**	**.0334**	**.0328**
	Average	***.0326***	***.0352***	***.0341***
Grand Average		*.0481*	*.0488*	*.0519*

With regard to pairings of group sizes and variances, remember that Othman et al. (2004) found that positive pairings produced conservative values, while negative pairings generated liberal values. In contrast, the procedures we examined resulted in higher empirical estimates of error for positive pairings when data were either normal or *g*-and-*h* distributed. Thus, the current results are not in accord with those of Othman et al. (2004). Only when examining the procedures with chi-squared data resulted in deflated empirical estimates for positive pairings.

4.2 Four groups case

The $J = 4$ empirical Type I error values are contained in Table 4. The grand average values are similar and close to the nominal .05 value, with T_n exhibiting the smallest discrepancy between empirical and nominal values, (i.e., .0496 vs. .05).

Across distributional shapes, there were variations in the procedures that could be designated as “best”. However, when the results from the most extreme distribution (g = 0.5 and h = 0.5) were set aside we found that the S_n procedure produced an average value (.0576) closest to the nominal value compared to MAD_n (.0578) and T_n (.0593).

Table 4: Type I Error Rates.

Distribution	Pairing	*MOM-H* with corresponding scale estimators		
		MAD_n	S_n	T_n
Normal	+ve	.0486	.0478	.0486
	-ve	.0520	.0540	.0542
	Average	*.0503*	*.0509*	*.0514*
$\chi^2_{(3)}$	+ve	**.0646**	**.0642**	**.0694**
	-ve	**.0660**	**.0642**	**.0650**
	Average	***.0653***	***.0642***	***.0672***
g-and-*h*	+ve	**.0292**	**.0268**	**.0286**
	-ve	**.0286**	**.0308**	**.0316**
	Average	***.0289***	***.0288***	***.0301***
Grand Average		*.0482*	*.0480*	*.0496*

Specifically, when data were normal in shape, the empirical *p*-values for all procedures were well controlled, with MAD_n (.0503) emerging as the best procedure. On the other hand, when data were chi-squared distributed, the best results belonged to the S_n procedure (.0642), while for *g*-and-*h* distributed data, the T_n procedure emerged as best (i.e., .0301). As well, we note that across distributions, the empirical estimates of Type I error for the chi-squared and the *g*-and-*h* distributions inclined towards liberal and conservative values respectively.

Lastly, we note that the empirical *p*-values obtained from all procedures tested under the symmetric distributions were concordant with the findings reported by Othman et al. (2004) with respect to pairings of group sizes and variances. However, for the skewed distributions that we investigated, mixed results were obtained. In particular, when data were chi-squared distributed, MAD_n resulted in higher empirical *p*-values and T_n resulted in smaller empirical *p*-values for negative pairings, while the S_n procedure resulted in equivalent values for both pairings. In contrast, when data were *g*-and-*h* distributed, when the pairing was negative, MAD_n resulted in a lower Type I error estimate, while the T_n and S_n procedures resulted in higher rates of error.

5 Conclusion

In this paper, we investigated the Type I error rates of three methods that can used to compare measures of the "typical score" across independent groups when data are neither normal nor homoscedastic. The procedures that were compared differed according to the estimate of scale that was incorporated as the trimming criterion for the *MOM-H* statistic originally described by Wilcox and Keselman (2003). The *MOM-H* statistic empirically determines the amount of data, if any, that should be trimmed from each tail of the empirical distribution. Thus, as noted by Othman et

al. (2004), *MOM-H* can be most reliable for examining differences between groups when data are nonnormal. In our investigation, we adopted *MOM-H* with the robust scale estimators suggested by Rousseuw and Croux, (1993).

Our results indicated that all three procedures were robust with respect to Type I error control even though data were nonnormal and heterogeneous; that is, all rates were within Bradley's stringent interval criterion (i.e., within the interval .045-.055). The minute variabilities between the empirical *p*-values across the procedures indicate that the procedures were on par with one another. However, when averaging rates of Type I error across the conditions investigated, the rank order of the tests with respect to the deviation of empirical from nominal rates, indicate that *MOM-H* with T_n was best, thus making it the procedure we recommend to researchers.

Acknowledgement

The authors would like to acknowledge the work that led to this paper is partially funded by the Fundamental Research Grant Scheme of the Universiti Sains Malaysia and the Social Sciences and Humanities Research Council of Canada.

References

[1] Babu, G.J., Padmanabhan, A.R., and Puri, M.L. (1999): Robust one-way ANOVA under possibly non-regular conditions. *Biometrical Journal*, **41**, 321-339.

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

[3] Efron, B. and Tibshirani, R.J. (1993): *An Introduction to the Bootstrap*. Chapman & Hall Inc.

[4] Hoaglin, D.C. (1985): Summarizing shape numerically: The *g*-and-*h* distributions. In D. Hoaglin, F.Mosteller, and J. Tukey (Eds.): *Exploring Data Tables, Trends, and Shapes*. New York: Wiley.

[5] Lix, L.M. and Keselman, H.J., (1998): To trim or not to trim: Tests of location equality under heteroscedasticity and non-normality. *Educational and Psychological Measurement*, **58**, 409-429

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

[7] Keselman, H.J., Huberty, C.J., Lix, L.M., Olejnik, S., Cribbie, R.A., Donahue, B., Kowalchuk, R.K., Lowman, L.L., Petoskey, M.D., Keselman, J.C., and 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

[8] Keselman, H.J., Kowalchuk, R.K., Algina, J., Lix, L.M., and 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.

[9] Keselman, H.J., Othman, A.R., Wilcox, R.R., and Fradette, K. (2004): The new and improved two-sample t-test. *Psychological Science*, **15**(1), 57–51.

[10] Keselman, H.J., Wilcox, R.R., Algina, J., Fradette, K., and Othman, A.R. (2004): A power comparison of robust test statistics based on adaptive estimators. *Journal of Modern Applied Statistical Methods*, **3**, 27-38.

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

[12] Kulinskaya, E., Staudte, R.G., and Gao, H. (2003): Power approximations in testing for unequal means in a one-way ANOVA weighted for unequal variances. *Communications in Statistics - Theory and Methods*, **32,** 2353-2371.

[13] Othman, A.R., Keselman, H.J., Padmanabhan, A.R., Wilcox, R.R., and Fradette, K. (2004): Comparing measures of the "typical" score across treatment groups. *British Journal of Mathematical and Statistical Psychology*, **57**, 215-234.

[14] Rousseeuw, P.J. and Croux, C. (1993): Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, **88**, 1273-1283.

[15] SAS Institute Inc. (1999): *SAS/IML User's Guide version 8*. Cary. NC: SAS Institute Inc.

[16] Schrader, R.M. and Hettmansperger, T.P. (1980): Robust analysis of variance. *Biometrika*, **67**, 93-101.

[17] Syed Yahaya, S.S., Othman, A.R., and Keselman, H.J. (2004): Testing the equality of location parameters for skewed distributions using S1 with high breakdown robust scale estimators. In M.Hubert, G. Pison, A. Struyf and S. Van Aelst (Eds.): *Theory and Applications of Recent Robust Methods, Series: Statistics for Industry and Technology*. Basel: Birkhauser, 319 – 328.

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

[19] Wilcox, R.R. (1994): A one-way random effects model for trimmed means. *Psychometrika*, **59** (3), 289-306

[20] Wilcox, R.R. (1997): *Introduction to Robust Estimation and Hypothesis Testing*. New York : Academic Press.

[21] Wilcox, R.R. (2003): *Applying Contemporary Statistical Techniques*. New York : Academic Press.

[22] Wilcox, R.R., Charlin, V.L., and Thompson, K.L. (1986): New Monte Carlo Results on the Robustness of the ANOVA F, W and F* Statistics. *Communications in Statistics-Simulations*, **15**, 933-943.

[23] Wilcox, R.R. and Keselman, H.J. (2002): Power Analyses when comparing trimmed means. *Journal of Modern Applied Statistical Methods*, **1**, 24-31.

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