# THE DUAL FUNCTION OF TWO HIGHLY ROBUST ESTIMATORS OF SCALE.

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

[1]Faculty of Quantitative Sciences, Universiti Utara Malaysia, 06010 Sintok, Kedah, Malaysia; sharipah@uum.edu.my

[2]School of Distance Education, 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@Ms.UManitoba.CA

Two statistics for testing the equality of central tendency measures under conditions of variance heterogeneity and non-normality, i.e. $S_1$ and MOM-$H$, were compared in the context of a one-way completely randomized design. Both statistics were matched with two highly robust scale estimators: $MAD_n$ and $T_n$. For the $S_1$ statistic, $MAD_n$ and $T_n$ were used as scale estimators. For MOM-$H$, they were used as the trimming criterion for the modified one-step $M$-estimator (MOM), the central tendency measure for the MOM-$H$ statistic. The tests proved to be robust to our conditions of variance heterogeneity and non-normality.

Keywords: central tendency measures, robust scale estimators, Type I error, variance heterogeneity, non-normality, trimming

AMS 2000 Classification: 62K25

## 1.0 Introduction

In comparing measures of location for two or more groups, the classical methods, i.e., Student's two-sample $t$-test and the ANOVA $F$-test, are among the most commonly used statistical methods in the one-way independent groups design. However, these methods are adversely affected by non-normality, particularly when variances are heterogenous and group sizes are unequal (Lix and Keselman, 1998). Violating the assumptions associated with these methods will cause their Type I error and power rates to be distorted. The Type I error rates will be inflated (liberal) from the nominal value, resulting in spurious rejections, and power rates can be substantially reduced from the

theoretical value, resulting in true effects being missed. Even though it is well established that the ANOVA (and *t*) is (are) not robust if the homogeneity assumption does not hold (Wilcox et al., 1986), it (they) is (are) often used by applied researchers even when the data suggest that population variances are unequal (Kulinskaya et al., 2003). It is also well known that a slight departure from normality can have a substantial effect on power for these methods (Sawilowsky and Blair, 1992; Wilcox, 1995).

In an effort to overcome the deficiencies associated with the *t* and ANOVA *F*-tests, researchers continue to seek alternative methods. For example, Cochran (1937) suggested weighting the terms in the sum of squares explained by the respective inverses of the sample variances, and he provided a chi-squared test for equal means based on a transformation of the ANOVA *F*-test. However, the design has to be balanced (i.e., equal sample sizes per group). For unbalanced designs, James (1951) and Welch (1951) had suggested weighting the terms in the sum of squares explained by estimates of the inverses of the variances of the respective sample means. This weighted sum of squares has an approximate chi-squared distribution under the null hypothesis of equal population means for large sample sizes. Even if the problem of unequal variances could be overcome, non-normality may still be problematic for the classical and some alternative methods. The purpose of our paper therefore was to examine other methods that may be adopted to compare measures of the typical score across independent groups when data are neither normal in form nor variances equal across groups.

## 2.0 Test Statistics

In searching for an alternative approach in testing central tendency measures in the one-way independent groups design, we suggested two robust procedures. The first

was the $S_1$ statistic by itself, not in the adaptive manner as originally proposed by Babu et al. (1999). The second was the MOM-*H* statistic proposed by Othman et al. (2004). These statistics were combined with selected robust scale estimators. Based on the proposed robust scale estimators by Rousseeuw and Croux (1993), one of the scale estimators we used, $T_n$, has the highest breakdown point and a bounded influence function. Additionally, we were also interested in $MAD_n$, one of the most popular robust scale estimators, based on its robustness. We incorporated these scale estimators with $S_1$ and found that the method provided good Type I error control when data were generated from moderately skewed distributions (Syed Yahaya et al., 2004a; 2004b). Babu et al. (1999) proposed using the $S_1$ statistic in order to overcome non-normality and variance heterogeneity. With this statistic one does not need to trim data.

Another way of dealing with skewed data is by trimming data from the tails of a distribution. Working with actual data, Wilcox et al. (2000) found that power can be greatly increased and control over the probability of a Type I error can be better by comparing trimmed means versus means. However, there are practical concerns regarding trimming and accordingly with comparing trimmed means rather than the usual least squares means, as is the case with ANOVA. One issue is that by assumption, the amount of trimming is fixed prior to analyzing the data. Another concern is that trimming is typically assumed to be symmetric. Given these concerns, researchers are faced with the issue of determining the best percentage of trimming and from which tail(s) should the trimming occur. For example, if sampling is from a light-tailed distribution or normal distribution, it might be desirable to trim very few observations or perform no trimming

at all. If the distribution is skewed, a natural reaction is to trim more observations from the skewed side of the empirical distribution.

Wilcox et al. (2000) suggested modifying the one-step *M*-estimator (MOM) as a means of addressing the issue of how much of the data should be trimmed and from which tails, if any, the data should be trimmed. This central tendency estimator, like a trimmed mean, can be applied to a test statistic in order to investigate the equality of central tendency measures across treatment groups (Keselman et al., 2002; Othman et al., 2004). By using a statistic mentioned by Schrader and Hettsmansperger (1980), examined by He et al. (1990), and discussed by Wilcox (1997), Othman et al. (2004) proposed a method known as MOM-*H* which uses MOM as the central tendency measure.

## 2.1 $S_1$ Statistic

Consider the problem of comparing location parameters for skewed distributions. Let $Y_{ij} = (Y_{1j}, Y_{2j}, ..., Y_{n_j j})$ be a sample from an unknown distribution $F_j$ and let $M_j$ be the population median ($F_j$: $j$ = 1, 2, ... , $J$). To test $H_0$: $M_1 = M_2 = \ldots = M_J$ versus $H_1: M_i \neq M_j$ for at least one pair of $(i, j)$, the $S_1$ statistic is defined as

$$S_1 = \sum_{1 \le i < j \le J} |s_{ij}|, \qquad [2.1]$$

where

$$s_{ij} = \frac{(\hat{M}_i - \hat{M}_j)}{\sqrt{(\hat{\omega}_i + \hat{\omega}_j)}}, \qquad [2.2]$$

$\hat{M}_j$ = the median of group $j$,

$\hat{\omega}_j = \dfrac{\omega_j}{n_j}$,

$n_j$ = number of observations for group $j$, and

$$\omega_j = \left( \frac{1}{n_j} \sum_{j=1}^{J} \sum_{i=1}^{n_j} \left| Y_{ij} - \hat{M}_j \right| \right)^2 . \qquad [2.3]$$

$S_1$ is the sum of all possible differences of sample medians from the $J$ distributions divided by their respective sample standard error, $\hat{\omega}$. Therefore, if there are $J$ distributions, the number of possible differences equals $J(J-1)/2$.

When dealing with skewed distributions, the parameter of interest is therefore the population median. As stated in the formula, $S_1$ uses the median as its central measure. Therefore, in the case of skewed distributions, $S_1$ seems to be a suitable procedure for comparing the typical score (median) across independent groups.

## 2.2 MOM-*H* Statistic

MOM-*H* is a procedure that uses a statistic originally proposed by Schrader and Hettmansperger (1980) known as the *H* test. The test is defined as

$$H = \frac{1}{N} \sum_{j=1}^{J} n_j (\hat{\theta}_j - \hat{\theta}.)^2 , \qquad [2.4]$$

where $N = \sum_j n_j$ and

$$\hat{\theta}. = \sum_j \hat{\theta}_j / J .$$

This statistic is readily adaptable to any measure of central tendency but not recommended for means or trimmed means (Wilcox, 1997).

Othman et al. (2004) used the *H* statistic when comparing the typical score across treatment groups. However, they modified this statistic by replacing $\hat{\theta}$ with the MOM (denoted as $\hat{\theta}_M$). The modified test statistic is known as MOM-*H*, and the goal of this statistic is to test $H_0 : \theta_{M1} = \theta_{M2} = \ldots = \theta_{M_J}$ versus $H_1 : \theta_{Mi} \neq \theta_{Mj}$, for at least one pair of (*i*,

*j*). Keselman et al. (2002) found that rates of Type I error for MOM-*H* were not affected when data were skewed.

### 2.2.1 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

$$\hat{\theta}_{Mj} = \sum_{i=i_1+1}^{n_j - i_2} \frac{Y_{(i)j}}{n_j - i_1 - i_2} \quad , \qquad [2.5]$$

where

$Y_{(i)j}$ = the *i*th ordered observation in group *j*,

$i_1$ = the number of $Y_{ij}$ observations such that $(Y_{ij} - \hat{M}_j) < -2.24(\mathrm{MAD}_{n_j})$, and

$i_2$ = the number of $Y_{ij}$ observations such that $(Y_{ij} - \hat{M}_j) > 2.24(\mathrm{MAD}_{n_j})$.

### 2.2.2 Criterion for choosing the sample values

From Equation 2.5, 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 values, wherein $i_1 = i_2 = 0$, $\hat{\theta}_M$ is equal to the mean for the *j*th group. After eliminating the extreme values, calculate $\hat{\theta}_{Mj}$ and proceed with the calculation of the *H* statistic.

The next section will briefly outline the alternative scale estimators that were used as substitutes to the default scale estimators in the previous two statistical tests.

## 2.3 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 scale estimators defined in this paper have the optimum breakdown value of 0.5. These scale estimators possess explicit formulas 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 $med_i x_i$.

### 2.3.1 $MAD_n$

A very popular and robust scale estimator is the median absolute deviation about the median, given by

$$\text{MAD}_n = b\ med_i \left| x_i - med_j x_j \right| . \qquad [2.6]$$

The constant $b$ in the formula is needed to make the estimator consistent for the parameter of interest. For example, if observations are randomly sampled from a normal distribution with $b = 1$, the estimator does not estimate $\sigma$, the standard deviation, instead, it estimates $0.6745\sigma$. To posit $\text{MAD}_n$ in a more familiar context, it is typically rescaled so that it estimates $\sigma$ when sampling from a normal distribution. In this case, set $b = 1.4826$.

$\text{MAD}_n$ is simple and easy to compute and its extreme sturdiness makes it ideal for screening the data for extreme values in a quick way by computing

$$\frac{\left|x_i - med_j x_j\right|}{\text{MAD}_n} > K \qquad [2.7]$$

for each $x_i$ and flagging those $x_i$ as extreme when the statistic exceeds a certain cut off point. $\text{MAD}_n$ has the best possible breakdown point, and its influence function is bounded (Rousseeuw and Croux, 1993). Huber (1981) identified $\text{MAD}_n$ as the single most useful ancillary estimate of scale due to its high breakdown property.

### 2.3.2 $T_n$

Another promising scale estimator proposed by Rousseeuw and Croux (1993) which possesses the attractive properties of a robust scale estimator is $T_n$, defined as

$$T_n = 1.3800 \frac{1}{h} \sum_{k=1}^{h} \left\{ med_{j \neq i} \left| x_i - x_j \right| \right\}_{(k)} . \qquad [2.8]$$

It was proven that $T_n$ has the highest breakdown point (50%), a continuous influence function, and an efficiency of 52%, which makes it more efficient than $\text{MAD}_n$. This estimator has a simple and explicit formula which guarantees uniqueness and it is suitable for asymmetric distributions.

Taking into consideration all the attractive properties attached to the scale estimators, such as the breakdown point, continuous influence function, and efficiency, it was therefore decided that these estimators would be used as the scale estimator for $S_1$ and as a criterion for choosing the sample values for MOM-*H*.

## 2.3 Bootstrap Method

Since the sampling distributions of $S_1$ and MOM-*H* are unknown, Babu et al. (1999), followed by Othman et al. (2004), used the bootstrap percentile method for obtaining the *p*-values in their studies. According to Babu et al. (1999), the bootstrap method is known

to give a better approximation than the one based on the normal approximation theory, and this method is attractive, especially when samples are of moderate. Taking into consideration the intractability of the sampling distributions of $S_1$ and MOM-$H$, and the reliability of the bootstrap method, the percentile bootstrap method (see, e.g. Efron and Tibshirani, 1993) was utilized to assess the statistical significance ($p$-value) of each procedure in our study.

## 3.0 Procedures and Empirical Investigations

$S_1$ and MOM-$H$, were combined with the two highly robust scale estimators. Specifically, the following methods were investigated:

1. $S_1$ with $\text{MAD}_n$,
2. $S_1$ with $T_n$,
3. MOM-$H$ with $T_n$,

and the default methods for $S_1$ and MOM-$H$, that is,

4. $S_1$ with $\hat{\omega}$ and
5. MOM-$H$ with $\text{MAD}_n$.

In the remainder of this paper, each of these procedures will be referred to by its respective scale estimator, $\text{MAD}_n$ and $T_n$.

Four variables were manipulated: (1) number of groups, (2) population distribution, (3) degree of variance heterogeneity, and (4) pairing of unequal variances and group sizes.

Investigations were done on completely randomized designs containing two ($J$ = 2) and four unbalanced groups ($J$ = 4) since previous research had utilized similar designs (Lix and Keselman, 1998; Othman et al., 2004; Yuen, 1974). By analyzing these two

cases, we are able to compare the effect of the number of groups on the Type I error rates of the investigated procedures.

In investigating the effects of distributional shape on rates of Type I error, three distributions representing different levels of skewness were considered. The standard normal distribution represents a distribution with zero skewness. In addition, two non-normal distributions were also analyzed. They were the chi-squared distribution with three degrees of freedom and the *g*-and-*h* distribution with $g = 0.5$ and $h = 0.5$. We chose the chi-squared distribution ($\chi_3^2$) to represent mild skewness and the *g*-and-*h* distribution (Hoaglin, 1985) to represent extreme skewness. The skewness and kurtosis values for the $\chi_3^2$ distribution are $\gamma_1 = 1.63$ and $\gamma_2 = 4.00$, respectively (Othman et al., 2004). On the other hand, the respective theoretical values for skewness and kurtosis of the *g*-and-*h* distribution are $\gamma_1 = \gamma_2 =$ undefined. The purpose of selecting the extreme g and h distribution is based on the assumption that if a method performs well under seemingly large departures from normality, then it can be safely assumed that the method will also perform well for distributions of lesser skewness.

Variance heterogeneity is one of the two major problems researchers typically encounter when testing the equality of location measures. The classical ANOVA *F*-test (as well as the two-sample *t*-statistic) is (are) known to yield misleading results when there exist different population variances (Kulinskaya et al., 2003). To investigate the effect of variance heterogeneity on Type I error rates, variances with a 1:36 ($J = 2$) or 1:1:1:36 ($J = 4$) ratio were assigned to the groups. Though this ratio may seem large, larger ratios have been reported in the literature (Keselman, Wilcox et al., 2004).

When unequal variances were paired with unequal sample sizes, negative and positive pairings were formed. A positive pairing involves pairing the largest number of observations with the largest variance and the smallest number of observations with the smallest group variance. For the negative pairing case, the largest group of observations was paired with the smallest group variance, while the smallest group of observations was paired with the largest group variance. It should be noted that pairings of sample sizes and variances do have an effect on Type I error rates (Keselman et al., 1998; Keselman, Othman et al., 2004; Othman et al., 2004). Positive and negative pairings typically produce conservative and liberal results, respectively (Othman et al., 2004).

The programs and simulations were run using the SAS/IML language. The random samples were generated using the SAS generator RANNOR (SAS Institute, 1999). The variates transformed to $\chi_3^2$ and *g*-and-*h* variates and then standardized. In examining the Type I error rates, the group location measures were set to zero and group variances factored into the variates. 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 $\alpha = .05$.

## 4.0 Results

To evaluate the particular conditions under which a test is insensitive to assumption violations, Bradley's (1978) liberal criterion of robustness was employed. According to this criterion, a test is considered robust if its empirical rate of Type I error ($\hat{\alpha}$) lies in the $0.5\alpha \leq \hat{\alpha} \leq 1.5\alpha$ interval. Therefore, for the 5% level of statistical significance used in this study, a test was considered robust, in a particular condition, if its empirical Type I error fell within the $(.025 - .075)$ interval.

## 4.1 $S_1$ Procedures for $J = 2$

The empirical Type I error rates are contained in Table 4.1. The Grand Average in the last row of the table, which represents the overall performance of a procedure, shows that

Table 4.1: Type I Error Rates ($J = 2$; unbalanced design)

| Distribution | Pairing | $S_1$ with corresponding scale estimators | | |
|---|---|---|---|---|
| | | $MAD_n$ | $T_n$ | $\hat{\omega}$ |
| | | | | |
| Normal | +ve | .0524 | .0534 | .0448 |
| | -ve | .0564 | .0532 | .0422 |
| | Average | *.0544* | *.0533* | *.0435* |
| Chi-sq | +ve | .0408 | .0406 | .0426 |
| | -ve | .0440 | .0374 | .0390 |
| | Average | *.0424* | *.0390* | *.0408* |
| *g*-and-*h* | +ve | .0294 | .0286 | .0192 |
| | -ve | .0314 | .0242 | .0156 |
| | Average | *.0304* | *.0264* | *.0174* |
| Grand Average | | *.0424* | *.0396* | *.0339* |

the corresponding *p*-values satisfy Bradley's (.025 to .075) interval for robustness. Therefore, all the procedures were found to be robust. The average value for the $MAD_n$ (.0424) procedure was the closest to the nominal level. In comparing the original $S_1$ (i.e. with $\hat{\omega}$) against the modified $S_1$ procedures, it clearly can be seen that the $MAD_n$ procedure performed exceptionally well across the three distributions, producing Type I error rates which were closest to the .05 nominal value. With regard to variance and group size pairings, except for $MAD_n$, the other procedures generated higher *p*-values for positive pairs across the three distributions investigated.

## 4.2 $S_1$ Procedures for $J = 4$

Table 4.2 contains the Type I error values for the four groups unbalanced design. One can observe from the table that none of the grand average $p$-values were within the (.025-.075) interval; that is, none of the procedures can be considered robust under Bradley's criterion of robustness.

Table 4.2: Type I Error Rates ($J = 4$; unbalanced design)

| Distribution | Pairing | $S_1$ with corresponding scale estimators | | |
|---|---|---|---|---|
| | | $MAD_n$ | $T_n$ | $\hat{\omega}$ |
| | | | | |
| Normal | +ve | .0248 | .0244 | .0278 |
| | -ve | .0268 | .0260 | .0302 |
| | Average | *.0258* | *.0252* | *.0290* |
| Chi-square | +ve | .0236 | .0264 | .0246 |
| | -ve | .0324 | .0330 | .0278 |
| | Average | *.0280* | *.0297* | *.0262* |
| *g*-and-*h* | +ve | .0188 | .0174 | .0078 |
| | -ve | .0206 | .0194 | .0102 |
| | Average | *.0197* | *.0184* | *.0090* |
| Grand Average | | *.0245* | *.0244* | *.0214* |

However, if we compare these values with the values of the original $S_1$ ($\hat{\omega}$) procedure, the modified $S_1$ procedures using scale estimators $MAD_n$ and $T_n$ generated better empirical Type I error rates. Although, the overall performance of the modified procedures proves to be better than the original $S_1$ procedure, the $p$-values of .0245 and .0244 for $MAD_n$ and $T_n$ respectively, are slightly conservative with regard to Bradley's criterion.

In general, our results indicate that the $MAD_n$ procedure was as good as the $T_n$ procedure when date were obtained from skewed distributions. Although $MAD_n$ adopts a symmetric view on dispersion, the results indicate that it functions effectively under

skewed conditions and was found to be the best procedure under extreme conditions of non-normality. In contrast, the original $S_1$ statistic (with $\hat{\omega}$) only performed well under symmetric conditions, and was the least effective procedure under extreme conditions.

## 4.3 MOM-*H* Procedures for $J = 2$

All the $J$ = 2 Type I error values contained in Table 4.3 are within the (.025-.075) interval. Accordingly, all the procedures are robust under Bradley's criterion for robustness.

Table 4.3: Type I Error Rates ($J = 2$; unbalanced design)

| Distribution | Pairing | MOM-*H* with corresponding scale estimators | |
|---|---|---|---|
| | | $MAD_n$ | $T_n$ |
| | | | |
| Normal | +ve | .0496 | .0502 |
| | -ve | .0470 | .0482 |
| | Average | *.0483* | *.0492* |
| Chi-sq | +ve | .0626 | .0718 |
| | -ve | .0642 | .0732 |
| | Average | *.0634* | *.0725* |
| *g*-and-*h* | +ve | .0328 | .0354 |
| | -ve | .0324 | .0328 |
| | Average | *.0326* | *.0341* |
| Grand Average | | *.0481* | *.0519* |

Based on the Grand Average values, the $MAD_n$ procedure generated an average value closest to the .05 nominal level, with a $p$-value of .0481. However, across distributional shapes, $T_n$ seems to be the better procedure for symmetric and extremely skewed distributions, while $MAD_n$ performed best for the mildly skewed distribution investigated. All the values obtained when sampling from the symmetric distribution were not only robust, but they also satisfied Bradley's stringent criterion of robustness

which requires the values to be within a (.045-.055) interval. For the chi-squared distribution, all the empirical Type I error rates tended to be above the nominal .05 value; nonetheless, they were still within the interval. The procedures examined under the *g*-and-*h* distribution also produced good Type I error rates, ranging from .0324 to .0370.

With regard to variance and group size pairings, the procedures, examined under the normal and *g*-and-*h* distribution, generated higher *p*-values for positive pairings. These results are not consistent with the results reported by Othman et al. (2004), where they observed that positive and negative pairings produced conservative and liberal results, respectively. Only the procedures tested under the chi-squared distribution produced lower *p*-values for the positive pairing case.

## 4.4 MOM-*H* Procedures for $J = 4$

Table 4.4 contains the empirical Type I error rates for the four groups case. Of particular significance was the finding that all *p*-values were within Bradley's (.025-.075) interval.

Table 4.4: Type I Error Rates ($J = 4$; unbalanced design)

| Distribution | Pairing | MOM-*H* with corresponding scale estimators | |
|---|---|---|---|
| | | $MAD_n$ | $T_n$ |
| | | | |
| Normal | +ve | .0486 | .0486 |
| | -ve | .0520 | .0542 |
| | Average | *.0503* | *.0514* |
| Chi-sq | +ve | .0646 | .0694 |
| | -ve | .0660 | .0650 |
| | Average | *.0653* | *.0672* |
| *g*-and-*h* | +ve | .0292 | .0286 |
| | -ve | .0286 | .0316 |
| | Average | *.0289* | *.0301* |
| Grand Average | | *.0482* | *.0496* |

The Grand Average values, are consistent and close to the nominal level of .05, with the $T_n$ procedure being closest, i.e., with a $p$-value of .0496. Across distributional shapes, there was variation in which procedure could be described as best. When data were obtained from the normal distribution, the $p$-values for the procedures were well controlled, with $MAD_n$ (.0503) emerging as the best procedure. Likewise, for the chi-square distribution, $MAD_n$ provided the best control with an empirical Type I error rate of .0653. For the $g$-and-$h$ distribution, $T_n$ provided better control (i.e., .0301). In addition, we found that the $p$-values for the chi-squared and the $g$-and-$h$ distributions tended generally to be liberal and conservative, respectively.

For the variance and sample size pairings, it should be noted that the $p$-values obtained from all the procedures examined under the symmetric distributions are in agreement with the findings from Othman et al. (2004). However, for the skewed distributions, we obtained mixed results. For example, for chi-squared distributed data, $MAD_n$ resulted in higher $p$-values, while $T_n$ resulted in lower $p$-values for negative pairings. In contrast, when data were $g$-and-$h$ distributed, and the pairing was negative, $MAD_n$ resulted in lower $p$-values, while larger p-values were found for the $T_n$ procedures.

## 5.0 Conclusion

In our investigation we compared a number of procedures which can be used to compare the typical score across independent groups of subjects when data are non-normal and variances are unequal. Two statistics, $S_1$ and MOM-$H$, exhibited good control of their Type I error rates; however, MOM-$H$ had better rates of error for both designs investigated ($J = 2$ and $J = 4$). The MOM-$H$ procedures, when data were normal, satisfied Bradley's stringent interval of robustness criterion (.045-.055), whereas most of the $S_1$

statistics were within this interval only when $J = 2$. For chi-squared distributed data, even though all the MOM-$H$ and $S_1$ procedures were robust, the Type I error rates differed between the two statistics; MOM-$H$ empirical rates tended to towards the liberal end of the continuum while the $S_1$ values tended toward the conservative side of .05. Because we believe it is generally important to control the rate of Type I error, procedures with conservative values are preferable to those with liberal values. Therefore, under mild conditions of skewness, the $S_1$ statistics we investigated would be preferable. For data that is more substantially non-normal (i.e., skewed -- the $g$-and-$h$ distribution), MOM-$H$ performed exceptionally well, providing good Type I error rates. On the other hand, none of the $S_1$ procedures were robust for $J = 4$, and only two procedures, $MAD_n$ and $T_n$ were robust for $J = 2$ with $p$-values of .0304 and .0264, respectively. Therefore, when working with extremely skewed data, the MOM-$H$ procedures seem preferable.

Our goal was to search for some alternative methods for testing the equality of location measures when data are obtained from skewed distributions. In this final section, we would like to share some of the advances that emerged from our investigation. Modifications to the $S_1$ and MOM-$H$ statistics successfully improved the performance of the two statistics in terms of Type I error control. The original $S_1$ procedure performed well when data were obtained from a symmetric distribution; however, its rate of Type I error was not well controlled when the degree of non-normality was more extreme. On the other hand, $S_1$ with $T_n$ or $S_1$ with $MAD_n$ proved to be much more effective with regard to Type I error control and thus should be considered as viable statistics to be adopted, particularly for testing the equality of two groups.

When there are more than two groups, that is, when $J = 4$, we recommend the MOM-$H$ procedures - the best results were found for the MOM-$H$ with $T_n$. This procedure performed remarkably well for symmetric and extremely skewed distributions. However, it should be noted that, when observations are mildly skewed, the use of this procedure might result in slightly inflated Type I error rates (from the .05 level). Nevertheless, the $p$-values observed in this study showed that they are still within Bradley's definition of robustness. Therefore, when dealing with mildly skewed distributions any of the $S_1$ procedures will be a good alternative. Similarly, for $J = 2$, the MOM-$H$ procedures performed exceptionally well under symmetric and extremely skewed distributions. However, for mildly skewed distributions, we suggest any of the $S_1$ procedures, especially $S_1$ combined with the $MAD_n$ scale estimator. When researchers suspect that their data is extremely skewed, in a manner similar to the characteristics of the $g$-and-$h$ distribution ($g = 0.5$ and $h = 0.5$), then clearly, it will be advantageous to adopt one of the MOM-$H$ procedures.

It is our impression that applied researchers would prefer a method that compared treatment performance across groups with a measure for the typical score which is based on as much of the original data as possible. $S_1$ will be the best choice for this purpose. Moreover, no trimming or transforming of the data is needed when using this statistic, meaning that one can save all the information that might have been lost if trimming had been applied. However, if the data need to be trimmed, one can avoid unnecessary trimming by using one of the MOM-$H$ procedures. These procedures empirically determine whether observations, if any, should be trimmed, as well as where (which tail) the data should be trimmed from.

**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**

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.

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

Cochran, W.G. (1937). Problems arising in the analysis of a series of similar experiments. *Journal of the Royal Statistical Society.* (suppl.4): 102-118.

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

He, X., Simpsom, D.G., and Portnoy, L.S. (1990). Breakdown robustness of test. *Journal of American Statistical Association.* **85**: 446-452.

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.* Wiley, New York.

Huber, P.J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics.* **35**: 73-101.

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

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**(3): 350-386.

Keselman, H.J., Wilcox, R.R., Othman, A.R., Fradette, K. (2002). Trimming, transforming statistics, and bootstrapping: circumventing the biasing effects of heteroscedasticity and non normality. *Journal of Modern Applied Statistical Methods.* **1**(2): 288-399.

Keselman, H.J., Wilcox, R.R., Algina, J., Fradette, K., Othman, A.R. (2004). A power comparison of robust test statistics based on adaptive estimators. *Journal of Modern Applied Statistical Methods*. **3**(1): 27-38.

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

Kulinskaya, E., Staudte, R.G., 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

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**: 409-42.

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**(2): 215-234

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

SAS Institute Inc. (1999). *SAS/IML User's Guide version 8*. SAS Institute Inc, Cary, NC.

Sawilowsky, S.S. and Blair, R.C. (1992). A more realistic look at the robustness and Type II error properties of the *t*-test to departures from population normality. *Psychological Bulletin*. **111**: 352-360.

Schrader. R.M. and Hettmansperger, T.P. (1980). Robust analysis of variance based upon a likelihood ratio criterion. *Biometrika*. **67**(1): 93-101

Syed Yahaya, S.S., Othman, A.R. and Keselman, H.J. (2004a). Testing the equality of location parameters for skewed distributions using $S_1$ 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*. Birkhauser, Basel. 319-328.

Syed Yahaya, S.S., Othman, A.R., and Keselman, H.J. (2004b). An alternative approach for testing location measures in the one way independent group design. *Proceedings of the International Conference on Statistics and Mathematics and Its Applications in the Development of Science and Technology*. Bandung, Indonesia. 201-207.

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

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). *Introduction to Robust Estimation and Hypothesis Testing*. Academic Press, New York.

Wilcox, R.R., Charlin, V.L., Thompson, K.L. (1986). New Monte Carlo results on the robustness of the ANOVA F, W and F* statistics. *Communications in Statistics-Simulations*. **15**(4): 933-943.

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. and Keselman, H.J. (2003). Repeated measures one-way ANOVA based on a modified one-step *M*-estimator. *British Journal of Mathematical and Statistical Psychology*. **56**(1): 15-35.

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