Comparing Means under Heteroscedasticity and Nonnormality:

Further Exploring Robust Means Modeling

Alyssa Counsell
Department of Psychology
Ryerson University

Phil Chalmers
Department of Educational Psychology
University of Georgia

Robert A. Cribbie
Quantitative Methods Program
Department of Psychology
York University

Comparing Means under Heteroscedasticity and Nonnormality:

Further Exploring Robust Means Modeling

The traditional independent samples analysis of variance (ANOVA) is a popular statistical analysis in psychology because researchers commonly seek to examine mean differences across multiple groups. Like all parametric statistical tests, certain assumptions must be met to validly interpret the results of the traditional ANOVA, namely independence of observations, normality of population distributions, and equal population variances. While the assumption of independence is an issue at the research design level, normality and equal variance are important statistical assumptions that researchers should always examine when conducting their ANOVA.

Research suggests that the aforementioned assumptions are rarely satisfied with the types of data typically collected within psychology and other social science fields (Blanca, Arnau, Lopez-Montiel, Bono, & Bendayan, 2011; Golinski & Cribbie, 2009; Keselman et al., 1998, Micceri, 1989; Wilcox, 1990a, 1990b). In fact, researchers continue to adopt more traditional approaches despite a wealth of research highlighting that the Type I error rates and power of the ANOVA are affected by violating assumptions (e.g., Boneau, 1960; Glass, Peckham, & Sanders, 1972), despite the availability of improved methods for comparing central tendencies under these circumstances (e.g., Cribbie, Fiksenbaum, Wilcox, & Keselman, 2012; Keselman, Algina, Lix, Wilcox, 1995, 2017; Keselman, Algina, Wilcox, & Deering, 2008). Recent research further suggests that few researchers examine these assumptions at all (Hoekstra, Kiers, & Johnson, 2012).

Assumption violation may result in inaccurate interpretations of true population differences for the traditional ANOVA procedure. For example, when one population's variance is much higher than another, especially with unequal sample sizes, the empirical probability of Type I errors deviates from the nominal level (Box, 1954, Brown & Forsythe, 1974a, 1974b; Wilcox, 1988). The manner in which Type I error rates are affected depends on the pairing of sample size and variance heterogeneity. Specifically, when small sample sizes are paired with large variance (i.e., negative/inverse pairing), empirical Type I error rates will be inflated relative to the nominal α, whereas when large sample sizes are paired with

large variance (i.e., positive/direct pairing), Type I error rates tend to be too conservative. In instances where the homogeneity of variance assumption has been met, deviations from normality tend to have little effect on the Type I error rates of the traditional ANOVA, but often decrease the statistical power (Harwell, Rubinstein, Hayes, & Olds, 1992; Lix, Keselman, & Keselman, 1996). When population variances are unequal and distributions are nonnormal, empirical Type I error rates are extremely aberrant (Cribbie et al., 2012).

Trimmed Welch Test with Winsorized Variances

Due to the routine violation of assumptions in psychological research, many researchers have proposed alternatives to the omnibus ANOVA *F* test. One method that has been found to maintain accurate Type I error control and retain power under variance heterogeneity and nonnormality is to use trimmed means and Winsorized variances in combination with a test that uses a non-pooled standard error and adjusted degrees of freedom (trimmed Welch; e.g., Cribbie et al., 2012; Keselman, Kowalchuk, & Lix, 1998; Keselman, Algina, Wilcox, & Kowalchuk, 2000; Keselman et al., 2008; Wilcox, Keselman, Muska, & Cribbie, 2000). The trimmed Welch test has been found to perform well even with extremely nonnormal distributions and disparate sample sizes and variances. Details on this test statistic are given below.

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

$$\bar{X}_t = \frac{1}{h}\sum_{i=\lambda+1}^{n-\lambda} X_i \quad (1)$$

The sample Winsorized mean is:

$$\bar{X}_W = \frac{1}{n}\sum_{i}^{N} Y_i \quad (2)$$

where:

$$Y_i = \begin{cases} X_{(\lambda+1)} \ if \ X_i \le X_{(\lambda+1)} \\ X_i \ \ if \ X_{(\lambda+1)} < X_i < X_{(n-\lambda)} \\ X_{(n-\lambda)} \ if \ X_i \ge X_{(n-\lambda)} \end{cases} \quad (3)$$

The sample Winsorized variance is:

$$s_W^2 = \frac{\sum_i (Y_i - \bar{X}_W)^2}{n-1} \quad (4)$$

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

$$q_j = \frac{(n_j - 1)s_{Wj}^2}{h_j(h_j - 1)}, \quad (5)$$

$$w_j = \frac{1}{q_j}, \quad (6)$$

$$U = \sum_j w_j, \quad (7)$$

$$\tilde{X} = \frac{1}{U}\sum_i w_j \bar{X}_{tj}, \quad (8)$$

$$A = \frac{1}{J-1}\sum_j w_j(\bar{X}_{tj} - \tilde{X})^2, \quad (9)$$

$$B = \frac{2(J-2)}{J^2-1}\sum_j \frac{(i - \frac{w_j}{U})^2}{h_j - 1}, and \quad (10)$$

$$F_t = \frac{A}{B+1} \quad (11)$$

The null hypothesis when using sample trimmed means is H_o: $\mu_{t1} = ... = \mu_{tJ}$ (i.e., the population trimmed means are equal), and is rejected if $F_t \ge F_{\alpha, J-1, \nu_{Wt}}$, where:

$$\nu_{Wt} = \frac{1}{\frac{3}{J^2-1}\sum_j \frac{(1-\frac{w_j}{U})^2}{h_j - 1}} \quad (12)$$

Fan and Hancock (2012) note two major criticisms with using a non-pooled standard error test in combination with trimming extreme observations. First, researchers may be hesitant to use trimming because it involves temporarily removing a portion of their data. Second, the null hypothesis relates to

trimmed, not ordinary, population mean differences. Lastly, they argue that Type I error rates and power of these techniques may not be satisfactory with larger degrees of nonnormality. The first two criticisms hold if researchers are only interested in comparing the full distributions of the populations, however if researchers are interested in comparing the 'bulk' of the distributions while limiting the effects of outliers then these criticisms do not hold. In other words, moving from a traditional null hypothesis (e.g., H_0: $\mu_1 = \mu_2$) to a robust null hypothesis (e.g., H_0: $\mu_{t1} = \mu_{t2}$, where μ_t represents the trimmed population mean) has little effect on the overall testing strategy (i.e., it simply eliminates the extreme scores from the analysis) and is typically preferred when the outlying cases have undue influence on the results of the analyses. With regard to the final criticism, the trimmed Welch has been found to be superior to alternative procedures when distributions are nonnormal and population variances are unequal (Cribbie et al., 2012). These potential limitations led Fan and Hancock to propose a new structural equation modelling (SEM; Bollen, 1989) based approach entitled robust means modeling.

Robust Means Modeling

Robust means modeling (RMM; Fan & Hancock, 2012) is a SEM technique inspired by Sorbom's (1974) structured means modeling (SMM). Specifically, it is a special case of SMM where the means being compared are observed variables (e.g., an ANOVA model) instead of latent variables (Fan & Hancock, 2012). The SMM approach can be represented in matrix form by the following model:

$$\boldsymbol{x} = \boldsymbol{\upsilon_k} + \boldsymbol{\Lambda_k}\boldsymbol{\xi} + \boldsymbol{\delta} \tag{13}$$

where $\boldsymbol{x}$ is a p x 1 vector of observed indicators of a latent variable, $\boldsymbol{\xi}$; υ_k is a p x 1 vector of intercepts; $\boldsymbol{\Lambda}_k$ is a p x 1 vector of factor loadings λ; and $\boldsymbol{\delta}$ is a p x 1 vector of errors. The model for RMM is a simpler version of the SMM model because there are no latent variables ($\boldsymbol{\xi}$) and therefore no factor loadings ($\boldsymbol{\Lambda}_k$) leaving only: $x = \upsilon_k + \delta$ (Fan & Hancock, 2012). The null hypothesis remains H_0: $\upsilon_1 = \upsilon_2 = \ldots = \upsilon_K$, where υ represents the population intercepts/means, but the method for comparing the means differs. Specifically, the means are constrained to be equal in the SEM model and the variances are free to be estimated, thereby removing the homogeneity of variance assumption. The SMM model can be

estimated through a weighted combination of the multi-group maximum likelihood (ML) fit functions:

$$F_{ML} = \sum_{k=1}^{K} \left(\frac{n_k}{N}\right) F_k\left(\boldsymbol{S}_k, \boldsymbol{m}_k, \widehat{\boldsymbol{\Sigma}}_k, \widehat{\boldsymbol{\mu}}_k\right) \quad (14)$$

where n_k is the sample size of the *k*th group, *N* is the total sample size for all groups, $\boldsymbol{S}_k$ is the *k*th group's observed covariance matrix, $\boldsymbol{m}_k$ is the *k*th group's observed mean vector, $\widehat{\boldsymbol{\Sigma}}_k$ is the *k*th group's model-implied covariance matrix, $\widehat{\boldsymbol{\mu}}_k$ is the *k*th group's model-implied vector of means and F_k is the *k*th group's ML fit function defined as:

$$F_k = [\ln\left|\widehat{\boldsymbol{\Sigma}}_k\right| + tr\left(\boldsymbol{S}_k\widehat{\Sigma}_k^{-1}\right) - \ln|\boldsymbol{S}_k| - p] + (\boldsymbol{m}_k - \widehat{\boldsymbol{\mu}}_k)'\widehat{\Sigma}_k^{-1}(\boldsymbol{m}_k - \widehat{\boldsymbol{\mu}}_k) \quad (15)$$

where *p* is the number of observed variables (i.e., 1 for RMM).

F_{ML} can be used to calculate a test statistic that quantifies evidence against the null hypothesis of mean equality. Specifically $T_{ML} = (N-1)F_{ML}$ with degrees of freedom (*df*) = $Kp(p+3)/2 - q$, where *K* is the number of groups, *p* is the number of observed variables, and *q* is the number of parameters estimated for the model (Fan & Hancock, 2012). The only parameters estimated in the RMM model are the *K* population variances plus one population mean (constrained to be equal across the *K* groups). Therefore the *df* in the RMM model simplifies considerably to *K* - 1. T_{ML} follows a χ^2distribution when data are normal, but it becomes biased as data become less normally distributed.

Although traditional ML estimation requires multivariate normality to produce unbiased results, there are a number of modified estimation procedures designed to alleviate issues stemming from nonnormality (e.g., Browne, 1984; Satorra & Bentler, 2001; Yuan & Bentler, 1999). For our study we chose to test many of the original RMM procedures in Fan and Hancock's (2012) study. They are described below.

Asymptotically distribution free method (ADF). One of the first modifications to ML is Browne's (1984) ADF method. It is also known as arbitrary generalized least squares (AGLS) or weighted least squares (WLS). Unlike traditional ML, the ADF method does not require the multivariate normality assumption as a condition for its use. The ADF method is based on the generalized least squares approach,

but uses a different weight matrix that allows for nonnormal data. It can be written as the following weighted fit function for multiple groups:

$$F_{ADF} = \sum_{k=1}^{K} (\boldsymbol{s_k} - \hat{\boldsymbol{\sigma}}_{\boldsymbol{k}})' W_k^{-1} (\boldsymbol{s_k} - \hat{\boldsymbol{\sigma}}_k) \tag{16}$$

where $\boldsymbol{s_k}$ is the p^* x 1 vector of first and second moments of the distribution of observed means, variances, and covariances ($p^* = p(p+3)/2$), $\hat{\boldsymbol{\sigma}}_{\boldsymbol{k}}$ is the p^* x 1 vector of model-implied first and second moments, and W_k^{-1} is a p^* x p^* weight matrix of higher moments. For more details about the weight matrix see Browne (1984) or Muthén (1989). Using the ADF method, one can obtain test statistic

$$T_{ADF} = (N-1)F_{ADF} \tag{17}$$

which is distributed as χ^2 with $K - 1$ *df*. In theory, the ADF method solves estimation issues arising from models with nonnormal data, but simulation studies demonstrate that it requires very large sample sizes and may be limited in the number of variables in the SEM model to obtain stable estimates for the weight matrix (e.g., Curran, West, & Finch, 1996; Finch, West, and MacKinnon, 1997; Muthén, & Kaplan, 1992; Olsson, Foss, Troye, & Howell, 2000).

Modified ADF methods. Given the issues discussed above concerning the ADF method, researchers have proposed modifications to correct for estimation issues resulting from small sample sizes. For example, Fan and Hancock (2012) recommended two modified ADF methods by Yuan and Bentler (1997; 1999). The first statistic (YB1; Yuan & Bentler, 1997) modifies the ADF statistic as follows:

$$T_{YB1} = \frac{T_{ADF}}{(1 + T_{ADF} N^{-1})} \tag{18}$$

The T_{YB1} follows a χ^2 distribution and has the same *df* as the ADF model (*K*-1 for the RMM model).

Yuan and Bentler's (1999) second modified ADF statistic (YB2) follows an *F* distribution and can be expressed by the following equation:

$$T_{YB2} = T_{ADF} \frac{(N - (Kp^* - q))}{(N-1)(Kp^* - q)} \tag{19}$$

with numerator $df = Kp^*-q$ and denominator $df = N - (Kp^*-q)$. In RMM, the *df* simplifies to *K*-1 for the

numerator and $N - K + 1$ for the denominator *df* (Fan & Hancock).

Scaling corrections to ML. The Satorra-Bentler (SB; Satorra & Bentler, 1988) rescaled test statistic is another popular alternative to traditional ML estimation, which was extended to include mean testing (Satorra, 1992). The new statistic, $T_{SB} = T_{ML}\hat{c}^{-1}$ where $\hat{c}^{-1}$ is a scaling factor that takes into account the model, estimation procedure, and degree of kurtosis. It is approximately distributed as χ^2 with the same *df* as T_{ML} and includes the use of robust standard errors. For technical details about the scaling constant see Satorra (1992) or Satorra and Bentler (2001). The SB rescaled test has been found to perform well in general, and better than the ADF methods with smaller sample sizes (e.g., Hu, Bentler, & Kano, 1992; Curran, West, & Finch, 1996). Given these options, RMM is a promising method as it includes robust estimation techniques to combat issues with nonnormal data and allows for distinct model estimates of population variances.

Performance of the RMM Methods

Fan and Hancock (2012) evaluated the performance of several RMM procedures in comparison to four modified ANOVA procedures along with the traditional ANOVA *F*-test as a reference. The alternative ANOVA procedures included the Welch test (Welch, 1951), Brown and Forsythe method (Brown & Forsythe, 1974c), Alexander-Govern method (Alexander & Govern, 1994), and James' second order test (James, 1951). In Fan and Hancock's study, trimmed means and Winsorized variances were incorporated into each of the four ANOVA alternatives. Under various conditions of nonnormality and unequal population variances, Fan and Hancock (2012) found that the RMM procedures outperformed the traditional methods and alternatives with regard to Type I error rates and power when moderate to extreme amounts of nonnormality were combined with unequal sample sizes and variances. Their study reported very liberal Type I error rates for the modified ANOVA tests including the trimmed Welch, and due to inaccurate Type I error rates, power results were not reported. It was, however, noted that the power of the RMM methods was higher than the ANOVA-based methods, although the power difference between the approaches decreased as sample size increased.

Fan and Hancock's (2012) results contrasted with previous research demonstrating that the Welch

test with trimmed means and Winsorized variances has accurate Type I error rates and adequate power results (e.g., Cribbie et al, 2012; Lix et al., 1996). Fan and Hancock found slight differences in performance across the RMM methods, depending on the condition, but overall the pattern of results was similar for the procedures. For Type I error rates, the RMM methods all provided good results, only deviating substantially from the nominal level in a few conditions. The results were significantly better than the ANOVA-based methods. The RMM method with the highest power was Browne's (1984) ADF method, followed by the two Yuan and Bentler (1997; 1999) statistics. Based on the overall performance of all of the methods under study, Fan and Hancock recommended the two Yuan and Bentler adjusted ADF methods over the ANOVA-based methods and other RMM approaches.

Study Objectives

Given the promising results for RMM procedures, we sought to extend the findings of Fan and Hancock (2012) in two ways. First, we simulated data from two different families of nonnormal distributions not investigated by Fan and Hancock -- the *g* and *h* distribution (Hoaglin, 1985) and the χ^2distribution. We also included results on the performance of the RMM procedures when the distribution shapes differed across groups (e.g., one group had positively skewed data, another group had normally distributed or negatively skewed data, etc.). Lastly, like Fan and Hancock (2012), we included the trimmed Welch because it is widely recommended for comparing population means under nonnormality and variance heterogeneity (Cribbie et al., 2012; Wilcox, 2017). The poor performance of the trimmed Welch in Fan and Hancock was unexpected and deserves further investigation. The expanded conditions of the current paper will allow further comparisons between the trimmed Welch and the RMM procedures.

Methodology

A Monte Carlo study was constructed to evaluate the performance (i.e., power and Type I error rates) of traditional ANOVA-based methods and RMM tests for comparing independent group means across many conditions of nonnormality and variance heterogeneity. The ANOVA-based methods included the traditional ANOVA (for baseline comparisons only), Welch's (1951) heteroscedastic

procedure with both usual means and variances (Welch) and trimmed means (20% symmetric trimming) and Winsorized variances (T Welch). The RMM methods included the traditional maximum-likelihood (ML) approach based on a χ^2test, a maximum likelihood-based Satorra-Bentler corrected test (SB), the asymptotically distribution free (ADF) test, and the two sample-size adjusted ADF methods due to Yuan and Bentler (YB1,YB2). The study used the open source software *R* (*R* Core Team, 2014). The simulation results were organized with the `SimDesign` package (Chalmers, 2016) and the RMM models were all evaluated using `lavaan`, an R package for the analysis of latent variable models (Rosseel, 2012).

Several variables were investigated in the simulation study, including the number of groups (*K* = 2 or 4), mean pattern (for investigating Type I error rates and power), sample sizes, population distribution shapes, and variance heterogeneity. Average group sample sizes included 10, 50 and 200 with both equal and unequal sample size conditions. Both equal and unequal variance conditions were included. The largest to smallest variance ratio was 16:1, which represents extreme levels of variance heterogeneity (Keselman et al., 1998). Two different heterogeneous variance conditions were included, one for a positively paired variance and sample size and one with a negatively paired variance and sample size. In addition to generating data from the normal (Gaussian) distribution, we simulated data from the χ^2distribution (with 3 *df*, skewness = 1.64, kurtosis = 4.00) and the *g* and *h* distribution (Hoaglin, 1985) with a positively skewed distribution (*g* = 1, *h* = 0, skewness = 6.18, kurtosis = 113.94) and its negatively skewed counterpart (*g* = -1, *h* = 0). These distributions are expected to represent the moderate (χ^2) to extremely skewed (*g/h*) distributions that behavioural science researchers would encounter (Wilcox, 1995). We explored conditions where all groups have the same population distribution shape, as well as conditions with mixtures of population distribution shapes (e.g., first population normal and second population positively skewed for *K* = 2).

In total we explored 420 unique conditions (300 conditions with the *g* and *h* distribution and 120 conditions with the χ^2distribution). The specific conditions for the simulation study are presented in Table 1 and were selected to match common design conditions in psychological research. To generate

pseudo-random normal variates, we used the R generator 'rnorm' (R Development Core Team, 2016). 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:

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

where $g = 1/-1$ and $h = 0$. 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 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).

The nominal Type I error rate (α) was set at .05 for all conditions. Finally, 5000 replications were conducted for each condition. The code used for running the simulations is available at http://?.? (omitted for blind review).

Table 1
Simulation Conditions

Distributions	Normal, *g* and *h* distribution (Positive, Negative Skew), χ^2	
Dist. Patterns	Same Distribution Shape or Different for Half of the Groups	
T1 Mean Pattern	All population means = 0	
	K=2	*K*=4
Variance Pattern	1,1 or 1,16	1,1,1,1 or 1,4,9,16
Avg. *n* = 10	10,10; 4,16; 16,4	10,10,10,10; 4,8,12,16; 16,12,8,4
Power Mean Pattern	0, 1.325	0, 0.493, 0.986, 1.479
Avg *n* = 50	50,50; 20, 80; 80, 20	50,50,50,50; 20,40,60,80; 80,60,40,20
Power Mean Pattern	0, .566	0, 0.211, 0.422, 0.633
Avg. *n* = 200	200, 200; 80, 320; 320, 80	200,200,200,200; 80,160,240,320; 320,240,160,80
Power Mean Pattern	0, .281	0, 0.105, 0.209, 0.314

Note: The mean patterns for power conditions were calculated for each of the three sample size conditions such that the power would be approximately .80 under normality, equal *n*s and equal group variance; For Type I error rates, the raw population means were zero, but the trimmed population means were used for calculating Type I error rates for the trimmed Welch test.

Results

Due to the large number of conditions, only a subset of the results is presented below. Specifically, we present the results for the moderate (average *n* =50) and the large (average *n* = 200) sample size condition with four groups for each of the distribution types. These conditions were chosen to highlight any simulation conditions that had an effect on the Type I error rates. The results when *K* = 2 mirror those when *K* = 4. The full simulation results can be obtained from the first author.

Estimation Issues for the RMM Methods

Nonconvergence rates were minimal for RMM models as they converged in over 99.9% of the replications across all of the conditions. However, with smaller sample sizes (average *n* of 10) the ADF methods exhibited problems with nonpositive definite matrices in the majority of conditions (rates as high as 88%). This was no longer an issue when the average *n* per group increased to 50 or 200.

Type I Error Rates

The nominal Type I error rate was set at .05 for all investigated conditions and empirical rates were considered acceptable if they fell within Bradley's (1978) liberal bounds (i.e., α +/- .5α). All of the tests were found to have accurate Type I error rates when all of the groups' data follow a normal

distribution with equal group variance and sample sizes. However, once the groups' data did not follow a normal distribution (e.g., extremely positively or negatively skewed), many of the investigated tests no longer demonstrated accurate error rates. The only method found to maintain accurate empirical Type I error rates across all of the investigated conditions was the trimmed Welch ANOVA.

***g* and *h* distribution**. Tables 2 and 3 display the empirical Type I error rates for each of the tests when data were generated from the *g* and *h* distributions for average group sample sizes of 50 and 200, respectively. The accuracy of the Type I error rates for the RMM methods improves as sample size increases. When the average sample size per group was 50 (as can be seen in Table 2), the RMM methods' empirical error rates were found to be more liberal than the nominal α level under several simulation conditions. For example, in cases where the sample size and variance were negatively paired, rates were as high as .180 when the groups' distribution shapes were the same, and as high as .183 when the distribution shapes differed. When the average group sample size increased to 200 (see Table 3), the Type I error rates for the RMM methods become closer to the nominal α level, but are still somewhat liberal (e.g., as high as .10 in negative pairing conditions). As demonstrated in the tables, the error rates for the RMM approaches were very similar to one another regardless of the method used (e.g., traditional ML versus YB2).

Table 2

g and h Distribution: Omnibus Type I Error Rates for K = 4 and Average n = 50

Distribution	*n*	*σ*	ANOVA	Welch	T Welch	ML	SB	ADF	YB1	YB2
0,0,0,0	50,50,50,50	1,1,1,1	0.055	0.053	0.056	0.059	0.059	0.063	0.054	0.057
	50,50,50,50	1,2,3,4	0.067	0.051	0.058	0.055	0.055	0.057	0.052	0.054
	20,40,60,80	1,1,1,1	0.048	0.051	0.057	0.058	0.058	0.063	0.055	0.056
	20,40,60,80	1,2,3,4	**0.017**	0.049	0.052	0.052	0.052	0.055	0.049	0.051
	80,60,40,20	1,2,3,4	**0.202**	0.052	0.062	0.058	0.058	0.064	0.057	0.058
1,1,1,1	50,50,50,50	1,1,1,1	0.041	0.059	0.049	0.067	0.067	0.068	0.060	0.063
	50,50,50,50	1,2,3,4	0.075	**0.102**	0.062	**0.107**	**0.107**	**0.112**	**0.103**	**0.105**
	20,40,60,80	1,1,1,1	0.049	0.075	0.059	**0.087**	**0.087**	**0.090**	**0.081**	**0.084**
	20,40,60,80	1,2,3,4	0.026	0.071	0.048	**0.078**	**0.078**	**0.082**	0.073	**0.076**
	80,60,40,20	1,2,3,4	**0.197**	**0.156**	0.075	**0.169**	**0.169**	**0.180**	**0.166**	**0.172**
2,2,2,2	50,50,50,50	1,1,1,1	0.043	0.057	0.050	0.066	0.066	0.069	0.058	0.061
	50,50,50,50	1,2,3,4	0.069	**0.101**	0.061	**0.108**	**0.108**	**0.112**	**0.102**	**0.105**
	20,40,60,80	1,1,1,1	0.048	**0.081**	0.061	**0.093**	**0.093**	**0.098**	**0.089**	**0.091**
	20,40,60,80	1,2,3,4	0.025	0.067	0.053	0.074	0.074	**0.076**	0.069	0.072
	80,60,40,20	1,2,3,4	**0.202**	**0.150**	0.075	**0.163**	**0.163**	**0.171**	**0.164**	**0.167**
0,0,1,1	50,50,50,50	1,1,1,1	0.048	0.065	0.046	0.071	0.071	0.073	0.066	0.068
	50,50,50,50	1,2,3,4	**0.077**	**0.110**	0.060	**0.117**	**0.117**	**0.121**	**0.112**	**0.115**
	20,40,60,80	1,1,1,1	0.056	0.050	0.048	0.055	0.055	0.061	0.054	0.056
	20,40,60,80	1,2,3,4	0.029	**0.082**	0.054	**0.091**	**0.091**	**0.094**	**0.085**	**0.087**
	80,60,40,20	1,2,3,4	**0.184**	**0.143**	0.068	**0.151**	**0.151**	**0.160**	**0.151**	**0.152**

0,0,2,2	50,50,50,50	1,1,1,1	0.050	0.063	0.049	0.070	0.070	0.072	0.065	0.067
	50,50,50,50	1,2,3,4	**0.080**	**0.112**	0.059	**0.120**	**0.120**	**0.124**	**0.113**	**0.117**
	20,40,60,80	1,1,1,1	0.064	0.059	0.059	0.068	0.068	0.074	0.066	0.069
	20,40,60,80	1,2,3,4	0.026	0.072	0.051	**0.078**	**0.078**	**0.080**	0.075	**0.076**
	80,60,40,20	1,2,3,4	**0.195**	**0.156**	0.069	**0.165**	**0.165**	**0.175**	**0.165**	**0.167**
1,1,2,2	50,50,50,50	1,1,1,1	0.060	**0.126**	0.054	**0.133**	**0.133**	**0.139**	**0.128**	**0.131**
	50,50,50,50	1,2,3,4	**0.090**	**0.144**	0.066	**0.151**	**0.151**	**0.158**	**0.146**	**0.149**
	20,40,60,80	1,1,1,1	0.064	**0.148**	0.062	**0.157**	**0.157**	**0.162**	**0.154**	**0.156**
	20,40,60,80	1,2,3,4	0.029	**0.131**	0.060	**0.142**	**0.142**	**0.145**	**0.134**	**0.138**
	80,60,40,20	1,2,3,4	**0.200**	**0.163**	0.075	**0.174**	**0.174**	**0.183**	**0.175**	**0.177**

Note: Dist = 0 is the normal distribution, Dist = 1 is a positively skewed distribution with $g = 1$ and $h = 0$, and Dist = 2 is the negatively skewed distribution for $g = 1$ and $h = 0$; Rates outside of Bradley's liberal bounds are bolded.

Table 3

g and h Distribution: Omnibus Type I Error Rates for K = 4 and average n = 200

Distribution	*n*	σ	ANOVA	Welch	T Welch	ML	SB	ADF	YB1	YB2
0,0,0,0	200,200,200,200	1,1,1,1	0.050	0.052	0.054	0.052	0.052	0.052	0.052	0.052
	200,200,200,200	1,2,3,4	0.073	0.050	0.051	0.052	0.052	0.053	0.051	0.051
	80,160,240,320	1,1,1,1	0.057	0.056	0.052	0.057	0.057	0.058	0.057	0.058
	80,160,240,320	1,2,3,4	**0.019**	0.050	0.050	0.051	0.051	0.052	0.050	0.050
	320,240,160,80	1,2,3,4	**0.195**	0.051	0.051	0.052	0.052	0.053	0.052	0.052
1,1,1,1	200,200,200,200	1,1,1,1	0.048	0.058	0.048	0.060	0.060	0.061	0.058	0.059
	200,200,200,200	1,2,3,4	0.068	0.073	0.050	0.074	0.074	0.075	0.073	0.073
	80,160,240,320	1,1,1,1	0.047	0.067	0.049	0.068	0.068	0.068	0.067	0.067
	80,160,240,320	1,2,3,4	**0.021**	0.061	0.050	0.062	0.062	0.062	0.061	0.061
	320,240,160,80	1,2,3,4	**0.189**	**0.098**	0.057	**0.100**	**0.100**	**0.101**	**0.100**	**0.100**
2,2,2,2	200,200,200,200	1,1,1,1	0.046	0.055	0.057	0.058	0.058	0.058	0.056	0.056
	200,200,200,200	1,2,3,4	0.071	**0.076**	0.059	**0.078**	**0.078**	**0.079**	**0.077**	**0.078**
	80,160,240,320	1,1,1,1	0.046	0.067	0.050	0.069	0.069	0.070	0.068	0.069
	80,160,240,320	1,2,3,4	**0.024**	0.062	0.049	0.064	0.064	0.065	0.063	0.063
	320,240,160,80	1,2,3,4	**0.201**	**0.095**	0.056	**0.098**	**0.098**	**0.100**	**0.097**	**0.098**
0,0,1,1	200,200,200,200	1,1,1,1	0.052	0.063	0.050	0.065	0.065	0.066	0.063	0.063
	200,200,200,200	1,2,3,4	0.070	0.074	0.050	0.075	0.075	0.075	0.074	0.075
	80,160,240,320	1,1,1,1	0.057	0.055	0.053	0.058	0.058	0.059	0.057	0.057
	80,160,240,320	1,2,3,4	**0.024**	0.068	0.052	0.070	0.070	0.071	0.068	0.069
	320,240,160,80	1,2,3,4	**0.194**	**0.095**	0.051	**0.097**	**0.097**	**0.100**	**0.097**	**0.098**
0,0,2,2	200,200,200,200	1,1,1,1	0.049	0.056	0.055	0.057	0.057	0.057	0.056	0.056

	200,200,200,200	1,2,3,4	0.063	0.052	0.050	0.054	0.054	0.056	0.053	0.054
	80,160,240,320	1,1,1,1	0.052	**0.078**	0.056	**0.080**	**0.080**	**0.082**	**0.079**	**0.079**
	80,160,240,320	1,2,3,4	**0.024**	0.055	0.051	0.057	0.057	0.057	0.055	0.056
	320,240,160,80	1,2,3,4	**0.197**	0.058	0.050	0.061	0.061	0.062	0.060	0.061
1,1,2,2	200,200,200,200	1,1,1,1	0.048	**0.076**	0.052	**0.077**	**0.077**	**0.078**	**0.076**	**0.076**
	200,200,200,200	1,2,3,4	0.074	**0.085**	0.053	**0.086**	**0.086**	**0.087**	**0.085**	**0.086**
	80,160,240,320	1,1,1,1	0.054	**0.096**	0.053	**0.097**	**0.097**	**0.099**	**0.096**	**0.097**
	80,160,240,320	1,2,3,4	**0.023**	**0.086**	0.054	**0.087**	**0.087**	**0.088**	**0.086**	**0.087**
	320,240,160,80	1,2,3,4	**0.211**	**0.106**	0.062	**0.107**	**0.107**	**0.110**	**0.108**	**0.108**

χ^2 distribution. Tables 4 and 5 display the empirical error rates when data follow a χ^2 distribution with three degrees of freedom for average group sample sizes of 50 and 200, respectively. The most notable finding is that the Type I error rates for all of the procedures are better than those observed in similar conditions but with data generated from the *g* and *h* distribution. As can be seen in the tables, the RMM methods' error rates improve with the larger sample size condition whereby they only fall outside of Bradley's liberal bounds when the average *n* is 50 in the negative pairing conditions, and are still less than 2α. When the sample size increases to 200, the RMM error rates are accurate in all of the variance-sample size pairings. The trimmed Welch procedure's error rates are accurate across all conditions regardless of sample size.

Table 4

χ^2 Distribution: Omnibus Type I Error Rates for K = 4 and Average n = 50

Distribution	*n*	σ	ANOVA	Welch	T Welch	ML	SB	ADF	YB1	YB2
1,1,1,1	50,50,50,50	1,1,1,1	0.050	0.057	0.051	0.063	0.063	0.067	0.059	0.061
	50,50,50,50	1,2,3,4	0.067	0.065	0.054	0.068	0.068	0.071	0.065	0.067
	20,40,60,80	1,1,1,1	0.053	0.060	0.056	0.070	0.070	0.074	0.066	0.068
	20,40,60,80	1,2,3,4	**0.023**	0.054	0.056	0.058	0.058	0.060	0.055	0.057
	80,60,40,20	1,2,3,4	**0.212**	**0.083**	0.070	**0.092**	**0.092**	**0.101**	**0.092**	**0.095**
0,0,1,1	50,50,50,50	1,1,1,1	0.051	0.054	0.049	0.059	0.059	0.062	0.055	0.057
	50,50,50,50	1,2,3,4	0.072	0.069	0.051	0.073	0.073	**0.076**	0.070	0.072
	20,40,60,80	1,1,1,1	0.046	0.049	0.054	0.058	0.058	0.064	0.056	0.059
	20,40,60,80	1,2,3,4	**0.021**	0.055	0.058	0.061	0.061	0.065	0.056	0.058
	80,60,40,20	1,2,3,4	**0.205**	**0.077**	0.064	**0.083**	**0.083**	**0.093**	**0.085**	**0.088**

Note: Dist=1 is the χ^2 distribution with 3 degrees of freedom, and Dist=0 is the normal distribution

Table 5

χ^2 Distribution: Omnibus Type I Error Rates for K = 4 and Average n = 200

Distribution	*n*	σ	ANOVA	Welch	T Welch	ML	SB	ADF	YB1	YB2
1,1,1,1	200,200,200,200	1,1,1,1	0.051	0.053	0.052	0.054	0.054	0.054	0.053	0.053
	200,200,200,200	1,2,3,4	0.069	0.055	0.049	0.056	0.056	0.056	0.055	0.056
	80,160,240,320	1,1,1,1	0.052	0.055	0.050	0.057	0.057	0.058	0.057	0.057
	80,160,240,320	1,2,3,4	**0.021**	0.049	0.053	0.049	0.049	0.051	0.049	0.050
	320,240,160,80	1,2,3,4	**0.196**	0.058	0.050	0.059	0.059	0.060	0.060	0.060
0,0,1,1	200,200,200,200	1,1,1,1	0.048	0.049	0.048	0.050	0.050	0.051	0.049	0.050
	200,200,200,200	1,2,3,4	0.065	0.048	0.051	0.049	0.049	0.049	0.048	0.048
	80,160,240,320	1,1,1,1	0.051	0.051	0.057	0.053	0.053	0.055	0.053	0.053
	80,160,240,320	1,2,3,4	**0.019**	0.053	0.046	0.055	0.055	0.055	0.054	0.054
	320,240,160,80	1,2,3,4	**0.201**	0.061	0.055	0.064	0.064	0.065	0.064	0.064

Power Rates

Fan and Hancock (2012) did not report power results for the trimmed Welch because their simulation study reported inaccurate error rates for the test. As this was not found in our simulation, we present power results for the same conditions investigated above. Power rates are in bold when the error rates for the same conditions in the previous section fell outside of Bradley's liberal bounds. The population means are different for the average n = 50 and average n = 200 conditions because the mean pattern reflects power rates of approximately .80 for homoscedastic and normally distributed data with equal sample sizes per group. Taking this approach means that there is no power increase by increasing sample size from 50 to 200, because the conditions use different population means to assess power.

***g* and *h* distribution.** Tables 6 and 7 present the power results under the same conditions used for Type I error rates for the two sample sizes when data follow the *g* and *h* distribution. When all assumptions have been met, the trimmed Welch has somewhat lower power than the other procedures by about 8% (which is expected because of the reduced effective sample size with trimming). With skewed data, however, the trimmed Welch demonstrates comparable, and for many conditions, superior power rates compared to the RMM methods. This pattern of results was true even when the RMM methods were found to have liberal error rates. Note that for all procedures, unequal variances drastically decreases power to detect population mean differences. As was seen with Type I error rates, the RMM procedures exhibited similar power rates to one another, such that one procedure did not consistently outperform the others.

χ^2 distribution. Tables 8 and 9 present the power results for outcomes that follow a χ^2 distribution with three *df*. A similar pattern of power results was observed as those discussed above when data were generated from the *g* and *h* distribution. Namely, the power results for the trimmed Welch and RMM approaches were quite similar across the conditions and the different RMM procedures were almost identical to one another.

Table 6

Power Results for g and h distributions with K = 4 and average n = 50

Distribution	*n*	*σ*	ANOVA	Welch	T Welch	ML	SB	ADF	YB1	YB2
0,0,0,0	50,50,50,50	1,1,1,1	0.810	0.803	0.731	0.814	0.814	0.818	0.805	0.809
	50,50,50,50	1,2,3,4	0.156	0.190	0.176	0.206	0.206	0.214	0.194	0.200
	20,40,60,80	1,1,1,1	0.687	0.675	0.595	0.696	0.696	0.708	0.690	0.697
	20,40,60,80	1,2,3,4	**0.054**	0.187	0.165	0.203	0.203	0.209	0.191	0.195
	80,60,40,20	1,2,3,4	**0.311**	0.150	0.138	0.165	0.165	0.177	0.164	0.169
1,1,1,1	50,50,50,50	1,1,1,1	0.833	0.881	0.999	0.888	0.888	0.893	0.882	0.885
	50,50,50,50	1,2,3,4	0.111	**0.116**	0.488	**0.131**	**0.131**	**0.138**	**0.122**	**0.126**
	20,40,60,80	1,1,1,1	0.756	0.846	0.984	**0.858**	**0.858**	**0.862**	**0.853**	**0.856**
	20,40,60,80	1,2,3,4	0.029	0.217	0.513	**0.232**	**0.232**	**0.237**	0.219	**0.225**
	80,60,40,20	1,2,3,4	**0.275**	**0.085**	0.320	**0.101**	**0.101**	**0.109**	**0.099**	**0.102**
2,2,2,2	50,50,50,50	1,1,1,1	0.820	0.875	0.999	0.882	0.882	0.884	0.876	0.879
	50,50,50,50	1,2,3,4	0.252	**0.445**	0.598	**0.457**	**0.457**	**0.465**	**0.448**	**0.454**
	20,40,60,80	1,1,1,1	0.726	**0.779**	0.997	**0.795**	**0.795**	**0.802**	**0.790**	**0.792**
	20,40,60,80	1,2,3,4	0.119	0.362	0.576	0.374	0.374	**0.383**	0.367	0.373
	80,60,40,20	1,2,3,4	**0.407**	**0.455**	0.506	**0.469**	**0.469**	**0.481**	**0.467**	**0.471**
0,0,1,1	50,50,50,50	1,1,1,1	0.846	0.841	0.951	0.855	0.855	0.860	0.844	0.849
	50,50,50,50	1,2,3,4	**0.100**	**0.112**	0.326	**0.124**	**0.124**	**0.129**	**0.115**	**0.118**
	20,40,60,80	1,1,1,1	0.740	0.717	0.857	0.744	0.744	0.757	0.733	0.742
	20,40,60,80	1,2,3,4	0.030	**0.127**	0.289	**0.138**	**0.138**	**0.145**	**0.131**	**0.134**
	80,60,40,20	1,2,3,4	**0.250**	**0.100**	0.231	**0.120**	**0.120**	**0.126**	**0.114**	**0.117**

0,0,2,2	50,50,50,50	1,1,1,1	0.796	0.845	0.911	0.855	0.855	0.857	0.848	0.852
	50,50,50,50	1,2,3,4	**0.252**	**0.414**	0.440	**0.426**	**0.426**	**0.433**	**0.416**	**0.422**
	20,40,60,80	1,1,1,1	0.711	0.722	0.821	0.739	0.739	0.749	0.734	0.739
	20,40,60,80	1,2,3,4	0.118	0.347	0.335	**0.360**	**0.360**	**0.366**	0.350	**0.354**
	80,60,40,20	1,2,3,4	**0.396**	**0.414**	0.376	**0.434**	**0.434**	**0.447**	**0.433**	**0.438**
1,1,2,2	50,50,50,50	1,1,1,1	0.768	**0.868**	0.989	**0.874**	**0.874**	**0.876**	**0.868**	**0.871**
	50,50,50,50	1,2,3,4	**0.263**	**0.431**	0.547	**0.439**	**0.439**	**0.448**	**0.434**	**0.437**
	20,40,60,80	1,1,1,1	0.704	**0.826**	0.968	**0.835**	**0.835**	**0.841**	**0.832**	**0.835**
	20,40,60,80	1,2,3,4	0.120	**0.438**	0.546	**0.451**	**0.451**	**0.457**	**0.441**	**0.445**
	80,60,40,20	1,2,3,4	**0.412**	**0.416**	0.475	**0.432**	**0.432**	**0.444**	**0.430**	**0.435**

Bolded values indicate conditions where the Type I error rates were unacceptable.

Table 7

Power Results for g and h distributions with K = 4 and average n = 200

Distribution	*n*	*σ*	ANOVA	Welch	T Welch	ML	SB	ADF	YB1	YB2
0,0,0,0	200,200,200,200	1,1,1,1	0.803	0.799	0.729	0.801	0.801	0.803	0.799	0.800
	200,200,200,200	1,2,3,4	0.145	0.196	0.174	0.200	0.200	0.202	0.197	0.199
	80,160,240,320	1,1,1,1	0.700	0.696	0.623	0.701	0.701	0.703	0.699	0.700
	80,160,240,320	1,2,3,4	**0.052**	0.192	0.164	0.195	0.195	0.196	0.193	0.193
	320,240,160,80	1,2,3,4	**0.318**	0.172	0.151	0.177	0.177	0.180	0.176	0.177
1,1,1,1	200,200,200,200	1,1,1,1	0.808	0.839	0.999	0.840	0.840	0.842	0.839	0.840
	200,200,200,200	1,2,3,4	0.123	0.135	0.514	0.137	0.137	0.139	0.136	0.137
	80,160,240,320	1,1,1,1	0.736	0.790	0.993	0.795	0.795	0.797	0.793	0.794
	80,160,240,320	1,2,3,4	**0.040**	0.198	0.536	0.202	0.202	0.203	0.198	0.200
	320,240,160,80	1,2,3,4	**0.273**	**0.090**	0.393	**0.094**	**0.094**	**0.097**	**0.092**	**0.094**
2,2,2,2	200,200,200,200	1,1,1,1	0.819	0.847	0.999	0.849	0.849	0.850	0.847	0.848
	200,200,200,200	1,2,3,4	0.224	**0.354**	0.593	**0.357**	**0.357**	**0.358**	**0.355**	**0.356**
	80,160,240,320	1,1,1,1	0.700	0.718	0.998	0.724	0.724	0.727	0.721	0.722
	80,160,240,320	1,2,3,4	**0.084**	0.272	0.552	0.276	0.276	0.278	0.273	0.275
	320,240,160,80	1,2,3,4	**0.378**	**0.342**	0.502	**0.347**	**0.347**	**0.349**	**0.346**	**0.347**
0,0,1,1	200,200,200,200	1,1,1,1	0.827	0.823	0.949	0.827	0.827	0.829	0.824	0.825
	200,200,200,200	1,2,3,4	0.117	0.141	0.362	0.145	0.145	0.146	0.141	0.143
	80,160,240,320	1,1,1,1	0.721	0.709	0.861	0.715	0.715	0.719	0.713	0.715
	80,160,240,320	1,2,3,4	**0.033**	0.140	0.307	0.144	0.144	0.145	0.140	0.141
	320,240,160,80	1,2,3,4	**0.268**	**0.102**	0.257	**0.105**	**0.105**	**0.108**	**0.105**	**0.105**
0,0,2,2	200,200,200,200	1,1,1,1	0.784	0.812	0.931	0.816	0.816	0.817	0.813	0.814

	200,200,200,200	1,2,3,4	0.212	0.329	0.419	0.332	0.332	0.333	0.330	0.331
	80,160,240,320	1,1,1,1	0.690	**0.709**	0.843	**0.715**	**0.715**	**0.718**	**0.714**	**0.715**
	80,160,240,320	1,2,3,4	**0.086**	0.276	0.321	0.279	0.279	0.280	0.277	0.278
	320,240,160,80	1,2,3,4	**0.365**	0.314	0.349	0.318	0.318	0.322	0.318	0.320
1,1,2,2	200,200,200,200	1,1,1,1	0.774	**0.832**	0.998	**0.835**	**0.835**	**0.836**	**0.832**	**0.833**
	200,200,200,200	1,2,3,4	0.216	**0.338**	0.557	**0.341**	**0.341**	**0.344**	**0.340**	**0.341**
	80,160,240,320	1,1,1,1	0.695	**0.783**	0.989	**0.786**	**0.786**	**0.787**	**0.785**	**0.786**
	80,160,240,320	1,2,3,4	**0.086**	**0.334**	0.541	**0.339**	**0.339**	**0.340**	**0.335**	**0.336**
	320,240,160,80	1,2,3,4	**0.370**	**0.314**	0.462	**0.318**	**0.318**	**0.322**	**0.318**	**0.320**

Bolded values indicate conditions where the Type I error rates were unacceptable.

Table 8

Power Rates for the χ^2 distribution with K = 4 and average n = 50

Distribution	*n*	σ	ANOVA	Welch	T Welch	ML	SB	ADF	YB1	YB2
1,1,1,1	50,50,50,50	1,1,1,1	0.802	0.802	0.827	0.814	0.814	0.820	0.803	0.809
	50,50,50,50	1,2,3,4	0.134	0.145	0.166	0.158	0.158	0.166	0.150	0.154
	20,40,60,80	1,1,1,1	0.715	0.733	0.714	0.751	0.751	0.758	0.743	0.747
	20,40,60,80	1,2,3,4	**0.041**	0.190	0.195	0.206	0.206	0.211	0.193	0.198
	80,60,40,20	1,2,3,4	**0.288**	**0.093**	0.122	**0.107**	**0.107**	**0.116**	**0.104**	**0.109**
0,0,1,1	50,50,50,50	1,1,1,1	0.812	0.797	0.784	0.813	0.813	0.817	0.800	0.806
	50,50,50,50	1,2,3,4	0.119	0.140	0.167	0.154	0.154	**0.161**	0.144	0.149
	20,40,60,80	1,1,1,1	0.704	0.684	0.656	0.708	0.708	0.719	0.697	0.705
	20,40,60,80	1,2,3,4	**0.039**	0.154	0.169	0.169	0.169	0.175	0.159	0.164
	80,60,40,20	1,2,3,4	**0.285**	**0.104**	0.117	**0.125**	**0.125**	**0.135**	**0.120**	**0.125**

Bolded values indicate conditions where the Type I error rates were unacceptable.

Table 9

Power Rates for the χ^2 distribution with K = 4 and average n = 200

Distribution	*n*	σ	ANOVA	Welch	T Welch	ML	SB	ADF	YB1	YB2
1,1,1,1	200,200,200,200	1,1,1,1	0.806	0.803	0.841	0.806	0.806	0.807	0.804	0.804
	200,200,200,200	1,2,3,4	0.147	0.163	0.188	0.166	0.166	0.168	0.163	0.165
	80,160,240,320	1,1,1,1	0.692	0.700	0.743	0.706	0.706	0.708	0.703	0.705
	80,160,240,320	1,2,3,4	**0.045**	0.191	0.196	0.194	0.194	0.195	0.192	0.193
	320,240,160,80	1,2,3,4	**0.315**	0.130	0.143	0.134	0.134	0.136	0.133	0.135
0,0,1,1	200,200,200,200	1,1,1,1	0.799	0.793	0.789	0.796	0.796	0.798	0.793	0.795
	200,200,200,200	1,2,3,4	0.136	0.160	0.176	0.165	0.165	0.167	0.161	0.163
	80,160,240,320	1,1,1,1	0.704	0.695	0.674	0.703	0.703	0.707	0.701	0.702
	80,160,240,320	1,2,3,4	**0.040**	0.166	0.185	0.170	0.170	0.172	0.166	0.168
	320,240,160,80	1,2,3,4	**0.294**	0.127	0.137	0.131	0.131	0.133	0.130	0.132

Bolded values indicate conditions where the Type I error rates were unacceptable.

Conclusion

Given the popularity of comparing mean differences and prevalence of assumption violation in research in the behavioural sciences (e.g., Blanca et al., 2011; Golinski & Cribbie, 2009; Keselman et al., 1998; Micceri, 1989), it is important that researchers have viable alternatives to the traditional ANOVA. A recent study proposed another robust statistical tool for comparing mean differences called robust means modeling (Fan & Hancock, 2012). Given their surprising results regarding the trimmed Welch test, the current study sought to replicate their findings and extend their paper in a few important ways; namely by examining their Type I error and power rates with other families of distributions (e.g., *g*/*h*, χ^2) and exploring their performance when the populations have differing distribution shapes.

RMM Procedures

Although Fan and Hancock (2012) recommended the YB1 or YB2 approaches as the better performing tests, we found little deviation in the performance of the different RMM methods. Type I error rates for the procedures were similar and there was no noticeable power advantage of any other approaches under the conditions investigated. The degree of similarity between the regular ML approach and the ADF methods was somewhat surprising because ML requires the assumption of normally distributed data. If one were to choose between the methods, the regular ML approach or Satorra-Bentler corrected ML test might actually be preferable with smaller sample sizes because they did not exhibit any problems with nonpositive definite matrices, whereas this was sometimes an issue for the ADF methods.

Distribution shape had an effect on the performance of the RMM methods. Empirical Type I error rates were much better when data followed a χ^2 (with three *df*) distribution compared to the positively or negatively skewed *g* and *h* distribution. However, this may simply be due to severity of nonnormality as the χ^2 distribution is less skewed than the *g* and *h* distributions used in the current study. Sample size also influenced the empirical Type I error rates, as the tests became overly conservative with smaller sample sizes. Given that the methods use ML estimation and the ADF methods are notorious for requiring larger sample sizes, it is possible that the models produces more biased estimates in smaller sample sizes, which manifested in the poorer Type I error rates.

Trimmed Welch Versus RMM

The most noteworthy finding is the difference between the performance of the trimmed Welch ANOVA and the RMM methods in the current study compared to what was reported in Fan and Hancock (2012). Specifically, they reported inconsistent and often extremely liberal Type I error rates for the trimmed Welch, whereas we found that the rates were very stable around the nominal α level. In fact, the trimmed Welch was the only procedure with empirical Type I error rates inside an acceptable range under all of the conditions tested. Our results regarding the Type I error rates of the Welch test on trimmed means agree with the results of several previous simulation studies including Cribbie, Wilcox, Bewell & Keselman (2007), Cribbie et al. (2012), Lix and Keselman (2006), and Wilcox (1995), and therefore we are confident that the Welch test on trimmed means is not overly liberal with heteroscedastic and/or skewed distributions. Additionally the trimmed Welch had comparable or higher power than the RMM tests, including conditions where the RMM's Type I error rates were more liberal than the nominal α rate.

A last topic worth discussing when comparing the trimmed Welch procedure to the RMM approaches is that of effect sizes (ES). Reporting statistical significance tests does not allow for an indication of the magnitude of group differences. Researchers who conduct an ANOVA often present the raw group means (or mean differences) as their ES measure when data are normally distributed (or have similar distribution shapes) and group variances are approximately equal. However, reporting raw mean differences may not be the best choice under assumption violation, as means are sensitive to nonnormality and outliers. Reporting trimmed means, however, is an intuitive and appropriate ES when conducting the trimmed Welch. It has an easy interpretation for applied researchers who are accustomed to reporting means or mean differences. In contrast, an appropriate measure of ES in conjunction with the RMM procedures is unclear. Reporting raw means with RMM methods is somewhat inconsistent as they do not account for the nonnormality or heterogeneity of the data.

Summary

In contrast to Fan and Hancock (2012), our simulation study found that RMM methods do not demonstrate better Type I error control and power than the trimmed Welch ANOVA. Whereas the Type I error rates of the RMM method often deviated from the nominal α level, the rates for the trimmed Welch were acceptable under all conditions. Further, the trimmed Welch had comparable if not higher power than the RMM methods when assumptions have been violated. Given the ease of interpretation and choice of ES using the trimmed Welch, its excellent Type I error control and often superior power under a variety of conditions including nonnormal distributions and unequal sample sizes and variances, we recommend that researchers use the trimmed Welch procedure for comparing the means of independent groups in situations where nonnormality and unequal variances are an issue.

References

Alexander, R. A., & Govern, D. M. (1994). A new and simpler approximation for ANOVA under variance heterogeneity. *Journal of Educational Statistics, 19,* 91-101.

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

Blanca, M. J., Arnau, J., Lopez-Montiel, D., Bono, R., & Bendayan, R. (2011) Skewness and kurtosis in real data samples. *Methodology, 92,* 78-84.

Bollen, K. A. (1989). Structural Equations with Latent Variables. NY: Wiley.

Boneau, C. A. (1960). The effects of violations of assumptions underlying the *t* test. *Psychological Bulletin, 57,* 49-64.

Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification. *Annals of Mathematical Statistics, 25,* 290-302.

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

Brown, M. B., & Forsythe, A. B. (1974a). The ANOVA and multiple comparisons for data with heterogeneous variances. *Biometrics, 30,* 719-724.

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

Brown, M. B. & Forsythe, A. B. (1974c). Robust tests for the equality of variances. *Journal of the American Statistical Association, 69,* 364-367.

Browne, M. W. (1984). Asymptotically distribution free methods in the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology, 37,* 127–141.

Chalmers, R. P. (2016). `SimDesign`: Structure for organizing Monte Carlo simulation designs [software manual]. Retrieved from https://CRAN.R-project.org/package=SimDesign (R package

version 1.3)

Cribbie, R. A., Fiksenbaum, L., Wilcox, R. R. & Keselman, H. J. (2012). Effects of nonnormality on test statistics for one-way independent groups designs. *British Journal of Mathematical and Statistical Psychology, 65,* 56-73.

Curran, P. J., West, S. G, & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods, 1,* 16-29.

Fan, W. & Hancock, G. R. (2012). Robust means modeling: An alternative for hypothesis testing of independent means under variance heterogeneity and nonnormality. *Journal of Educational and Behavioral Statistics, 37,* 137-156.

Finch, J. F., West, S. G., & MacKinnon, D. P. (1997). Effects of sample size and nonnormality on the estimation of mediated effects in latent variable models. *Structural Equation Modeling, 4,* 87-107.

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.

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.C. Hoaglin, F. Mosteller, and J.W. Tukey (Eds.), *Exploring data tables, trends, and shapes,* (pp. 461-513). New York: Wiley

Hoekstra, R., Kiers, H. A. L, & Johnson, A. (2012). Are assumptions of well-known statistical techniques checked, and why (not)? *Frontiers in Psychology, 3*, 1-9.

Hu, L., Bentler, P.M., & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin, 112,* 351-362.

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

Keselman, H. J., Algina, J., Lix, L. M., Wilcox, R. R., & Deering, K. (2008). A generally robust approach for testing hypotheses and setting confidence intervals for effect sizes. *Psychological Methods, 13*, 110-129.

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. A., Donahue, B., Kowalchuk, R. K., …, & 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 revisited: An update based on trimmed means. *Psychometrika, 63*, 145–163.

Lix, L. M., & Keselman, H. J. (1998). To trim or not to trim: Tests of location equality under heteroscedasticity and nonnormality. . Educational and Psychological Measurement, , 409-42.

Lix, L. M., Keselman, J. C., & Keselman, H. J. (1996). Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance F test. *Review of Educational Research, 66*, 579–619.

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

Muthén, B. (1989). Multiple-group structural modeling with non-normal continuous variables. *British Journal of Mathematical and Statistical Psychology, 42*, 55–62.

Muthén, B., & Kaplan, D. (1992). A comparison of some methodologies for the factor analysis of non-normal Likert variables: A note on the size of the model. *British Journal of Mathematical and Statistical Psychology, 45*, 19-30.

Olsson, U. H, Foss, T., Troye, S., & Howell, R. (2000). The performance of ML, GLS, and WLS

estimation in structural equation modeling under conditions of misspecification and nonnormality. *Structural Equation Modeling: A Multidisciplinary Journal, 7*, 557-595.

R Development Core Team. (2014). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria.

Rosseel, Y. (2012). `lavaan`: An R package for structural equation modeling. *Journal of Statistical Software*, *48* (2), 1–36. Retrieved from http://www.jstatsoft.org/v48/i02

Satorra, A. (1992). Asymptotic robust inferences in the analysis of mean and covariance structures. *Sociological Methodology*, *22*, 249–278.

Satorra, A.,& Bentler, P.M. (1988). Scaling corrections for chi-square statistics in covariance structure analysis. In American Statistical Association 1988. Proceedings of the Business and Economics Sections (pp. 308–313). Alexandria, VA: American Statistical Association.

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, *66*, 507-514.

Sorbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical and Statistical Psychology*, *27*, 229–239.

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

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

Wilcox, R. R. (1988). A new alternative to the ANOVA *F* and new results on James' second-order method. *British Journal of Mathematical and Statistical Psychology*, *41*, 109-117.

Wilcox, R. R. (1990a). Comparing the means of two independent groups. *Biometrical Journal*, *32*, 771-780.

Wilcox, R. R. (1990b). Comparing variances and means when distributions have non-identical shapes. *Communications in Statistics-Simulation and Computation*, *19*, 155-173.

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. (2017). *Introduction to robust estimation and hypothesis testing,* 4th ed. San Diego, CA: Academic Press.

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.

Yuan, K. H., & Bentler, P.M. (1997). Mean and covariance structure analysis: Theoretical and practical improvements. *Journal of the American Statistical Association, 92,* 767–774.

Yuan, K. H., & Bentler, P. M. (1999). F tests for mean and covariance structure analysis. *Journal of Educational and Behavioral Statistics, 3,* 225–243.