Journal of Modern Applied Statistical Methods

Volume 18 | Issue 2 Article 2

7-17-2020

Empirical Comparison of Tests for One-Factor ANOVA Under Heterogeneity and Non-Normality: A Monte Carlo Study

Diep Nguyen
University of South Florida, diepdrm@gmail.com

Eunsook Kim
University of South Florida, ekim3@usf.edu

Yan Wang
University of Massachusetts, Yan_Wang1@uml.edu

Thanh Vinh Pham
University of South Florida, tvpham2@mail.usf.edu

Yi-Hsin Chen
University of South Florida, ychen5@usf.edu

See next page for additional authors

Recommended Citation

Nguyen, D., Kim, E., Wang, Y., Pham, T. V., Chen, Y.-H., & Kromrey, J. D. (2019). Empirical comparison of tests for one-factor ANOVA under heterogeneity and non-normality: A Monte Carlo study. Journal of Modern Applied Statistical Methods, 18(2), eP2906. doi: 10.22237/jmasm/1604190000

Empirical Comparison of Tests for One-Factor ANOVA Under Heterogeneity and Non-Normality: A Monte Carlo Study

Authors

Diep Nguyen, Eunsook Kim, Yan Wang, Thanh Vinh Pham, Yi-Hsin Chen, and Jeffrey D. Kromrey

Journal of Modern Applied Statistical Methods
November 2019, Vol. 18, No. 2, eP2906.
doi: 10.22237/jmasm/1604190000

בס"ד

ISSN 1538 − 9472

Empirical Comparison of Tests for One-Factor ANOVA Under Heterogeneity and Non-Normality: A Monte Carlo Study

Diep Nguyen
University of South Florida
Tampa, FL

Eunsook Kim
University of South Florida
Tampa, FL

Yan Wang
University of Massachusetts
Lowell, MA

Thanh Vinh Pham
University of South Florida
Tampa, FL

Yi-Hsin Chen
University of South Florida
Tampa, FL

Jeffrey D. Kromrey
University of South Florida
Tampa, FL

Although the Analysis of Variance (ANOVA) *F* test is one of the most popular statistical tools to compare group means, it is sensitive to violations of the homogeneity of variance (HOV) assumption. This simulation study examines the performance of thirteen tests in one-factor ANOVA models in terms of their Type I error rate and statistical power under numerous (82,080) conditions. The results show that when HOV was satisfied, the ANOVA F or the Brown-Forsythe test outperformed the other methods in terms of both Type I error control and statistical power even under non-normality. When HOV was violated, the Structured Means Modeling (SMM) with Bartlett or SMM with Maximum Likelihood was strongly recommended for the omnibus test of group mean equality.

Keywords: Analysis of variance, homogeneity, heterogeneity, non-normality, type I error control, statistical power

Introduction

Analysis of variance (ANOVA) is a common method used to compare the means of several groups. Although there are many statistical tests for ANOVA, none are suitable for every research situation (Lix et al., 1996). The traditional ANOVA *F* test is one of the most common statistical procedures to test the equality of several independent group means (Tomarken & Serlin, 1986). However, the *F* test is sensitive to violations of the homogeneity of variance (HOV) assumption (Rogan & Keselman, 1977). Several alternative tests (described below) have been

doi: 10.22237/jmasm/1604190000 | Accepted: July 1, 2018; Published: July 17, 2020.
Correspondence: Diep Nguyen, diepnguyen@usf.edu

suggested in response to this problem and can be classified into two groups: a group of tests using an ANOVA-type approach and a group of tests using a Structured Means Modeling (SMM) framework (Sörbom, 1974).

It was shown in simulation studies these alternatives can control the Type I error rates, given that the data are normally distributed and sample size is sufficiently large, even though population variances are heterogeneous. However, these tests become liberal when data are non-normal and heterogeneous (e.g., Fan & Hancock, 2012; Wilcox, 1988). Harwell et al. (1992) used meta-analytic methods to review 28 simulation studies on the ANOVA *F* test, Welch (1947) and the nonparametric Kruskal-Wallis (Kruskal & Wallis, 1952) tests when the homogeneity of variance and normality assumptions were not met. Although the sensitivity of the *F* and the Kruskal-Wallis tests to unequal variances was highlighted, it was also reported even with equal sample sizes the Welch test with two independent groups (Welch, 1947) was only robust to variance heterogeneity in the case of nearly normal population distributions.

Lix et al. (1996) extended the study of Harwell et al. (1992) with the addition of the Brown-Forsythe and James' second-order tests, as well as examining the Welch (1951) test for ANOVA models instead of the Welch (1947) test. Employing meta-analytic techniques to quantitatively assess alternative ANOVA tests under non-normality and heterogeneity of variances, they provided guidelines for applied researchers regarding under which data-analytic conditions a specific test should be used. Lix et al. (1996) confirmed the *F* test is not the test of choice when the variances are unequal, especially in combination with unequal group sizes. Both the Welch (1951) and James' second-order tests outperformed the *F*, Brown-Forsythe, and Kruskal-Wallis tests under the violation of HOV and normality assumptions. The two tests should only be bypassed when the population is moderately to highly skewed, and, in the case of the Welch test, also when total sample size is small, or one group size is very small.

Fan and Hancock (2012) examined eleven approaches to compare several independent group means. They investigated the performance of five ANOVA-based tests and six Robust Means Modeling (RMM) tests. The ANOVA-based methods included the *F* test and its alternatives (i.e., Welch test, Brown-Forsythe test, James' second-order test, and Alexander-Govern test). The RMM tests are based on SEM framework in which equal variances across groups are not assumed. Fan and Hancock included the maximum likelihood (ML) estimation method (T_{ML}), asymptotic distribution free (ADF) estimation method (TADF), Satorra and Bentler's (1988) rescaled test statistics (TSB), Yuan and Bentler's estimation methods (1997, 1999) that make corrections to T_{ADF} for small sample sizes (T_{YB1}

and T_{YB2}), and Bartlett's correction to the ML test statistic (T_{BC}). Results of Fan and Hancock (2012) showed that even though both ANOVA-based and RMM approaches provided reasonable control of Type I error rates under normal distributions, the RMM approaches were superior to the ANOVA-type tests under asymmetric non-normality and heterogeneous variances. Although the focus of Fan and Hancock was on introducing and examining the RMM approach, their study did not cover other available statistics such as the weighted least squares approach, the multilevel model with heterogeneous variances, and the Wilcox test.

As highlighted in Fan and Hancock's (2012), it is important to have guidelines on selecting an appropriate approach for their research scenarios, but there is a lack of extensive studies that investigate available test statistics for between-subjects ANOVA. Therefore, the purpose of this study is to examine the performance of thirteen available approaches to test the equality of several independent group means in terms of Type I error control and statistical power under various experimental situations. The test statistics investigated in this study are: ANOVA *F* test, Alexander-Govern test, Brown-Forsythe test, James' second-order test, Welch test, Weighted Least Squares test, Wilcox-centered test, SMM approach with Maximum Likelihood (ML) estimation, SMM approach with asymptotic distribution free (ADF) estimation, SMM with Bartlett's correction to the ML test statistic, SMM with Yuan and Bentler 1 (SMM with YB1), SMM with Yuan and Bentler 2 (SMM with YB2), and multilevel modeling approach (i.e. PROC MIXED in SAS). This simulation study includes comprehensive conditions with design factors that cover a variety of possible research situations.

Statistical Methods for Testing Mean Differences

ANOVA *F* Test

The ANOVA *F* (also called OLS) test is a common statistical method to test the equality of several independent group means, is defined as:

$$F = \frac{\sum_j n_j \left(\bar{X}_j - \bar{X}\right)^2 \Big/ \left(J-1\right)}{\sum_j \left(n_j - 1\right) S_j^2 \Big/ \left(N-J\right)},$$

where $j = 1, 2, \ldots, J$ for groups, n_j, $\bar{X}j$, and S_j^2 are the size, mean, and variance of group j, respectively, and $\bar{X}$ is the grand mean. The F statistic follows the F

distribution with (J – 1) and (N – J) degrees of freedom. As mentioned above, the F test is sensitive to the violation of the homogeneity of variance assumption.

Alexander-Govern (AG) Test

Alexander-Govern's approximation test (Alexander & Govern, 1994) defined a weight (w_j) for each group by

$$w_j = \frac{1/S_j^2}{\sum_1^J 1/S_j^2},$$

were $S_j = S_{\bar{x}_j}$ is the standard error of group j. The variance-weighted estimate of the common mean (X^+) is calculated by: $X^+ = \sum_1^J w_j \bar{X}_j$. For each of J groups, the t statistic is defined as

$$t_j = \frac{\bar{X}_j - X^+}{S_j}.$$

t_j is distributed as Student's t with v_j (= n_j – 1) degrees of freedom. A normalizing transformation of t_j to get z_j is conducted by

$$z_j = c + \frac{\left(c^3 + 3c\right)}{b} - \frac{\left(4c^7 + 33c^5 + 240c^3 + 855c\right)}{\left(110b^2 + 8bc^4 + 1000b\right)},$$

where $a = v_j - .5$; $b = 48a^2$; $c = \left[a \ln\left(1 + t_j^2 / v_j\right)\right]^{1/2}$. z_j is used to calculate the A statistic by

$$A = \sum_1^J z_j^2.$$

A is distributed as chi-square with (J – 1) degrees of freedom. The Alexander-Govern test has been suggested in the study of Schneider and Penfield (1997) as the best alternative to the ANOVA F test in the case of heterogeneous variances based on its good control of Type I error and high statistical power.

Brown-Forsythe (BF) Test

The Brown-Forsythe test (Brown & Forsythe, 1974) is a modification of the ANOVA *F* test:

$$F^* = \frac{\sum_j n_j \left(\bar{X}_j - \bar{X}\right)^2}{\sum_j \left(1 - n_j / N\right) S_j^2}.$$

F^* has an *F*-distribution with $(J - 1)$ and f degrees of freedom, where f is defined by the Satterthwaite approximation

$$\frac{1}{f} = \sum_j c_j^2 \Big/ \left(n_j - 1\right) \quad \text{and} \quad c_j = \frac{\left(1 - n_j / N\right) S_j^2}{\sum_j \left(1 - n_j / N\right) S_j^2}.$$

Although the BF test is known to have reasonable control of Type I error rates in several conditions (e.g. various levels of skewed distributions), it is not as good as the Welch test or James' second-order test with large variance heterogeneity in balanced designs (Lix et al., 1996) and small sample sizes (Wilcox, 1988). The BF test is also reported with lower statistical power estimates than the Welch and James' second-order tests (Fan & Hancock, 2012).

James' Second Order Test

The test statistic for James' test (James, 1951) is defined as:

$$Q = \sum_j w_j \left(\bar{X}_j - X_w\right)^2,$$

where

$$w_j = \frac{n_j}{S_j^2} \quad \text{and} \quad X_w = \frac{\sum_j w_j \bar{X}_j}{\sum_j w_j}.$$

The obtained value of *Q* is compared to a carefully adjusted critical value of χ^2 with $(J - 1)$ degrees of freedom (James, 1951). Although several studies recommended this test due to its good performance, this test is also reported to have inadequate

control of Type I error in some cases of asymmetric non-normal distributions or small sample sizes (Lix et al., 1996; Wilcox, 1988).

Welch Test

Welch (1951) proposed a modification of the F test that assumes the populations are independent and normally distributed but does not require equal population variances. The test statistic is defined as

$$F' = \frac{\sum_j w_j \frac{\left(\bar{X}_j - \bar{X}'\right)^2}{J-1}}{1+\frac{2(J-2)}{J^2-1}\sum_j\left[\left(1-\frac{w_j}{u}\right)^2(n_j-1)\right]},$$

where

$$w_j = \frac{n_j}{S_j^2}, \quad u = \sum_j w_j, \quad \bar{X}' = \sum_j \frac{w_j \bar{X}_j}{u}.$$

The distribution of F can be approximated using $v_b = J - 1$ and

$$\frac{1}{v_w} = \left(\frac{3}{J^2-1}\right)\sum_j\left[\frac{\left(1-\frac{w_j}{u}\right)^2}{n_j-1}\right].$$

The Welch test has been known to be relatively robust to different degrees of variance heterogeneity when sample sizes are equal and fairly large. However, its Type I error control becomes inadequate in the cases of variance heterogeneity associated with very small and unequal group sizes under certain types of skewed data (Lix et al., 1996) as well as with unequal group sizes or large number of group (Wilcox, 1988).

Wilcox Test

The Wilcox method (Wilcox, 1988) was contrasted with James (1951) method. The author made an improvement (Wilcox, 1989) in their original test and its modification covers the following settings

$$D_j = n_j / S_j^2,$$
$$W_S = \sum D_j,$$
$$\tilde{Y} = \sum D_j \tilde{Y}_j / W_s,$$

where

$$\tilde{Y}_j = \frac{X_{n_j,j}}{n_j} + \sum_{i=1}^{n_j-1} \left(1 - \frac{1}{n_j}\right) \frac{X_{ij}}{\left(n_j + 1\right)}$$

and $i = 1, 2, ..., n_j$ for individuals. The null hypothesis is rejected when $H_m = \Sigma D_j(\tilde{Y}_j - \tilde{Y})^2$ exceeds the $(1 - \alpha)$ quantile of a chi-square distribution with $(J - 1)$ degrees of freedom. The Wilcox test has been shown to result in poor Type I error control if the population grand mean differs from zero (Hsiung et al., 1994). In the current study, thus, the test was conducted after grand mean centering in each sample and called as Wilcox-centered.

Weighted Least Squares (WLS)

This method weights each observation by the inverse of its variance (Montgomery & Peck, 1992):

$$w_j = \frac{1}{S_j^2},$$

where w_j and S_j^2 are the weight and sample variance for group j and then uses generalized least squares to minimize

$$\sum_{j=1}^{J} \sum_{i=1}^{n_j} w_j \left(X_{ij} - \bar{X}_j\right)^2.$$

Structured Means Modeling Approach with Maximum Likelihood Estimation (SMM with ML)

When the SMM approach is applied to the between-subjects testing of measured variable mean equality, indicator $\mathbf{X}$ can be expressed as $\mathbf{X} = \mathbf{v}_k + \boldsymbol{\delta}$ where $\mathbf{v}_k$ is a $p \times 1$ vector of intercept values, $\boldsymbol{\delta}$ is a $p \times 1$ vector of normal errors, and p is the number of observed variables. The null hypothesis is tested by constraining population means to be equivalent although still allowing for variances of $\boldsymbol{\delta}$ to be heterogeneous. Estimation within SMM can be handled by using maximum likelihood. The F_{ML} is the ML fit function. The test statistic T_{ML} is a function of F_{ML} as $T_{\mathrm{ML}} = (N - 1)\, F_{\mathrm{ML}}$, with degrees of freedom equal to $Jp(p + 3) / 2 - q$, where J is the number of groups, and q is the number of parameter estimates across all groups.

SMM Approach with Asymptotic Distribution Free (ADF) Estimation (SMM with ADF)

When the variables are continuous but not multivariate normally distributed, Browne (1982, 1984) proposed asymptotic distribution free estimation (ADF) for the covariance structure and Muthén (1989) expanded ADF including both mean and covariance structures. Using a Generalized Least Square (GLS)-type fit function, the ADF fit function is defined as

$$F_{\mathrm{ADF}} = \sum_{j=1}^{J} \left(\mathbf{s}_j - \boldsymbol{\sigma}_j\right)' \mathbf{W}_j^{-1} \left(\mathbf{s}_j - \boldsymbol{\sigma}_j\right)$$

where, for each group J, $\mathbf{s}_j$ is the combined vector consisting of p elements of the observed means ($\mathbf{s}_1$) and $p(p + 1) / 2$ elements of the variance covariance matrix ($\mathbf{s}_2$), $\boldsymbol{\sigma}_j$ is the model implied counterpart of $\mathbf{s}_j$, and $\mathbf{W}$ represents the ADF weight matrix as an estimator of the asymptotic covariance matrix of $\mathbf{s}$. When this fit function is multiplied by $2n$ (where n is the total sample size), it follows the chi-square distribution with $(J - 1)$ degrees of freedom.

SMM with Bartlett's Correction to the ML Test Statistic (SMM with Bartlett)

Bartlett (1950) suggested a correction to the ML test statistic, which is translated to

$$T_{\text{BC}} = \left(N - \frac{p}{3} - \frac{2\mathbf{m}}{3} - \frac{11}{6} \right) F_{\text{ML}},$$

with degrees of freedom = $Jp^* - q$, where $p^* = p(p + 3) / 2$; N = total sample size; p = number of observed variables, and q = number of parameters estimated across all groups. In the context of one-way ANOVA, the SMM model now only has one observed variable and no latent factor.

Yuan and Bentler

Yuan and Bentler (1997, 1999) suggested test statistics T_{YB1} and T_{YB2} that make corrections to T_{ADF} for small sample sizes. Specifically,

$$T_{\text{YB1}} = \frac{T_{\text{ADF}}}{1 + \frac{T_{\text{ADF}}}{N}},$$

where $T_{\text{ADF}} = (N - 1) / F_{\text{ADF}}$, which follows a central χ^2 distribution with the same model degrees of freedom as T_{ADF} (when H_0 is true). Their second modification to ADF appeals to the F distribution:

$$T_{\text{YB2}} = \frac{N - \left(Jp^* - q \right)}{\left(N - 1 \right)\left(Jp^* - q \right)} T_{\text{ADF}}$$

with numerator and denominator degrees of freedom of $(Jp^* - q)$ and $(N - (Jp^* - q))$, respectively, and $p^* = p(p + 3) / 2$. Both T_{YB1} and T_{YB2} are the two best performing tests based on results of Fan and Hancock (2012) and are included in this study. The other SMM-based test statistics are also recommended by Fan and Hancock because of their outperformance over the ANOVA-based approaches, especially when data are non-normal.

Multilevel Model with Heterogeneous Variances

A mixed model may be fit with unequal residual variances to analyze data from ANOVA designs with heterogeneous variances (Littell et al., 2006). The model for a one factor ANOVA design can be written as

$$X_{ij} = \mu + \alpha_j + \varepsilon_{ij}, \quad \text{where} \quad \varepsilon_{ij} \sim \mathrm{N}\left(0, \sigma_j^2\right).$$

In this ANOVA model, variance for each group is estimated separately and may be fitted using ML or restricted ML (REML) estimation. The SAS procedure PROC MIXED provides a straightforward approach for fitting such a model. In this procedure, the heterogeneous variance solution is obtained by selecting the GROUP = option on the REPEATED statement (although a repeated-measures design is not used). Thus,

$$\text{REPEATED/group} = \text{IV},$$

where IV is the name of independent variable. For such analyses, the Satterthwaite degrees of freedom estimate should be used (Satterthwaite, 1946). This is obtained using the DDFM = SATTERTHWAITE option on the MODEL statement in PROC MIXED.

Method

A simulation was conducted to control and manipulate design factors, which included: number of groups (4 and 6), average number of observations per group (10 and 20), sample size pattern (*N*-pattern; see Table 1), variance pattern (described in Table 2), mean pattern (equal, progressive, one extreme, and split), maximum group variance ratio (1:1, 4:1, 8:1, and 16:1), effect size (0, .10, .25, and .4), and population shape ($\gamma 1 = 0.00$ and $\gamma 2 = 0.00$, $\gamma 1 = 1.00$ and $\gamma 2 = 3.00$, $\gamma 1 = 1.50$ and $\gamma 2 = 5.00$, $\gamma 1 = 2.00$ and $\gamma 2 = 6.00$, $\gamma 1 = 0.00$ and $\gamma 2 = 25.00$, and $\gamma 1 = 0.00$ and $\gamma 2 = -1.00$, where $\gamma 1$ and $\gamma 2$ represent skewness and kurtosis, respectively). Non-normal populations were generated by implementing Fleishman's transformation (Fleishman, 1978). Tables 1 and 2 show sample size pattern and variance pattern simulation factors, respectively, in detail.

There were four mean patterns: (1) equal pattern mean where all population means were equal; (2) progressive with all population means equally spaced; (3) one extreme where one mean differed from the others, (4) split where half the group means were equal to each other but different from the other half.

Table 1. Sample size patterns

		Sample size							
		Progressive *N*		**Equal *N***		**Split *N***		**One extreme**	
K=6	1	5	10	10	20	5	10	6	12
	2	7	14	10	20	5	10	6	12
	3	9	18	10	20	5	10	6	12
	4	11	22	10	20	15	30	6	12
	5	13	26	10	20	15	30	6	12
	6	15	30	10	20	15	30	30	60
Average *N*		10	20	10	20	10	20	10	20
K=4	1	7	14	10	20	5	10	6	12
	2	9	18	10	20	5	10	6	12
	3	11	22	10	20	15	30	6	12
	4	13	26	10	20	15	30	22	44
Average *N*		10	20	10	20	10	20	10	20

Note: *K* = number of groups, Progressive *N* = progressive increase of sample size, Split *N* = half of groups has the same sample size

The performance of the thirteen ANOVA approaches was examined at three nominal alpha levels: .01, .05, and .10. For effect size = 0 (i.e., null or Type I error conditions), there were 144 (2*3*4*6) conditions with equal variances and 2,592 (2*3*4*6*3*6) conditions with variance heterogeneity. For effect size = .10, .25, and .40 (i.e., power conditions), there were 1,296 (2*3*4*3*3*6) homogeneous conditions and 23,328 (2*3*4*6*3*3*3*6) heterogeneous conditions. Thus, there were a total of 82,080 simulation conditions across three alpha levels in this study. Type I error control and statistical power were evaluated as the simulation outcomes. For Type I error, we further investigated robustness using Bradley's (1978) liberal criterion. This criterion is set at 0.5α around nominal alpha. For instance, a test is considered robust when the Type I error rate falls between .025 (= 0.5*.05) and .075 (= 1.5*.05) at alpha level of .05. Finally, eta-squared analyses were conducted to explore the significant impacts of design factors on variability in the estimated Type I error and statistical power. Cohen's (1992) moderate effect size of .058 was set as a cutoff value for eta-squared analyses.

Table 2. Variance patterns

		Population variances									
		Progressive			Split			One extreme			Equal
Max variance ratio		1:4	1:8	1:16	1:4	1:8	1:16	1:4	1:8	1:16	1:1
K=6	1	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	2	1.6	2.4	4.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	3	2.2	3.8	7.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	4	2.8	5.2	10.0	4.0	8.0	16.0	1.0	1.0	1.0	1.0
	5	3.4	6.6	13.0	4.0	8.0	16.0	1.0	1.0	1.0	1.0
	6	4.0	8.0	16.0	4.0	8.0	16.0	4.0	8.0	16.0	1.0
K=4	1	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	2	2.0	3.3	6.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	3	3.0	5.7	11.0	4.0	8.0	16.0	1.0	1.0	1.0	1.0
	4	4.0	8.0	16.0	4.0	8.0	16.0	4.0	8.0	16.0	1.0
		Progressive inv.			Split inv.			One extreme inv.			Equal
Max variance ratio		1:4	1:8	1:16	1:4	1:8	1:16	1:4	1:8	1:16	1:1
K=6	1	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	2	1.6	2.4	4.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	3	2.2	3.8	7.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	4	2.8	5.2	10.0	4.0	8.0	16.0	1.0	1.0	1.0	1.0
	5	3.4	6.6	13.0	4.0	8.0	16.0	1.0	1.0	1.0	1.0
	6	4.0	8.0	16.0	4.0	8.0	16.0	4.0	8.0	16.0	1.0
K=4	1	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	2	2.0	3.3	6.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	3	3.0	5.7	11.0	4.0	8.0	16.0	1.0	1.0	1.0	1.0
	4	4.0	8.0	16.0	4.0	8.0	16.0	4.0	8.0	16.0	1.0

Note: For example, “Progressive” means that the population variances increased in a progressive way among groups; “Progressive inv.” refers to the same variance patterns as in “Progressive” but in the reverse group order

Data Sources

Continuous data were generated using a random number generator, RANNOR in SAS/IML statistical software, using a different seed value for each execution of the program. For each condition in the simulation, 5,000 samples were generated. The use of 5,000 replications aimed to reach a maximum standard error of an observed proportion (e.g., Type I error rate estimate) of .003, and a 95% confidence interval no larger than $\pm$ .006 (Robey & Barcikowski, 1992).

Results

The simulation results for the performance of all thirteen methods are presented in two sections regarding Type I error control and statistical power. In each section, we examined these tests under homogeneous conditions (where population variances were equal) and heterogeneous conditions (i.e., unequal population variances). Because we observed a similar pattern across the three nominal alpha levels (α = .01, .05, and .10), we present only the results at the nominal level of .05.

Type I Error Rate Estimates with Homogeneous Conditions

Presented in Figure 1 are the boxplots of the Type I error rate distributions across all simulation conditions with equal variances at the nominal alpha level of .05. Under the homogeneous conditions, the ANOVA *F* test (i.e., OLS) and the BF test showed the best performance. Among the other approaches, SMM with Bartlett, Wilcox-centered and SMM with ML controlled Type I error adequately.

Presented in Table 3 are the Type I error rates of all methods by three significant design factors. Because this study includes many design factors, we only present selected design factors that are substantially related to the variability of Type I error rates based on the eta-squared analyses: method (η^2 = .45), shape (η^2 = .13), and *N*-pattern (η^2 = .08). As observed in Table 3, the OLS and BF controlled Type I error around .05 across all conditions under the homogeneity of variance assumption. Type I error rates of WLS and SMM with ADF, on the contrary, were almost always above .07. The Wilcox-centered test showed reasonable Type I error control for all conditions under the equal variances except one condition of severe non-normality (skewness = 2, kurtosis = 6) and one extreme sample size pattern. For the SMM methods, Bartlett, and ML controlled Type I error reasonably except two conditions of severe non-normality (skewness = 2, kurtosis = 6) and unequal sample sizes. A very similar pattern was observed with James, Welch, and AG tests. As shown in Figure 2, the Type I error rates were inflated when the population shape was severely non-normal.

The proportion of conditions that satisfied Bradley's liberal criterion was calculated for each method at the alpha level of .05. Similar with the results presented in Figure 1, the ANOVA *F* test (OLS) and the BF were the most robust with all conditions meeting Bradley's criterion. Following were the SMM with Bartlett, Wilcox-centered, and SMM with ML methods with satisfied proportions of nearly 94%, 94%, and 92%, respectively, among all homogeneous conditions. The next good performers in terms of Type I error control were James, Welch, YB1, AG, and YB2 tests with proportions meeting Bradley's liberal criteria of 85%, 82%,

81%, 80%, and 70%, respectively. The test with poorest performance in controlling Type I error rates were ADF and WLS with 28% and 20% of conditions satisfied Bradley's criterion.

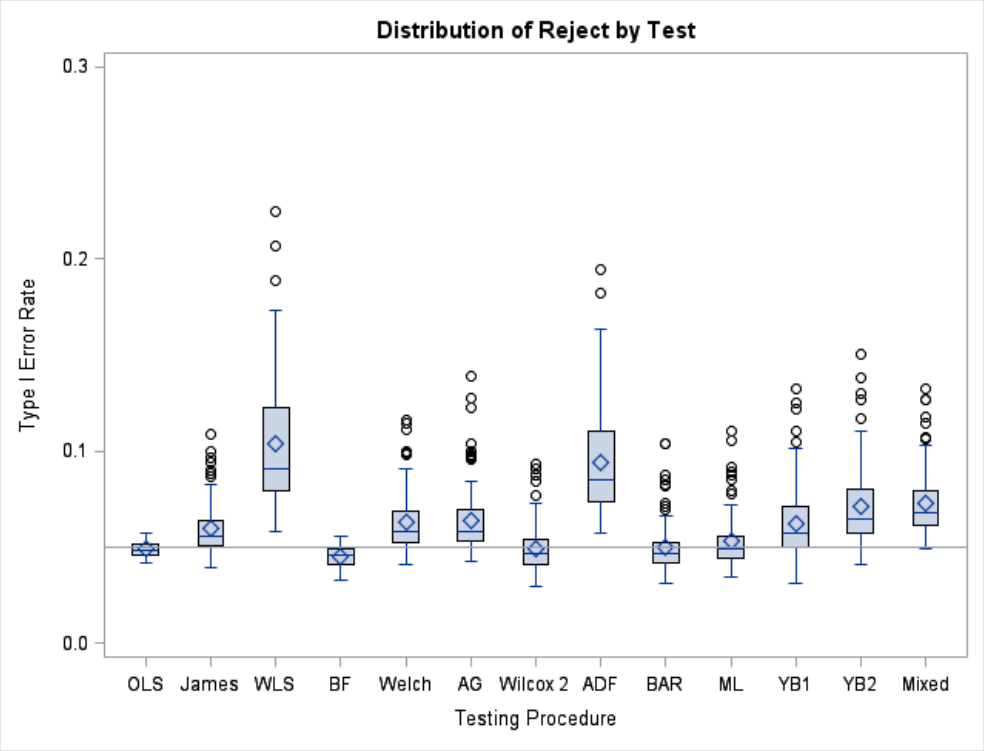

Figure 1. Distributions of Type I error estimates of the thirteen ANOVA tests with homogeneous conditions; OLS = ANOVA *F* test using ordinary least squares; James = James' second-order; WLS = Weighted Least Squares; BF = Brown-Forsythe; AG = Alexander-Govern; Wilcox2 = Wilcox-centered; ADF = SMM approach with asymptotic distribution free estimation; BAR = structured mean modeling with Bartlett's correction to the maximum likelihood test statistic; ML = structured mean modeling with maximum likelihood estimation; YB1 = structured mean modeling with Yuan and Bentler 1; YB2 = structured mean modeling with Yuan and Bentler 2; Mixed = multilevel modeling in SAS

Table 3. Type I error rates of thirteen robust ANOVA tests by selected simulation factors at nominal alpha of .05 with homogeneous conditions

		Method												
Shape	**N_pattern**	**OLS**	**James**	**WLS**	**BF**	**Welch**	**AG**	**Wilcox2**	**ADF**	**BAR**	**ML**	**YB1**	**YB2**	**Mixed**
1(0, 0)	equal	**0.05**	**0.05**	**0.07**	**0.05**	**0.05**	**0.05**	**0.04**	**0.07**	**0.04**	**0.04**	**0.04**	**0.05**	**0.05**
	one extreme	**0.05**	**0.05**	0.11	**0.05**	**0.05**	**0.05**	**0.04**	0.09	**0.04**	**0.04**	**0.06**	**0.07**	**0.07**
	split	**0.05**	**0.05**	0.10	**0.05**	**0.06**	**0.05**	**0.04**	0.09	**0.04**	**0.04**	**0.06**	**0.07**	**0.07**
	progressive	**0.06**	**0.06**	0.09	**0.05**	**0.06**	**0.06**	**0.05**	0.08	**0.05**	**0.05**	**0.05**	**0.06**	**0.07**
2(1, 3)	equal	**0.05**	**0.05**	**0.07**	**0.04**	**0.05**	**0.05**	**0.04**	**0.07**	**0.04**	**0.04**	**0.04**	**0.05**	**0.05**
	one extreme	**0.05**	**0.05**	0.11	**0.04**	**0.05**	**0.06**	**0.04**	0.09	**0.04**	**0.04**	**0.06**	**0.07**	**0.07**
	split	**0.05**	**0.05**	0.10	**0.04**	**0.05**	**0.06**	**0.04**	0.09	**0.04**	**0.04**	**0.06**	**0.07**	**0.07**
	progressive	**0.05**	**0.05**	0.08	**0.04**	**0.05**	**0.05**	**0.04**	0.08	**0.04**	**0.04**	**0.05**	**0.05**	**0.06**
3(1.5, 5)	equal	**0.04**	**0.05**	**0.07**	**0.04**	**0.05**	**0.05**	**0.04**	**0.07**	**0.04**	**0.04**	**0.04**	**0.05**	**0.05**
	one extreme	**0.05**	**0.06**	0.12	**0.04**	**0.06**	**0.07**	**0.05**	0.11	**0.05**	**0.06**	**0.07**	0.08	0.08
	split	**0.05**	**0.05**	0.11	**0.04**	**0.06**	**0.06**	**0.05**	0.10	**0.04**	**0.05**	**0.06**	**0.07**	**0.07**
	progressive	**0.05**	**0.05**	0.09	**0.04**	**0.06**	**0.06**	**0.04**	0.08	**0.05**	**0.05**	**0.05**	**0.06**	**0.07**
4(2, 6)	equal	**0.04**	**0.06**	0.09	**0.04**	**0.06**	**0.07**	**0.05**	0.09	**0.06**	**0.06**	**0.06**	**0.07**	**0.07**
	one extreme	**0.05**	0.09	0.16	**0.04**	0.10	0.12	0.08	0.15	0.09	0.10	0.11	0.12	0.11
	split	**0.05**	0.09	0.15	**0.04**	0.09	0.10	**0.07**	0.13	0.08	0.08	0.10	0.11	0.10
	progressive	**0.05**	**0.07**	0.11	**0.04**	**0.07**	0.08	**0.06**	0.11	**0.06**	**0.07**	**0.07**	0.08	0.08
5(0, 25)	equal	**0.05**	**0.06**	0.08	**0.05**	**0.06**	**0.06**	**0.05**	0.08	**0.05**	**0.05**	**0.05**	**0.06**	**0.06**
	one extreme	**0.05**	**0.07**	0.12	**0.05**	**0.07**	**0.07**	**0.06**	0.11	**0.05**	**0.06**	0.08	0.09	0.09
	split	**0.05**	**0.07**	0.12	**0.04**	**0.07**	**0.07**	**0.06**	0.11	**0.06**	**0.06**	0.08	0.09	0.09
	progressive	**0.05**	**0.06**	0.09	**0.05**	**0.06**	**0.06**	**0.05**	0.09	**0.05**	**0.05**	**0.06**	**0.07**	**0.07**
6(0, –1)	equal	**0.05**	**0.06**	0.08	**0.05**	**0.06**	**0.06**	**0.04**	0.08	**0.05**	**0.05**	**0.05**	**0.06**	**0.06**
	one extreme	**0.05**	**0.06**	0.12	**0.05**	**0.07**	**0.07**	**0.06**	0.11	**0.05**	**0.05**	**0.07**	0.08	0.08
	split	**0.05**	**0.07**	0.12	**0.05**	**0.07**	**0.07**	**0.06**	0.10	**0.05**	**0.05**	**0.07**	0.08	0.09
	progressive	**0.05**	**0.06**	0.09	**0.05**	**0.06**	**0.06**	**0.05**	0.09	**0.05**	**0.05**	**0.06**	**0.06**	0.09

Note: The Type I error rates meeting the Bradley's criterion are in bold; Progress = Progressive sample size pattern

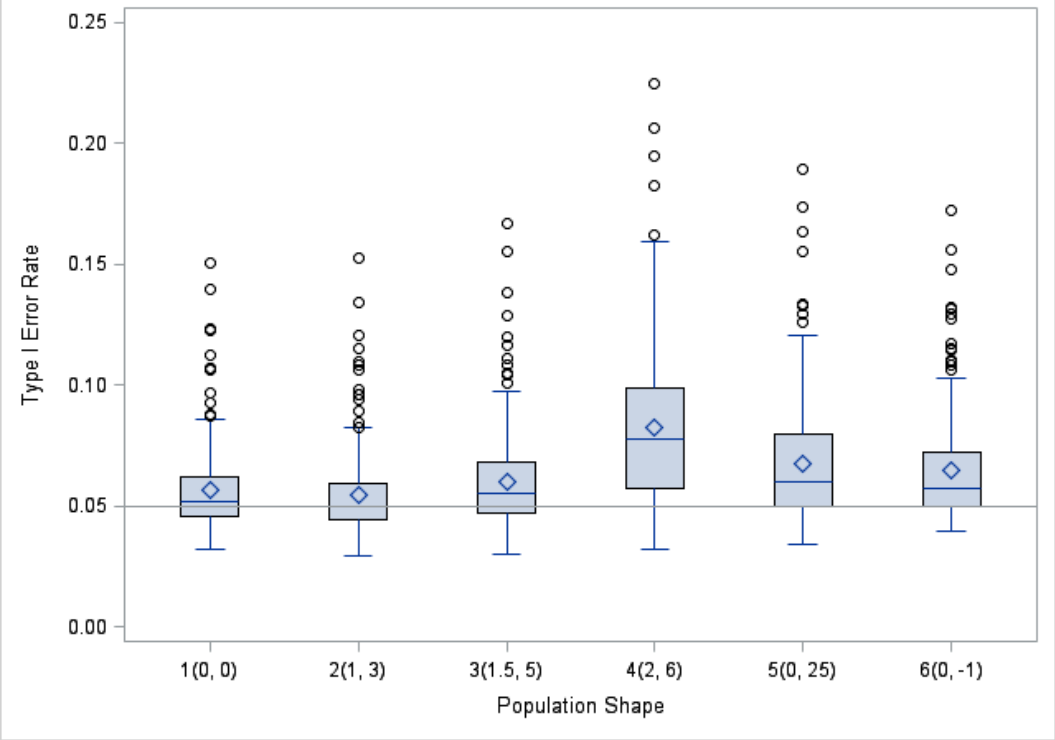

Figure 2. Distributions of Type I error rates of the thirteen ANOVA tests by population shape for homogeneous conditions; for population shapes, the values within parentheses are skewness and kurtosis, respectively; for example, 1(0, 0) indicates normal distribution with skewness = 0 and kurtosis = 0

Type I Error Rate Estimates with Heterogeneous Conditions

Under the heterogeneous conditions, the OLS method showed poor performance as expected. The Wilcox-centered, SMM with Bartlett and SMM with ML provided the best overall Type I error control as shown in Figure 3.

As shown in Figure 3, the Type I error rates of WLS and SMM with ADF were substantially high across all simulation conditions of heterogeneous variance. Similarly, the ANOVA *F* test (OLS) showed poor performance in controlling for Type I error: over control (or being conservative) when the large groups had the large variance and under control (or being liberal) when the large groups had the small variance. This phenomenon became more serious as the variance disparity across groups increased. In general, the SMM methods except the ADF showed adequate Type I error control on average. Particularly, the SMM with Bartlett and SMM with ML outperformed the other robust ANOVA tests. However, even these

best performing methods yielded inflated Type I error rates when the population shape was severely non-normal (i.e., skewness = 2, kurtosis = 6) in combination with the reversed variance patterns (i.e., the large group with the small variance). Following the SMM with Bartlett and SMM with ML, the Wilcox-centered and James controlled Type I error adequately. The Welch, AG, and the BF were the next good performers in terms of controlling for Type I error but showed increased Type I error rates (.08) when the variance heterogeneity was severe.

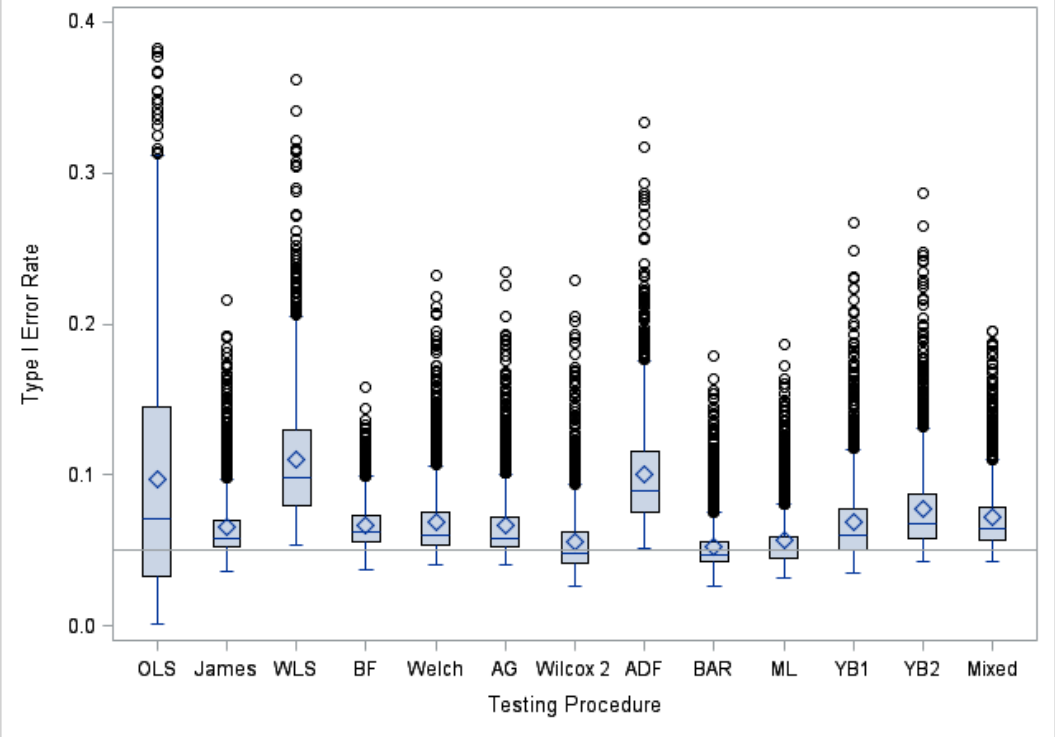

Figure 3. Distributions of Type I error estimates of the thirteen ANOVA tests for heterogeneous conditions; OLS = ANOVA *F* test using ordinary least squares; James = James' second-order; WLS = Weighted Least Squares; BF = Brown-Forsythe; AG = Alexander-Govern; Wilcox2 = Wilcox-centered; ADF = SMM approach with asymptotic distribution free estimation; BAR = structured mean modeling with Bartlett's correction to the maximum likelihood test statistic; ML = structured mean modeling with maximum likelihood estimation; YB1 = structured mean modeling with Yuan and Bentler 1; YB2 = structured mean modeling with Yuan and Bentler 2; Mixed = multilevel modeling in SAS

Table 4. Type I error estimates by variance pattern with heterogeneous conditions

	Method												
Variance pattern	**OLS**	**James**	**WLS**	**BF**	**Welch**	**AG**	**Wilcox2**	**ADF**	**BAR**	**ML**	**YB1**	**YB2**	**Mixed**
Extreme	**0.05**	**0.06**	0.10	0.08	**0.06**	**0.06**	**0.05**	0.09	**0.05**	**0.05**	**0.06**	**0.07**	**0.06**
Split	**0.04**	**0.06**	0.10	**0.07**	**0.06**	**0.06**	**0.05**	0.09	**0.05**	**0.05**	**0.06**	**0.07**	**0.07**
Progress	**0.03**	**0.06**	0.10	**0.06**	**0.06**	**0.06**	**0.05**	0.09	**0.05**	**0.05**	**0.06**	**0.07**	**0.07**
Extreme-R	0.16	**0.07**	0.11	0.08	**0.07**	**0.07**	**0.06**	0.10	**0.05**	**0.06**	**0.07**	0.08	**0.07**
Split-R	0.18	**0.07**	0.13	**0.06**	0.08	**0.07**	**0.07**	0.11	**0.06**	**0.06**	0.08	0.09	0.08
Progress-R	0.13	**0.07**	0.13	**0.05**	0.08	0.08	**0.07**	0.12	**0.06**	**0.06**	0.08	0.09	0.08

Note: The Type I error rates meeting the Bradley's criterion are in bold; Extreme-R = One Extreme Inversely, Split-R = Split Inversely, and Progress-R = Progress Inversely (see Table 2 for more details); OLS = ANOVA *F* test using ordinary least squares; James = James' second-order; WLS = Weighted Least Squares; BF = Brown-Forsythe; AG = Alexander-Govern; Wilcox2 = Wilcox-centered; ADF = SMM approach with asymptotic distribution free estimation; Bartlett = structured mean modeling with Bartlett's correction to the maximum likelihood test statistic; ML = structured mean modeling with maximum likelihood estimation; YB1 = structured mean modeling with Yuan and Bentler 1; YB2 = structured mean modeling with Yuan and Bentler 2; Mixed = multilevel modeling in SAS

The eta-squared analysis showed that variation in the Type I error rates was associated with the method (η^2 = .21), method and variance pattern interaction (η^2 = .15), population shape (η^2 = .14), and variance pattern (η^2 = .08). Table 4 presents the impact of variance pattern on the method in terms of Type I error control. The best performing methods (i.e., SMM with Bartlett, SMM with ML, and Wilcox-centered) controlled Type I error around .05 across all variance patterns. In addition, when groups were balanced (i.e., equal group size), all the tests but OLS, WLS, and SMM with ADF showed adequate Type I error on average. Similar to the homogeneous variance conditions, as the population shape departed from the normality, the Type I error inflation was more serious (see Figure 4).

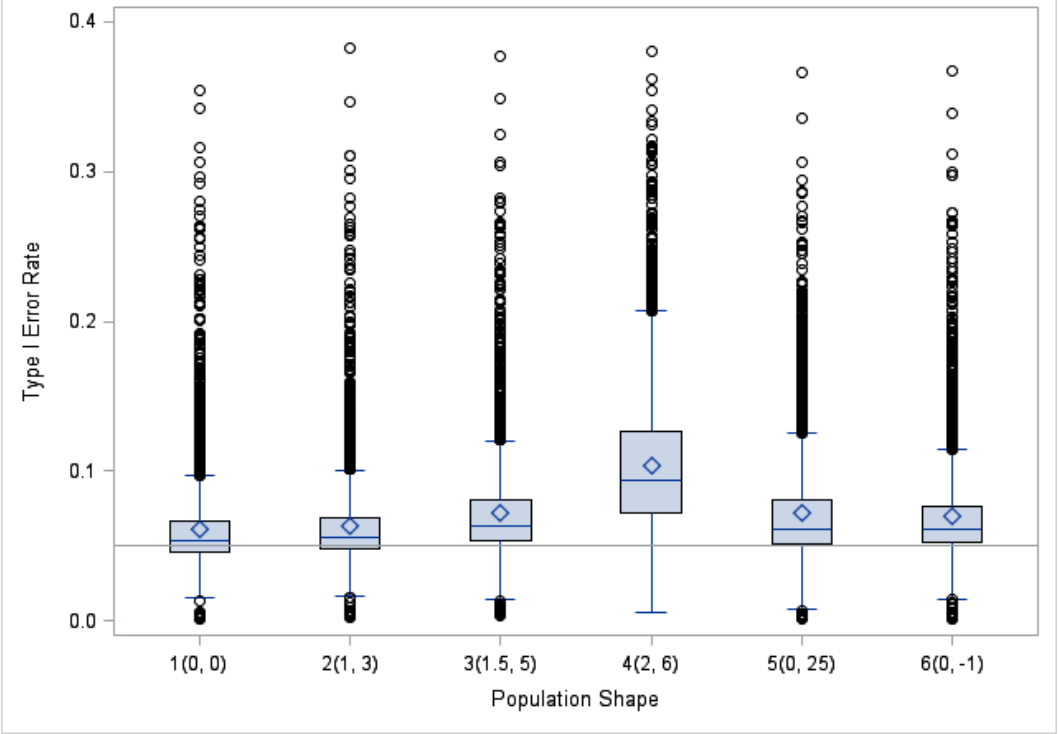

Figure 4. Distributions of Type I error rates of the thirteen ANOVA tests by population shape for heterogeneous conditions; for population shapes, the values within parentheses are skewness and kurtosis, respectively; for example, 1(0, 0) indicates normal distribution with skewness = 0 and kurtosis = 0

The proportions of simulation conditions with heterogeneous variances meeting the Bradley's criterion for Type I error rate were investigated. The SMM with Bartlett test showed the best performance (.90); followed by the SMM with ML (.88) and the Wilcox-centered (.84). In addition to three aforementioned methods, the James, Welch, AG, and BF tests had the improved proportions of .75 or higher that met the Bradley's criterion. The YB1 and YB2 had 73% and 62% conditions satisfied Bradley's criterion. The WLS, SMM with ADF tests, and OLS had the lowest proportions that met the Bradley's criterion for Type I error control with only 17%, 25%, and 36%, respectively.

Statistical Power with Homogeneous and Heterogeneous Conditions

Statistical power was estimated for the methods that provided adequate Type I error control across most conditions. Therefore, the ANOVA *F* (OLS), BF, Wilcox-centered, SMM with Bartlett, and SMM with ML methods were included in the power analysis under homogeneous conditions; the Wilcox-centered, SMM with Bartlett, and SMM with ML methods were included under heterogeneous conditions.

Regarding the power estimates under homogeneous conditions, the OLS, BF, SMM with Bartlett, SMM with ML, and Wilcox-centered all had relatively low power on average (.35, .34, .34, .34, and .33, respectively) with substantial variations within each method. The variations in power estimates were attributable to effect size ($\eta^2 = .57$), mean pattern ($\eta^2 = .10$), group size ($\eta^2 = .09$), and interaction of mean pattern and number of groups ($\eta^2 = .06$), based on eta-squared analyses.

Table 5. Power estimates by effect size, group size, and mean pattern for homogeneous conditions

	Effect size			Group size		Mean pattern		
	0.10	**0.25**	**0.40**	**10**	**20**	**Progressive**	**Partial null**	**Multiple null**
OLS	0.08	0.32	0.63	0.26	0.43	0.23	0.45	0.36
BF	0.08	0.31	0.61	0.25	0.42	0.22	0.44	0.35
Wilcox2	0.09	0.31	0.60	0.25	0.42	0.22	0.44	0.34
Bartlett	0.10	0.32	0.61	0.24	0.43	0.22	0.44	0.34
ML	0.09	0.31	0.60	0.26	0.43	0.23	0.45	0.35

Note: OLS = ANOVA *F* test using ordinary least squares; BF = Brown-Forsythe; Wilcox2 = Wilcox-centered; Bartlett = structured mean modeling approach with Bartlett estimation; ML = structured mean modeling approach with maximum likelihood estimation; Progressive = all means equally spaced; Partial Null = one extreme mean differing from the others; Multiple Null = half group means were equal but different from the other half

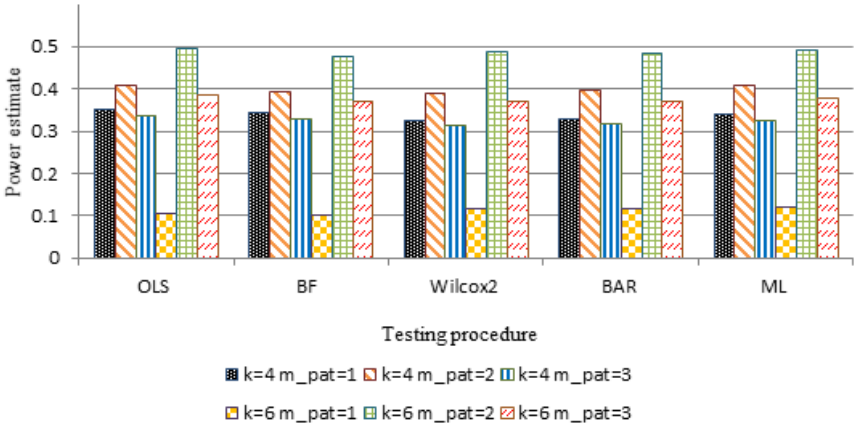

Figure 5. Power estimates by mean pattern and number of groups for homogeneous conditions; OLS = ANOVA *F* test using ordinary least squares; BF = Brown-Forsythe; Wilcox2 = Wilcox-centered; BAR = structured mean modeling approach with Bartlett estimation; ML = structured mean modeling approach with maximum likelihood estimation; m_pat = mean pattern; k = number of group; for mean pattern: m_pat = 1: progressive = all means equally spaced; m_pat = 2: partial null = one extreme mean differing from the others; m_pat = 3: multiple null = half group means were equal but different from the other half

---

Presented in Table 5 are power estimates by three significant design factors independently and Figure 5 shows the impact of the interaction between mean pattern and number of groups on power estimates. Power estimates of all five methods increased substantially as effect size increased and with large effect size (.40), power estimates reached .60 to .63. Larger group size would also lead to greater power estimates (e.g. .26 and .43 for group size 10 and 20, respectively, for OLS). Power estimates were much higher when the mean pattern is partial null (.44 - .45 for four methods), compared with progressive (.22 - .23) and multiple null (.34 - .36) mean patterns. However, the significant role of mean pattern in power estimates depended on the number of groups as shown in Figure 6. When mean pattern was partial null and combined with number of groups of 6, power estimates were the highest. On the other hand, large number of groups (i.e. 6 groups) associated with progressive mean pattern yielded lowest power for homogeneous conditions.

Similar to results under homogeneous conditions, average powers of Wilcox-centered, SMM with Bartlett, and SMM with ML were all relatively low (.39, .39,

and .40, respectively) when the variances were not equal. Substantial variations in power estimates were observed as well. Based on eta-squared analyses results, effect size (η^2 = .47), interaction of variance pattern and mean patter (η^2 = .09), group size (η^2 = .06), and interaction of number of groups and mean pattern (η^2 = .06) were associated with variation in power estimates across all three methods.

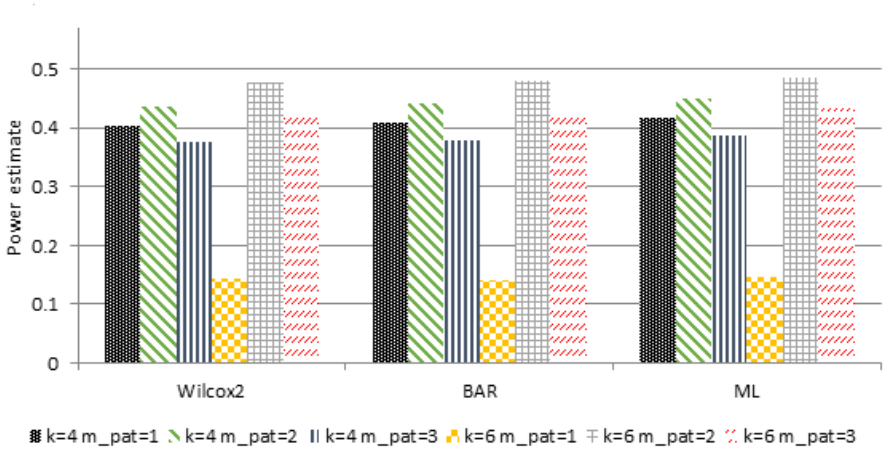

Figure 6. Power estimates by mean pattern and number of group for heterogeneous conditions; Wilcox2 = Wilcox-centered; BAR = structured mean modeling approach with Bartlett estimation; ML = structured mean modeling approach with maximum likelihood estimation; m_pat = mean pattern; k = number of group; for mean pattern: m_pat = 1: progressive = all means equally spaced; m_pat = 2: partial null = one extreme mean differing from the others; m_pat = 3: multiple null = half group means were equal but different from the other half

Table 6. Power estimates by effect size, group size, and mean pattern for heterogeneous conditions

	Effect size			Group size		Mean pattern		
	0.10	**0.25**	**0.40**	**10**	**20**	**Progressive**	**Partial null**	**Multiple null**
Wilcox2	0.11	0.39	0.66	0.30	0.48	0.27	0.48	0.40
Bartlett	0.12	0.40	0.67	0.31	0.48	0.28	0.49	0.41
ML	0.12	0.39	0.66	0.31	0.47	0.27	0.48	0.40

Note: Wilcox2 = Wilcox-centered; Bartlett = structured mean modeling approach with Bartlett estimation; ML = structured mean modeling approach with maximum likelihood estimation; Progressive = all means equally spaced; Partial Null = one extreme mean differing from the others; Multiple Null = half group means were equal but different from the other half

Table 7. Power estimates by variance pattern with mean pattern for heterogeneous conditions

	Variance pattern																	
	2			3			4			5			6			7		
	Mean pattern																	
	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
Wilcox2	0.25	0.22	0.48	0.25	0.35	0.36	0.25	0.34	0.37	0.30	0.60	0.49	0.30	0.66	0.34	0.29	0.61	0.37
Bartlett	0.26	0.24	0.49	0.27	0.36	0.39	0.26	0.35	0.39	0.29	0.60	0.48	0.29	0.65	0.31	0.27	0.60	0.35
ML	0.27	0.24	0.50	0.27	0.37	0.40	0.27	0.36	0.40	0.30	0.61	0.48	0.30	0.66	0.32	0.28	0.61	0.35

Note: Wilcox2 = Wilcox-centered; Bartlett = structured mean modeling approach with Bartlett estimation; ML = structured mean modeling approach with maximum likelihood estimation; for variance pattern: 2 = one extreme, 3 = split, 4 = progressive, 5 = one extreme inversely, 6 = split inversely, 7 = progressive inversely; for mean pattern: 1 = progressive = all means equally spaced, 2 = partial null = one extreme mean differing from the others, 3 = multiple null = half group means were equal but different from the other half

As presented in Table 6, similar to the pattern identified under homogeneous conditions, larger effect size and group size would lead to higher power estimates and the partial null mean pattern yielded the highest power among all mean patterns. Comparing Tables 5 and 6, we found that power estimates were slightly higher under heterogeneous conditions than homogeneous conditions (e.g., .12, .40, and .67 versus .10, .32, and .61 under effect size of .10, .25, and .40 for SMM with Bartlett, respectively). Similar to the findings under homogeneous conditions, the combination of the large number of groups (6) and the partial null mean pattern yielded the highest power for Wilcox-centered (.48), SMM with Bartlett (.48), and SMM with ML (.49), as shown in Figure 6. In addition, the partial null mean pattern combined with all three inverse variance patterns produced highest power estimates (.61 - .67). Interestingly, although the partial null mean pattern overall led to highest power estimates, this mean pattern associated with one extreme variance pattern yielded lowest power estimates (.23). All the power estimates for combinations of variance pattern and mean pattern are presented in Table 7.

Discussion

The performance of the thirteen robust ANOVA tests were studied under various simulation conditions. In addition to the traditional robust ANOVA (i.e., ANOVA-based) tests, the study examined the performance of SMM with different types of estimation methods. As found in Fan and Hancock (2012), the SMM methods, with the exception of ADF, performed relatively well compared to the ANOVA-based methods. Interestingly, among the SMM tests, the ML and its correction (i.e., Bartlett) outperformed the ADF and its corrections (i.e., SMM with YB1 and SMM with YB2). Although the assumption of normality underlies the ML, this study showed that the ML was fairly robust to the violation of this assumption. Thus, if the assumption was not severely violated, the ML controlled for Type I error reasonably. Even in the case of severe nonnormality, the performance of ML was not worse than that of many other methods. Consistent with the findings of Nevitt and Hancock (2004), the SMM with the Bartlett correction led to better Type I error control than the SMM with the ML estimation and performed best among the thirteen methods, particularly in small samples under the heterogeneity of variance.

It was somewhat surprising the SMM with the ADF estimation failed to control for Type I error even with homogeneous variance conditions. Because the SMM with the ADF estimation does not assume the normality of the outcome variable, superior performance of the ADF was expected under nonnormality (West et al., 1995). However, the ADF showed high Type I error rates across simulation

conditions in this study. As mentioned in the study of Curran et al. (1996), the ADF requires a large sample for the inverse of the weight matrix. Thus, this estimation method is possibly unfeasible with small samples such as what we investigated in this study (i.e., maximum average group size of 20). As suggested in the literature (e.g., Nevitt & Hancock, 2004; Yuan & Bentler, 1997), the two corrected estimation methods of the ADF for small samples (i.e., the YB1 and YB2) showed notably improved Type I error control. The SMM with YB1 slightly outperformed the SMM with YB2 across simulation conditions. Applied researchers using the SMM with ADF, SMM with YB1, and SMM with YB2 to test the group mean equality should be aware that these methods require at least 4 observations for each group and are expected to perform reasonably with large sample.

Under the heterogeneous conditions, we observed the interaction effect between variance pattern and sample size pattern on Type I error rates, which is well recognized as positive pairing and negative pairing in the ANOVA literature (e.g., Harwell et al., 1992; Lix et al., 1996). This interaction was more evident with the ANOVA F test (OLS) as the variance heterogeneity increased. That is, when the large group had the small variance (negative pairing), the tests became more liberal, yielding inflated Type I error rates. When the relation between variance and sample size patterns was reversed (i.e., large group with large variance or positive pairing), the OLS test became slightly conservative, showing over control of Type I error rates. We also confirmed that when group sizes were equal, Type I error was notably better controlled. Type I error rates in many robust tests were around the nominal level under balanced conditions even with heterogeneity of variance (Boneau, 1960). Thus, it is recommended that applied researchers pay attention to the pairing of group size and variance when comparing means across groups.

In summary, when homogeneity of variance was satisfied, the ANOVA F test using OLS and the BF test outperformed the other methods in terms of both Type I error control and power. Type I error rates of this test were not affected by other design factors with all conditions meeting Bradley's criterion, even under the severe nonnormality and unbalanced group sizes. The OLS or the BF, therefore, should be a choice when the variances are equal across groups. When homogeneity of variance was violated, the SMM with Bartlett or ML are strongly recommended for the omnibus test of group mean equality. When the average group size is 10 or above, the Wilcox-centered test and the James' second-order test can also be considered. However, it should be noted that even these best performing tests yielded inflated Type I error rates when the distribution was severely non-normal under heterogeneity of variance, although the Type I error rates of the Bartlett, ML, and Wilcox-centered were still lower than those of the other methods. It should also

be noted that, with the exception of the well-performing methods, nonnormality and unequal group sizes resulted in an increase in Type I error rates above the upper limit of Bradley's liberal criterion, even under homogeneous variance conditions. In addition, applied researchers should keep in mind that the maximum group size of this study was 20 and the performance of some methods could improve with larger group sizes (e.g., the SMM with ADF-based estimation methods).

As a final remark, no one test fits all (Lix et al., 1996). Thus, it is strongly recommended that researchers understand their data such as the degree of nonnormality, severity of heterogeneity, and pairing with group size for an informed decision of optimal tests for independent means tests (Lix et al., 1996).

References

Alexander, R. A., & Govern, D. M. (1994). A new and simpler approximation for ANOVA under variance heterogeneity. *Journal of Educational Statistics, 19*(2), 91-101. doi: 10.3102/10769986019002091

Bartlett, M. S. (1950). Tests of significance in factor analysis. *British Journal of Statistical Psychology, 3*(2), 77-85. doi: 10.1111/j.2044-8317.1950.tb00285.x

Boneau, C. A. (1960). The effects of violations of assumptions underlying the t test. *Psychological Bulletin, 57*(1), 49-64. doi: 10.1037/h0041412

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 34*(2), 144-152. doi: 10.1111/j.2044-8317.1978.tb00581.x

Brown, M. B., & Forsythe, A. B. (1974). The small sample behavior of some statistics which test the equality of several means. *Technometrics, 16*(1), 129-132. doi: 10.1080/00401706.1974.10489158

Browne, M. W. (1982). Covariance structures. In D. M. Hawkins (Ed.), *Topics in applied multivariate analysis* (pp. 72-141). Cambridge, UK: Cambridge University Press. doi: 10.1017/CBO9780511897375.003

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology, 37*(1), 62-83. doi: 10.1111/j.2044-8317.1984.tb00789.x

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155-159. doi: 10.1037/0033-2909.112.1.155

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*(1), 16-29. doi: 10.1037/1082-989X.1.1.16

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*(1), 137-156. doi: 10.3102/1076998610396897

Fleishman, A. I. (1978). A method for simulating non-normal distributions. Psychometrika, 43(4), 521-532. doi: 10.1007/BF02293811

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 and Behavioral Statistics, 17*(4), 315-339. doi: 10.3102/10769986017004315

Hsiung, T., Olejnik, S., & Huberty, C. J. (1994). Comment on a Wilcox test statistic for comparing means when variances are unequal. *Journal of Educational and Behavioral Statistics, 19*(2), 111-118. doi: 10.3102/10769986019002111

James, G. S. (1951). The comparison of several groups of observations when the ratios of the population variances are unknown. *Biometrika, 38*(3/4), 324-329. doi: 10.2307/2332578

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association, 47*(260), 583-621. doi: 10.1080/01621459.1952.10483441

Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., & Schabenberger, O. (2006). *SAS for mixed models* (2nd edition). Cary, NC: SAS Institute Inc.

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*(4), 579-619. doi: 10.3102/00346543066004579

Montgomery, D. C., & Peck, E. A. (1992). *Introduction to linear regression analysis* (2nd edition). New York: John Wiley & Sons, Inc.

Muthén, B. (1989). Multiple-group structural modeling with non-normal continuous variables. *British Journal of Mathematical and Statistical Psychology, 42*(1), 55-62. doi: 10.1111/j.2044-8317.1989.tb01114.x

Nevitt, J., & Hancock, G. (2004). Evaluating small sample approaches for model test statistics in structural equation modeling. *Multivariate Behavioral Research, 39*(3), 439-478. doi: 10.1207/S15327906MBR3903_3

Robey, R. R., & Barcikowski, R. S. (1992). Type I error and the number of iterations in Monte Carlo studies of robustness. *British Journal of Mathematical and Statistical Psychology, 45*(2), 283-288. doi: 10.1111/j.2044-8317.1992.tb00993.x

Rogan, J. C., & Keselman, H. J. (1977). Is the ANOVA F-test robust to variance heterogeneity when sample sizes are equal? An investigation via a coefficient of variation. *American Educational Research Journal, 14*(4), 493-498. doi: 10.3102/00028312014004493

Satorra, A., & Bentler, P. M. (1988). Scaling corrections for chi-square statistics in covariance structure analysis. In *Proceedings of the American statistical association* (Vol. 1, pp. 308-313). Alexandria, VA: American Statistical Assocation.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin, 2*(6), 110-114. doi: 10.2307/3002019

Schneider, P. J., & Penfield, D. A. (1997). Alexander and Govern's approximation: Providing an alternative to ANOVA under variance. *The Journal of Experimental Education, 65*(3), 271-286. doi: 10.1080/00220973.1997.9943459

Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical and Statistical Psychology, 27*(2), 229-239. doi: 10.1111/j.2044-8317.1974.tb00543.x

Tomarken, A. J., & Serlin, R. C. (1986). Comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures. *Psychological Bulletin, 99*(1), 90. doi: 10.1037/0033-2909.99.1.90

Welch, B. L. (1947). The generalization of 'Student's' problem when several different population variances are involved. *Biometrika, 34*(1/2), 28-35. doi: 10.2307/2332510

Welch, B. L. (1951). On the comparison of several means: An alternative approach. *Biometrika, 38*(3/4), 330-336. doi: 10.2307/2332579

West, S. G, Finch, J. F., & Curran, P. J. (1995). Structural equation models with nonnormal variables: Problems and remedies. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 56-75). Thousand Oaks, CA: Sage.

Wilcox, R. R. (1988). A new alternative to the ANOVA F and new results on James's second-order method. *British Journal of Mathematical and Statistical Psychology, 41*(1), 109-117. doi: 10.1111/j.2044-8317.1988.tb00890.x

Wilcox, R. R. (1989). Adjusting for unequal variances when comparing means in one-way and two-way fixed effects ANOVA models. *Journal of Educational and Behavioral Statistics, 14(*3), 269-278. doi: 10.3102/10769986014003269

Yuan, K. H., & Bentler, P. M. (1997). Mean and covariance structure analysis: Theoretical and practical improvements. *Journal of the American Statistical Association, 92*(438), 767-774. doi: 10.1080/01621459.1997.10474029

Yuan, K. H., & Bentler, P. M. (1999). *F* tests for mean and covariance structure analysis. *Journal of Educational and Behavioral Statistics, 24*(3), 225-243. doi: 10.3102/10769986024003225