Comparing Performance of Tests for One-factor ANOVA Models under Heterogeneity and Non-normality: a Monte Carlo Simulation Study

Thanh V. Pham, Eun Sook Kim, Yi-Hsin Chen, Yan Wang, Diep Nguyen, Jeffrey D. Kromrey
University of South Florida

Abstract

The Analysis of Variance (ANOVA) F test is one of the most common statistical methods to test group mean equivalence. However, it is sensitive to the violation of the assumption of homogeneity of variance. Several alternative tests have been developed in response to this problem of the ANOVA F test. These tests can be classified into two groups: tests using ANOVA-type approach and tests using Structured Means Modeling (SMM) technique. This simulation study examines the performance of fourteen available tests in one-factor ANOVA models in terms of their Type I error rate and statistical power under comprehensive conditions (total of 48,384), especially, under the violation of the assumption of homogeneity of variance. The results show that when the assumption of equal variance was satisfied, the ANOVA F test with Ordinary Least Square (OLS) excelled the other methods in terms of both Type I error control and power. When the homogeneity assumption was violated, the Brown-Forsythe, the SMM with Bartlett, and SMM with Maximum Likelihood tests are strongly recommended for the omnibus test of group mean equality.

Key words: Analysis of variance, Homogeneity, Heterogeneity, Non-normality, Type I error control, Statistical power, Structured Means Modeling.

1. Introduction

The traditional analysis of variance (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 assumption (Rogan & Keselman, 1977). Several alternative tests (described below) have been suggested in response to this problem. Simulation studies have shown that these alternatives can control the Type I error rates when data are normally distributed and population variances are heterogeneous. However, these tests become liberal when data are non-normal and heterogeneous (Fan & Hancock, 2012).

A different approach that does not require the assumption of the homogeneity of variance applies the technique called Structured Means Modeling (SMM). The SMM technique is developed from structural equation modeling (SEM) that allows group variances to be heterogeneous by freely estimating them. Moreover, various estimation methods robust to the violation of normality such as the Asymptotic Distribution Free (ADF) estimation (Browne, 1982) are available in addition to the maximum likelihood (ML) estimation in SEM (Fan & Hancock, 2012). Fan and Hancock (2012) showed that the SMM based tests performed better than ANOVA based tests in term of power and Type I error rate.

As highlighted in Fan and Hancock's (2012) study, it is important for applied researchers to have guidelines on selecting an appropriate approach for their research scenarios. However there was lack of extensive studies that investigate all test statistics for between-subject ANOVA. Therefore the purpose of this paper is to examine the performance of fourteen 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. This simulation study includes comprehensive conditions (more than 48,000) with numerous design factors that cover a variety of possible research situations such as several population shapes, various levels of variance heterogeneity, mean patterns, effect sizes, and variance ratios.

2. Theoretical Framework: Statistical Methods for Testing Mean Differences

2.1. ANOVA F Test

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

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

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.

2.2. Alexander and Govern test

Alexander and Govern approximation test (Alexander & Govern, 1994) defines a weight $\left(\mathrm{w_j}\right)$ for each group by $w_j = \frac{1/S_j{}^2}{\sum_1^J 1/S_j{}^2}$ where S_j is the standard error of group j. The variance-weighted estimate of the common mean ($\mathrm{X^+}$) 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. Normalizing transformation of t_j to get z_j by:

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

Where $a = v_j - .5$; $b = 48a^2$; $c = [a \ln(1 + t_j{}^2/v_j)]^{1/2}$. 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.

2.3. Brown-Forsythe (BF) test

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

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

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

$$\frac{1}{f} = \sum_j C_j{}^2/(n_j - 1)$$

and $c_j = \frac{(1-n_j/N)S_j{}^2}{\sum_j(1-n_j/N)S_j{}^2}$

2.4. James' second order test

The test statistic for James' test is defined as:

$$Q = \sum_{j=1}^{J} w_j \left(\bar{X}_j - X_w\right)^2$$

where $w_j = \frac{n_j}{S_j^2}$ and $X_w = \sum_{j=1}^{J} w_j \bar{X}_j / \sum_{j=1}^{J} w_j$. The obtained value of *Q* is compared to a carefully adjusted critical value of χ^2 with (*J*– 1) degrees of freedom (James, 1951).

2.5. 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 $w_j = n_j/s_j{}^2$; $u = \sum_j w_j$; $\bar{X}' = \sum_j \frac{w_j \bar{X}_j}{u}$. The distribution of *F* can be approximated using $v_B = J - 1$, and $\frac{1}{v_w} = (\frac{3}{J^2-1})\sum_j[\frac{\left(1-\frac{w_j}{u}\right)^2}{n_j-1}]$.

2.6. Wilcox test

The Wilcox method (Wilcox, 1988) was contrasted with James (1951) method. The author made an improvement (Wilcox, 1989) in his original test and its modification covers the following setting:
$D_j = n_j/{s_j}^2$,
$W_s = \sum D_j$,
$\tilde{Y} = \sum D_j \tilde{Y}_j / W_s$,

where i = 1, 2, ... , N for individuals, $\tilde{Y}_j = X_{n_j j}/n_j + \sum_{i=1}^{n_j-1}\left(1 - \frac{1}{n_j}\right) X_{ij}/(n_j + 1)$. The null hypothesis is rejected when $H_m = \sum D_j(\tilde{Y}_j - \tilde{Y})^2$ exceeds the (1 – α) 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, Olejnik, & Huberty, 1994). In this study, the test was conducted in both its improved formulation (Wilcox, 1989) and after grand mean centering in each sample (named as Wilcox 2 in the Results section).

2.7. Weighted Least Squares (WLS)

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

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

and then uses generalized least squares to minimize

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

2.8. Structured Mean Modeling (SMM) approach with Maximum Likelihood (ML) estimation (SMM with ML)

When the SMM approach is applied to the between-subject testing of measured variable mean equality, indicator x can be expressed as $x = v_k + \delta$ where v_k is a p x 1 vector of intercept values, δ is a p x 1 vector of normal errors. The null hypothesis is tested by constraining population means to be equivalent while still allowing for variances of δ 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_{ML} = (N-1) F_{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.

2.9. 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 Generalize Least Square-type fit function, the ADF fit function is defined as

$$F_{ADF} = 1/2\sum_{j=1}^{J}\left(s_j - \sigma_j\right)' W_j^{-1}(s_j - \sigma_j)$$

where for each group J, s_j is the combined vector consisting of p elements of the observed means (s_1) and $p(p+1)/2$ elements of the variance covariance matrix (s_2), σ_j is the model implied counterpart of s_j, and W represents the ADF weight matrix as an estimator of the asymptotic covariance matrix of 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.

2.10. SMM with Bartlett's correction to the ML test statistic (Bartlett)

Bartlett (1950) suggested a correction to the ML test statistic which is translated to:
$T_{BC} = (N-p/3-2m/3-11/6)\ F_{ML,}$
with degrees of freedom = Jp* - q; N= total sample size; p= number of observed variables, m= group's observed mean vector; q = number of parameters estimated across all groups.

2.11. 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,
$T_{\mathrm{YBI}} = \frac{T_{\mathrm{ADF}}}{1+\frac{T_{\mathrm{ADF}}}{\mathrm{N}}}$,
where T_{ADF} =(N-1) /F_{ADF} ,which follows a central $\chi 2$ 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:

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

with numerator and denominator degrees of freedom of (Jp* – q) and (N – (Jp* – q)), respectively. Both T_{YB1} and T_{YB2} are included in this study.

2.12. Multilevel model with heterogeneous variances

The SAS procedure PROC MIXED provides an elegant test for mean differences while adjusting for unequal variances. This heterogeneous variance solution is obtained with the "GROUP =" option on the REPEATED statement (even though a repeated-measures design is not used). That is,

repeated / group=IV;

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

3. Method

This study used a simulation approach, which allowed the control and manipulations of the design factors. The design factors included: number of groups (four and six groups), average number of observations per group (5, 10, and 20 observations per group), sample size pattern, variance pattern, mean pattern (equal, progressive, one extreme, and split), maximum group variance ratio (1, 4, 8 and 16), 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 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.

The performance of the fourteen ANOVA approaches was examined at nominal alpha levels: .01, .05, .10. This factorial design had a total of 48,384 (2 x 3 x 7x4 x 3 x 6 x 4 x 4) conditions.

Type I error rate 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-square analyses were conducted to explore the significant impacts of design factors on variability in the estimated Type I error. Cohen's (1992) moderate effect size of .05 was set as a cutoff value for eta-square analyses.

4. Data Sources

Continuous data for this study 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 provides a maximum standard error of an observed proportion (e.g., Type I error rate estimate) of .00158, and a 95% confidence interval no wider than ± .003 (Robey & Barcikowski, 1992).

5. Results

The simulation results for the performance of all fourteen 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 group variances were equal) and heterogeneous conditions (i.e., unequal group 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.

5.1. Type I Error Rate Estimates with Homogeneous Conditions

Boxplots were first examined to describe the distributions of Type I error rate estimates across all homogeneous conditions at each nominal alpha level. Figure 1 presents the boxplots of the rejection 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) showed the best performance. Among the other approaches, BF, Bartlett, and SMM with ML controlled Type I error adequately.

Table 3 presents the Type I error rates of all methods by three significant design factors. Because this study includes many design factors with 48,384 simulation conditions, we only present selected design factors that are substantially related to the variability of Type I error rates based on the eta-square analysis: method (η^2 = .32), method*group size (η^2 = .17), group size (η^2 = .11), *N*-pattern (η^2 = .06), and *N*-pattern*population shape (η^2 = .05). Note that the progressive, split, and one extreme *N*-patterns did not show a notable difference in terms of Type I error rates. Thus, only the equal *N*- and progressive *N*-patterns are presented. Also the average group size of 5 and 20 conditions are contrasted in Table 3.

As observed in Table 3 and Figure 2, the OLS and BF controlled Type I error around .05 across all conditions under the homogeneity of variance. On the contrary, the Type I error rates of WLS and ADF were almost always above .07. The Wilcox test showed reasonable Type I error control only when the group size was balanced (i.e., equal *N*-pattern), but did not work for unbalanced groups with the Type I error rates simply 1.00. For the SMM methods except the ADF (i.e., Bartlett, ML, YB1, and YB2), the Type I error rates were reasonably controlled even with the small group size. However, when the distribution was extremely leptokurtic (kurtosis = 25) and the group size was small, the Type I error rates were slightly inflated. Of note is that the ADF and corrected ADF (i.e., YB1 and YB2) required a minimum group size of 4 for the operation and thus did not run when the average group size was 5 and the groups were unbalanced (i.e., progressive, split, and one extreme *N*-patterns; see Table 1). The James, Welch, AG, Wilcox2, and Mixed methods failed to control for the Type I error rates when sample size was small and the groups were unbalanced. As shown in Figure 3, the Type I error inflation was more serious when the population shape was severely nonnormal.

According to Bradley's (1978) liberal criterion of robustness, the test is referred to as robust if its probability of Type I error falls within the range of .025 and .075 at the nominal α level of .05. Table 4 shows the proportion of conditions that satisfied Bradley's liberal criterion for each method at the alpha level of .05. Similar with the results presented in Figure 1, the ANOVA *F* test (OLS) was the most robust with all conditions meeting Bradley's criterion. Following were the BF, Bartlett and ML methods with satisfied proportions of nearly 98%, 88%, and 87%, respectively among all conditions. Note that for the ADF, YB1 and YB2, all homogeneous conditions including missing cases due to small group sizes were included in the computation of the Bradley's proportion (see the full dataset in Table 4).

In addition, we also examined the Bradley's robustness of the fourteen methods in the dataset without missing conditions (i.e. deleting all conditions where the three aforementioned methods did not produce estimates). When the simulated group size was at least 4 and the ADF, YB1, and YB2 yielded the Type I error rates, more methods demonstrated adequate Type I error control (see the subset of Table 4). Specifically, in addition to the OLS, BF, Bartlett, and ML methods, the James, Welch, AG, Wilcox2, and YB1 also witnessed acceptable results when the minimum group size was 4.

5.2. Type I Error Rate Estimates with Heterogeneous Conditions

Under the heterogeneous conditions, the OLS method showed poor performance as expected. The BF, Bartlett and ML provided the best overall Type I error control as shown in Figure 4.

The eta-square analysis showed that variation in the Type I error rates was associated with the method ($\eta^2 = .20$), method and group size interaction ($\eta^2 = .12$), group size ($\eta^2 = .11$), method and variance pattern interaction ($\eta^2 = .07$), variance pattern ($\eta^2 = .06$), and population shape ($\eta^2 = .06$). As observed in the homogeneous conditions, larger group size improved the Type I error control notably for some tests such as James, WLS, Welch, AG, Wilcox2, ADF, and Mixed (see Figure 5). Overall, the impact of variance pattern depended on group size as well as method (see Table 5). When the large variance was associated with the small groups (i.e., reversed variance patterns), Type I error was remarkably inflated. However, the best performing methods (i.e., Bartlett, ML, and BF) controlled Type I error around .05 across variance patterns. In addition, when groups were balanced (i.e., equal group size), all the tests but OLS, WLS, and 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.

As shown in Figure 5, the Type I error rates of WLS and 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 Bartlett and 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 nonnormal (i.e., skewness = 2, kurtosis =6 and skewness = 0, kurtosis = 25) in combination with the reversed variance patterns (i.e., the large group with the small variance). Again, the ADF, YB1, and YB2 did not run when the average group size was 5 and the groups were unbalanced because at least one group size was below 4. For the Wilcox, the Type I error rates were near or just 1.00 when the groups were unbalanced. Following the Bartlett and ML, the BF controlled Type I error adequately, but showed increased Type I error rates (.08 ~ .22) when the variance heterogeneity was severe with the one extreme or one extreme inversely pattern (16:1 or 1:16), the group size was small, and the distribution was nonnormal. For the James, Welch, AG, and Wilcox2, the Type I error rates increased even under the normality if the groups were unbalanced with a small group size. Among these robust ANOVA tests, the James performed slightly better than the others.

Finally, the proportions of simulation conditions with heterogeneous variances meeting the Bradley's criterion for Type I error rate are presented in Table 6. The SMM with Bartlett test showed the best performance (.85) followed by the SMM with ML (.83) and the BF (.78). Excluding the conditions in which the ADF, YB1, and YB2 did not yield the results of Type I error (i.e., the average group size was 5 and the groups were unbalanced), we observed notable improvement in Type I error control within the Bradley's liberal criterion for all the methods except that the BF kept the same high proportion. In addition to three aforementioned methods, the James, Welch, AG, Wilcox2, and SMM with YB1 had the improved proportions of .73 or higher that met the Bradley's criterion. Particularly, for the Wilcox2 test the proportions of simulation conditions meeting the Bradley's liberal criterion increased substantially from .62 to .81.

5.3. 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, SMM with Bartlett, and SMM with ML methods were included in the power analysis under homogeneous conditions; the BF, SMM with Bartlett, and SMM with ML methods were included under heterogeneous conditions. Figure 6 presents the boxplots of power estimates under homogeneous conditions. The OLS, BF, Bartlett, and ML all had relatively low power on average (.28, .26, .26, and .27, respectively), with substantial variations within each method.

The variations in power estimates were attributable to effect size ($\eta^2 = .40$), group size ($\eta^2 = .22$), effect size*group size ($\eta^2 = .11$), and mean pattern ($\eta^2 = .07$), based on eta-square analyses. Table 7 presents power estimates by three significant design factors independently and Figure 7 shows the impact of the interaction between effect size and group size on power estimates. Power estimates of all four methods increased

substantially as effect size increased and with large effect size (.40), power estimates reached .46-.51. Larger group size would also lead to greater power estimates (e.g., .14, .26, and .43 for group size 5, 10, and 20, respectively for OLS). Power estimates were much higher when the mean pattern is partial null (.34-.36 for four methods), compared with progressive (.18-.19) and multiple null (.26-.28) mean patterns. The significant role of effect size and group size was further supported by variations in power estimates due to their interaction effect, as shown in Figure 7. When effect size was .40 and group size was 20, power estimates were the highest (.75-.77 for four methods). With effect size of .25 and group size of 20 or effect size of .40 and group size of 10, power estimates reached between .40 and .50, whereas power was close to or below .20 for all other conditions.

Figure 8 shows the boxplots of power estimates under heterogeneous conditions. Similar to results under homogeneous conditions, on average powers of BF, Bartlett, and ML were all relatively low (.26, .26, and .27, respectively). Substantial variations in power estimates were observed as well. Based on eta-square analyses results, effect size ($\eta^2 = .35$), group size ($\eta^2 = .19$), effect size*group size ($\eta^2 = .08$), and mean pattern ($\eta^2 = .06$) were associated with variation in power estimates across all three methods.

As presented in Table 8, 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 7 and 8, we found that power estimates were slightly higher under heterogeneous conditions than homogeneous conditions (e.g., .10, .30, and .52 versus .08, .24, and .46 under effect size of .10, .25, and .40 for Bartlett). Not surprisingly, the combination of the largest effect size (.40) and the largest group size (20) yielded the highest power for BF (.74), Bartlett (.79), and ML (.79), as shown in Figure 9.

6. Discussion

This study investigated the performance of the fourteen robust ANOVA tests under various simulation conditions. In addition to the traditional robust ANOVA (i.e., ANOVA-based) tests, this study examined the performance of structured means modeling with different types of estimation methods. As found in Fan and Hancock (2012), the SMM methods except with the 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., YB1 and 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 fourteen methods, particularly in small samples under the heterogeneity of variance.

It was somewhat surprising that 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, Finch, & Curran, 1995). However, the ADF showed high Type I error rates across simulation conditions in this study. Because the ADF requires a large sample for the inverse of the weight matrix (Curran, West, & Finch, 1996), 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 YB1 slightly outperformed the YB2 across simulation conditions. Applied researchers using the SMM with ADF, YB1, and YB2 to test the group mean equality should be aware that these methods require at least 4 observations for each group.

Among many simulation factors examined in this study, the group size emerged as a primary factor related to the variability of Type I error and power rates. Generally, the increase of group size led to better control of Type I error, but the impact of group size on Type I error depended on the methods. For the well-performing methods such as the Bartlett, ML, and BF with both homogeneous and heterogeneous conditions and the OLS with homogeneous conditions, the Type I error rates were around the nominal level on average regardless of group size. On the other hand, for the other methods, as the group size increased, the Type I error rates were better controlled. Especially, the Type I error control of Wilcox2 improved notably in large samples and was comparable to that of the aforementioned well-performing methods.

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 paring and negative paring 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 in general became 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. We also confirmed that when group sizes were equal, Type I error was notably better controlled. Thus, in many robust tests Type I error rates 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 paring of group size and variance when comparing means across groups.

In summary, when homogeneity of variance was satisfied, the ANOVA *F* test using OLS excelled the other methods in terms of both Type I error control and power. Because the Type I error rates of OLS were not affected by the other design factors showing all conditions (100%) meeting the Bradley's liberal criterion even under the severe nonnormality and with unbalanced groups, this test should be a choice when the variances are equal across groups. When homogeneity of variance was violated, the BF and the SMM with Bartlett or ML are strongly recommended for the omnibus test of group mean equality. When the group size is large (at least 4 per group in this study or average group size 10 or above), the Wilcox2 test followed by 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 nonnormal under heterogeneity of variance although the Type I error rates of the Bartlett, ML, and BF were still lower than those of the other methods. It should also be noted that for many tests except the well-performing methods even with homogeneous conditions, nonnormality resulted in the increase in Type I error rates over the upper limit of the Bradley's liberal criterion. Also, 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).

Because homogeneity of variance is an important factor in the choice of an optimal test for the equality of group means, we suggest a conditional test of group means. That is, if the assumption of homogeneous variance is met, the ANOVA *F* test (OLS) is selected whereas the SMM with Bartlett, the SMM with ML, or BF is employed if the assumption

is violated. However, future research is called for to examine the performance of this conditional test. As a final remark, we reiterate what the robust ANOVA literature has found so far: 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 paring 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*, 91–101.

Bartlett, M. S. (1950). Tests of significance in factor analysis. *British Journal of Psychology: Statistical Section, 3*, 77–85.

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

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.

Browne, M. W. (1982). Covariance structures in DM Hawkins (Ed.), *Topics in applied multivariate analysis* (pp. 72–141).

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.

Cohen, J. (1992). A power primer. *Psychological bulletin*, *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*, 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*(1), 137-156.

Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika*, *43*(4), 521-532.

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

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

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.

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.

Montgomery, D. C., & Peck, E. A. (1992*). Introduction to Linear Regression Analysis* (2nd ed.). New York, NY: John Wiley & Sons, Inc.

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

Nevitt, J., & Hancock, G. (2004). Evaluating small sample approaches for model test statistics in structural equation modeling. *Multivariate Behavioral Research, 39*, 439–478.

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.

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.

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

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

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*, 109–117.

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.

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.

Tables and figures

Table 1

Sample Size Patterns

Sample Sizes												
	Progressive *N*			Equal *N*			Split *N*			One Extreme		
K=6												
1	2	5	10	5	10	20	2	5	10	3	6	12
2	3	7	14	5	10	20	2	5	10	3	6	12
3	4	9	18	5	10	20	2	5	10	3	6	12
4	6	11	22	5	10	20	8	15	30	3	6	12
5	7	13	26	5	10	20	8	15	30	3	6	12
6	8	15	30	5	10	20	8	15	30	15	30	60
Average N	5	10	20	5	10	20	5	10	20	5	10	20
K=4												
1	2	7	14	5	10	20	2	5	10	3	6	12
2	4	9	18	5	10	20	2	5	10	3	6	12
3	6	11	22	5	10	20	8	15	30	3	6	12
4	8	13	26	5	10	20	8	15	30	11	22	44
Average N	5	10	20	5	10	20	5	10	20	5	10	20

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

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	1	1	1	1	1	1	1	1	1
2	1.6	2.4	4	1	1	1	1	1	1	1
3	2.2	3.8	7	1	1	1	1	1	1	1
4	2.8	5.2	10	4	8	16	1	1	1	1
5	3.4	6.6	13	4	8	16	1	1	1	1
6	4	8	16	4	8	16	4	8	16	1
K=4										
1	1	1	1	1	1	1	1	1	1	1
2	2	3.3	6	1	1	1	1	1	1	1
3	3	5.7	11	4	8	16	1	1	1	1
4	4	8	16	4	8	16	4	8	16	1

(Cont'd)

Population Variances									
	Progressive Inversely			Split Inversely			One Extreme Inversely		
Max	4:1	8:1	16:1	4:1	8:1	16:1	4:1	8:1	16:1

Variance Ratio									
K=6									
1	4	8	16	4	8	16	4	8	16
2	3.4	6.6	13	4	8	16	1	1	1
3	2.8	5.2	10	4	8	16	1	1	1
4	2.2	3.8	7	1	1	1	1	1	1
5	1.6	2.4	4	1	1	1	1	1	1
6	1	1	1	1	1	1	1	1	1
K=4									
1	4	8	16	4	8	16	4	8	16
2	3	5.7	11	4	8	16	1	1	1
3	2	3.3	6	1	1	1	1	1	1
4	1	1	1	1	1	1	1	1	1

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

Table 3

Type I Error Rates of Fourteen Robust ANOVA Tests by Selected Simulation Factors with Homogeneous Conditions at Nominal Alpha of .05

Shape	Group size	*N*-pattern	Test												
			OLS	James	WLS	BF	Welch	AG	Wilcox	Wilcox2	ADF	BAR	ML	YB1	YB2	Mixed
(0,0)	5	Equal	**.05**	**.05**	.13	**.04**	**.06**	**.05**	**.05**	**.05**	.12	**.03**	**.04**	**.04**	**.06**	**.07**
		Progress	**.05**	.09	.24	**.05**	.11	.11	1.00	.13	.	**.04**	**.05**	.	.	.13
	20	Equal	**.05**	**.05**	**.07**	**.05**	**.05**	**.05**	**.04**	**.04**	**.06**	**.04**	**.04**	**.04**	**.05**	**.05**
		Progress	**.06**	**.06**	.08	**.05**	**.06**	**.06**	1.00	**.04**	**.07**	**.05**	**.05**	**.05**	**.06**	**.07**
(1,3)	5	Equal	**.05**	**.04**	.12	**.04**	**.04**	**.04**	**.04**	**.04**	.11	**.02**	**.03**	**.03**	**.05**	**.05**
		Progress	**.05**	**.07**	.22	**.04**	.08	.09	1.00	.10	.	**.04**	**.04**	.	.	.10
	20	Equal	**.05**	**.05**	**.07**	**.05**	**.05**	**.05**	**.04**	**.04**	**.07**	**.05**	**.05**	**.05**	**.05**	**.06**
		Progress	**.05**	**.05**	**.07**	**.04**	**.05**	**.05**	1.00	**.04**	**.07**	**.04**	**.05**	**.05**	**.05**	**.06**
(1.5, 5)	5	Equal	**.04**	**.04**	.12	**.03**	**.04**	**.05**	**.04**	**.04**	.11	**.02**	**.03**	**.03**	**.05**	**.06**
		Progress	**.05**	**.07**	.22	**.04**	.09	.09	1.00	.11	.	**.03**	**.04**	.	.	.11
	20	Equal	**.05**	**.05**	**.07**	**.04**	**.05**	**.05**	**.04**	**.04**	**.06**	**.04**	**.04**	**.04**	**.05**	**.05**
		Progress	**.05**	**.06**	.08	**.05**	**.06**	**.06**	1.00	**.05**	.08	**.05**	**.05**	**.06**	**.06**	**.07**
(2,6)	5	Equal	**.04**	**.05**	.14	**.02**	**.05**	**.06**	**.05**	**.05**	.13	**.04**	**.05**	**.04**	**.06**	**.06**
		Progress	**.05**	.10	.27	**.04**	.12	.15	1.00	.14	.	**.07**	.08	.	.	.14
	20	Equal	**.04**	**.06**	.08	**.04**	**.06**	**.06**	**.05**	**.05**	.08	**.06**	**.06**	**.06**	**.06**	**.07**
		Progress	**.05**	**.07**	.10	**.04**	**.07**	**.07**	1.00	**.06**	.09	**.06**	**.07**	**.07**	**.07**	.08
(0, 25)	5	Equal	**.05**	.08	.17	**.04**	.09	.09	.09	.09	.17	**.06**	**.07**	**.07**	.10	.11
		Progress	**.05**	.17	.33	**.05**	.19	.20	1.00	.21	.	.10	.11	.	.	.21
	20	Equal	**.05**	**.05**	**.07**	**.05**	**.05**	**.05**	**.04**	**.04**	**.07**	**.05**	**.05**	**.05**	**.05**	**.06**
		Progress	**.05**	**.05**	**.07**	**.05**	**.05**	**.05**	1.00	**.04**	**.07**	**.05**	**.05**	**.05**	**.06**	**.06**
(0, -1)	5	Equal	**.06**	**.07**	.16	**.05**	.08	**.07**	.08	.08	.15	**.05**	**.06**	**.06**	.09	.09
		Progress	**.05**	.12	.28	**.05**	.14	.15	1.00	.17	.	**.06**	**.07**	.	.	.16
	20	Equal	**.05**	**.05**	**.07**	**.05**	**.05**	**.05**	**.04**	**.04**	**.07**	**.05**	**.05**	**.05**	**.05**	**.06**
		Progress	**.05**	**.05**	.08	**.05**	**.06**	**.05**	1.00	**.04**	**.07**	**.05**	**.05**	**.05**	**.06**	**.06**

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

Table 4

Proportions of Homogeneous Conditions that Meet Bradley's Liberal Criteria

Test	OLS	James	WLS	BF	Welch	AG	Wilcox	Wilcox2	ADF	Bartlett	ML	YB1	YB2	Mixed
Full dataset	1	.75	.13	.98	.67	.63	.24	.71	.19	.88	.87	.61	.52	.57
Subset	1	.85	.17	.99	.83	.80	.32	.93	.25	.92	.92	.81	.69	.69

Table 5

Type I Error Estimates by Variance Pattern and Sample Size Pattern with Heterogeneous Conditions

Test	OLS	James	WLS	BF	Welch	AG	Wilcox	Wilcox2	ADF	BAR	ML	YB1	YB2	Mixed
Variance pattern														
Extreme	**.05**	**.06**	.14	**.07**	**.07**	**.07**	.76	**.07**	.09	**.05**	**.05**	**.06**	**.07**	**.07**
Split	**.04**	**.07**	.14	**.06**	**.07**	**.07**	.76	**.07**	.10	**.05**	**.05**	**.06**	**.07**	**.07**
Progress	**.03**	**.07**	.14	**.05**	**.07**	**.07**	.76	**.07**	.10	**.05**	**.05**	**.06**	**.07**	.08
Extreme-R	.17	.08	.17	.09	.09	.09	.76	.10	.11	**.05**	**.06**	**.07**	.08	.09
Split-R	.20	.10	.19	**.06**	.11	.10	.76	.12	.12	**.06**	**.07**	.08	.09	.10
Progress-R	.14	.09	.19	**.05**	.10	.11	.76	.12	.12	**.06**	**.07**	.08	.09	.10
N-pattern														
Equal	.08	**.06**	.11	**.06**	**.07**	**.06**	**.06**	**.06**	.11	**.05**	**.05**	**.06**	**.07**	**.07**
Extreme	.11	.08	.18	**.06**	.09	.09	1.00	.09	.11	**.06**	**.06**	.08	.09	.08
Split	.13	.09	.21	**.07**	.10	.11	1.00	.13	.11	**.06**	**.06**	.08	.09	.09
Progress	.10	.08	.16	**.07**	.09	.09	.99	.09	.09	**.05**	**.06**	**.06**	**.07**	.09

Note. The Type I error rates meeting the Bradley's criterion are in bold. Extreme-R=One Extreme Inversely, Split-R = Split Inversly, and Progress-R = Progress Inversely (see Table 2 for more details)

Table 6

Proportions of Conditions that Meet Bradley's Liberal Criteria (Heterogeneous Conditions)

Test	OLS	James	WLS	BF	Welch	AG	Wilcox	Wilcox2	ADF	Bartlett	ML	YB1	YB2	Mixed
Full dataset	.34	.66	.12	.78	.60	.62	.20	.62	.17	.85	.83	.55	.45	.58
Subset	.37	.77	.15	.78	.73	.76	.27	.81	.22	.89	.87	.73	.60	.68

Table 7
Power Estimates by Effect Size, Group Size, and Mean Pattern

	Effect Size			Group Size			Mean Pattern		
	.10	.25	.40	5	10	20	Progressive	Partial Null	Multiple Null
OLS	.07	.25	.51	.14	.26	.43	.18	.36	.28
BF	.07	.24	.47	.12	.25	.42	.18	.34	.27
Bartlett	.08	.24	.46	.12	.24	.43	.18	.35	.26
ML	.09	.26	.48	.14	.26	.43	.19	.36	.28

Note. OLS = ANOVA *F* test using ordinary least squares; BF = Brown-Forsythe; 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 8
Power Estimates by Effect Size, Group Size, and Mean Pattern

	Effect Size			Group Size			Mean Pattern		
	.10	.25	.40	5	10	20	Progressive	Partial Null	Multiple Null
BF	.09	.24	.46	.13	.25	.42	.18	.34	.27
Bartlett	.10	.30	.52	.15	.30	.48	.22	.40	.31
ML	.11	.32	.54	.17	.31	.48	.23	.41	.32

Note. BF = Brown-Forsythe; 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.

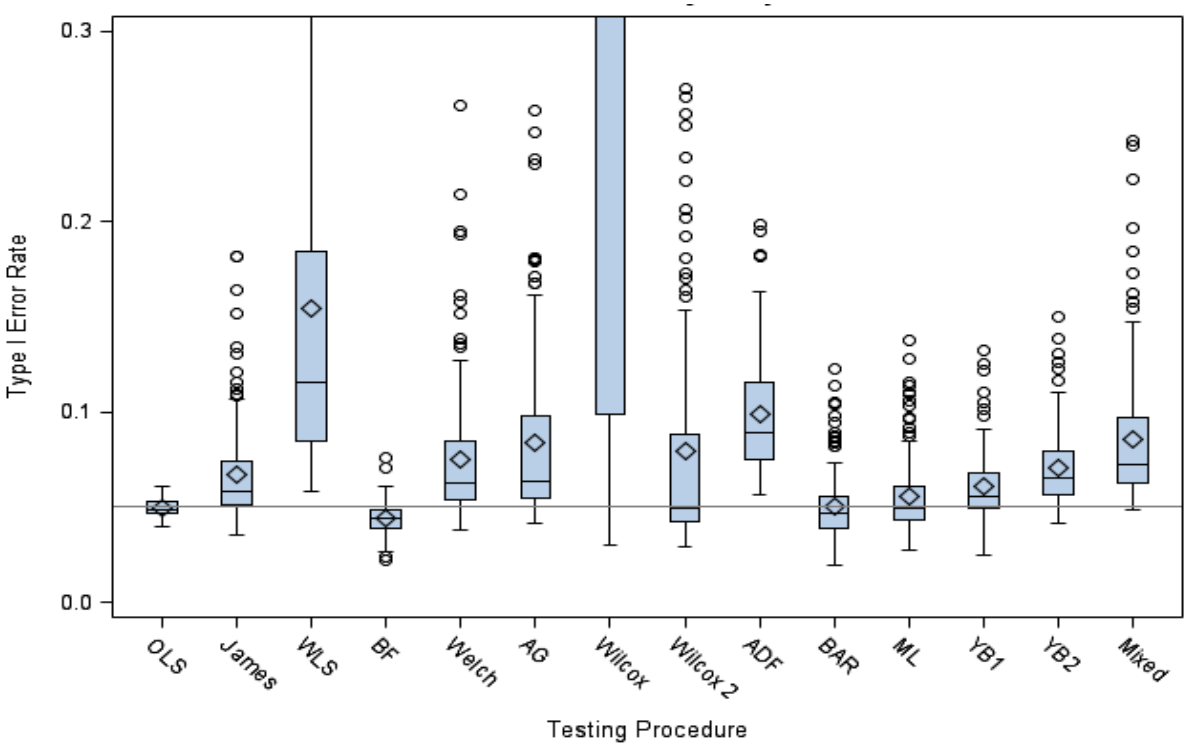

Figure 1. Distributions of Type I error estimates of the fourteen ANOVA tests under homogeneous conditions

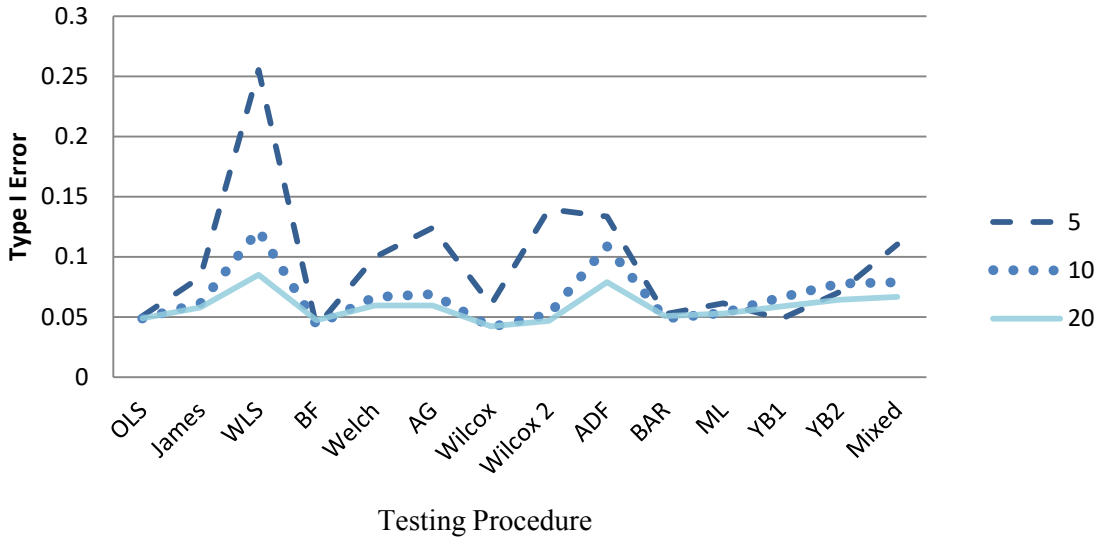

Figure 2. Type I error rates of the fourteen ANOVA tests by group size under homogeneous conditions. For the Wilcox test, only the balanced group conditions (*N*-pattern = 1) were included because the Type I error rates of Wilcox were 1.00 when the groups were unbalanced.

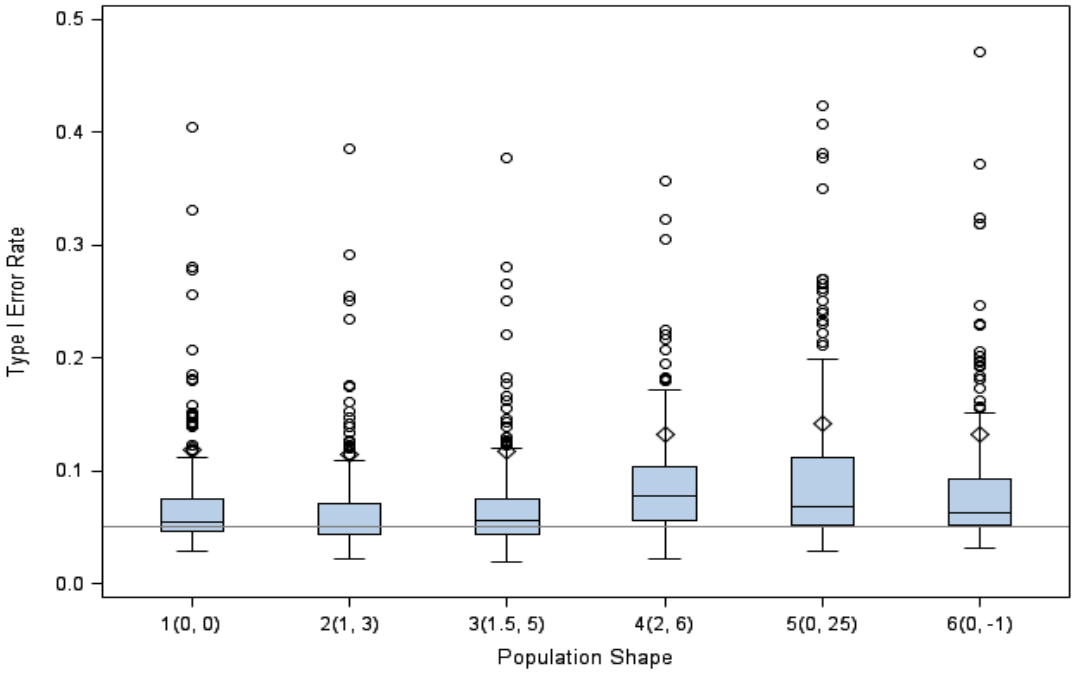

Figure 3. Distributions of Type I error rates of the fourteen ANOVA tests by population shape

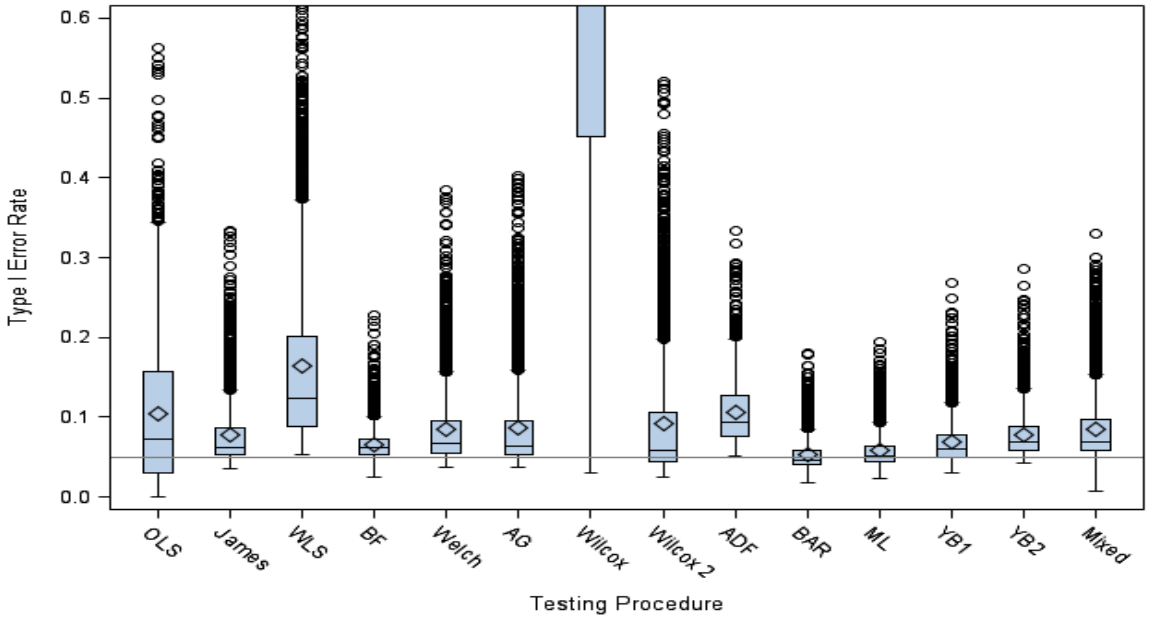

Figure 4. Distributions of Type I error estimates of the fourteen ANOVA tests under heterogeneous conditions

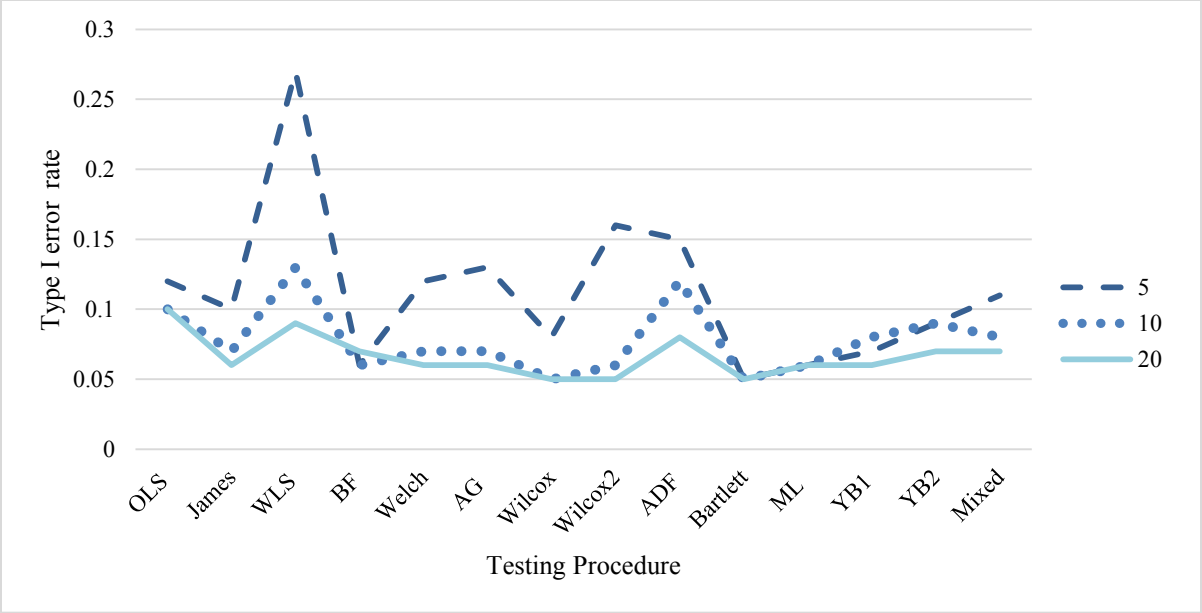

Figure 5. Type I error rates of the robust ANOVA tests by group size under heterogeneous conditions. For the Wilcox test, only the balanced group conditions (*N*-pattern = 1) were included because the Type I error rates of Wilcox were 1.00 when the groups were unbalanced.

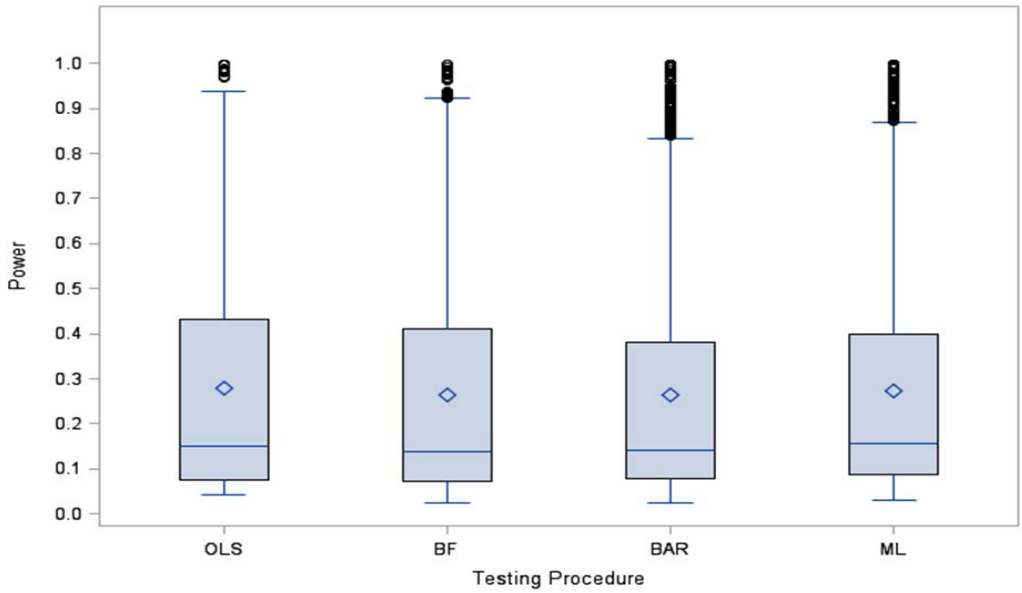

Figure 6. Boxplots of power estimates under homogeneous conditions. OLS = ANOVA *F* test using ordinary least squares; BF = Brown-Forsythe; BAR = structured mean modeling approach with Bartlett estimation; ML = structured mean modeling approach with maximum likelihood estimation.

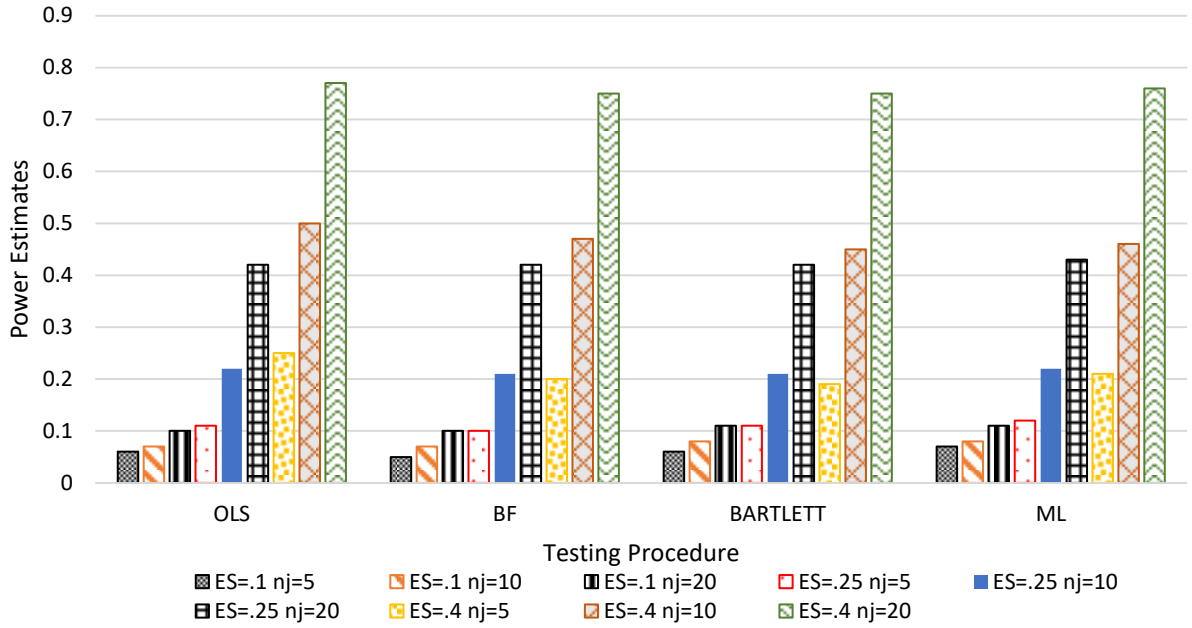

Figure 7. Power estimates by effect size and group size. OLS = ANOVA *F* test using ordinary least squares; BF = Brown-Forsythe; Bartlett = structured mean modeling approach with Bartlett estimation; ML = structured mean modeling approach with maximum likelihood estimation; ES = effect size; nj = group size.

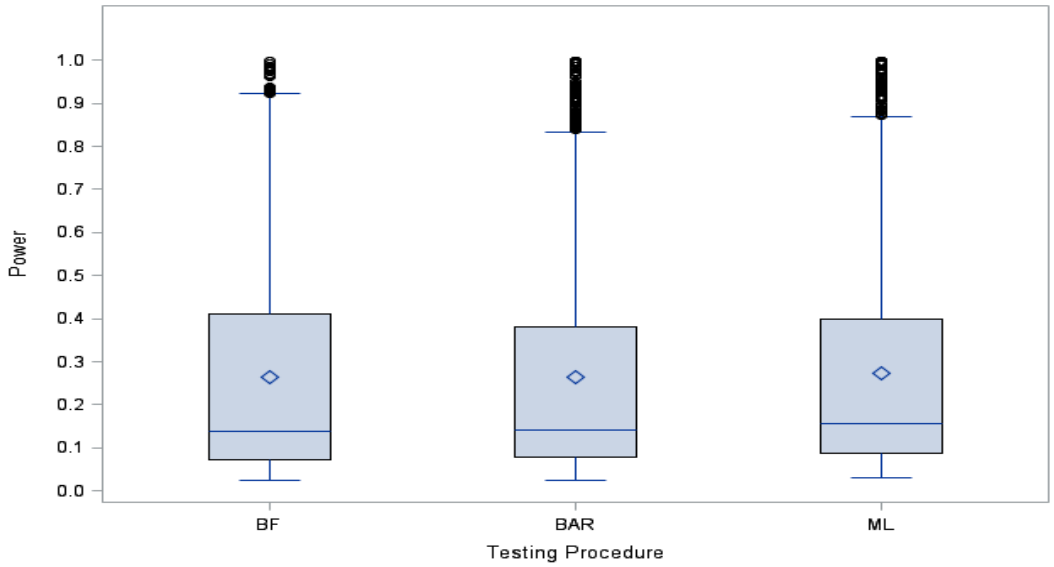

Figure 8. Boxplots of power estimates under heterogeneous conditions

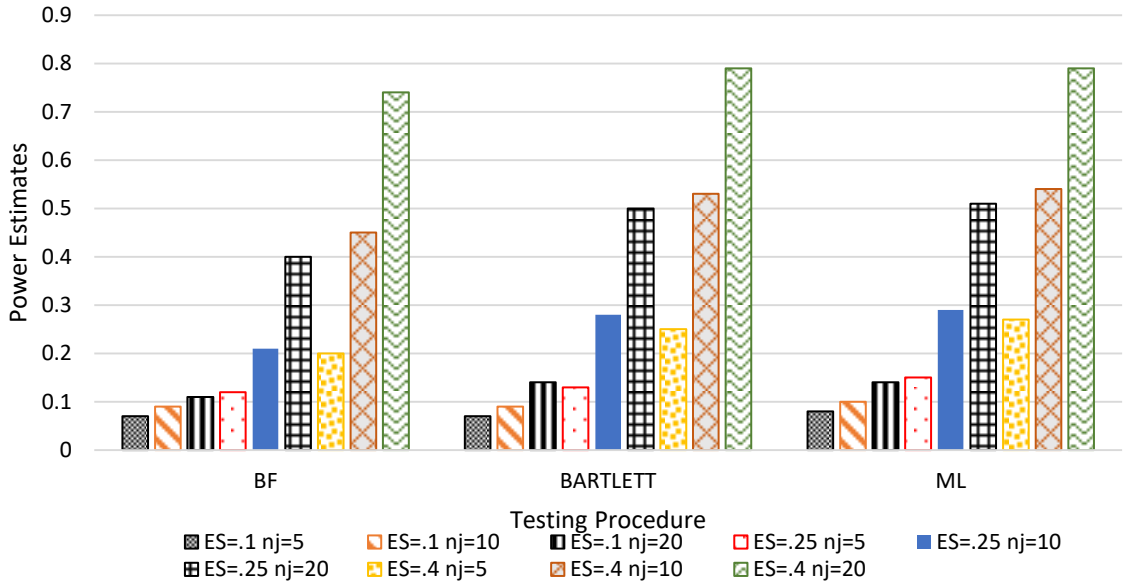

Figure 9. Power estimates by effect size and group size. BF = Brown-Forsythe; BARTLETT = structured mean modeling approach with Bartlett estimation; ML = structured mean modeling approach with maximum likelihood estimation; ES = effect size; nj = group size.