Journal of Modern Applied Statistical Methods

Volume 17 | Issue 2 Article 16

4-16-2018

Evaluating the Efficacy of Conditional Analysis of Variance under Heterogeneity and Non-Normality

Yan Wang
University of Massachusetts, Yan_Wang1@uml.edu

Thanh Pham
University of South Florida

Diep Nguyen
University of South Florida

Eun Sook Kim
University of South Florida

Yi-Hsin Chen
University of South Florida

See next page for additional authors

Evaluating the Efficacy of Conditional Analysis of Variance under Heterogeneity and Non-Normality

Authors
Yan Wang, Thanh Pham, Diep Nguyen, Eun Sook Kim, Yi-Hsin Chen, Jeffrey Kromrey, Zhiyao Yi, and Yue Yin

Journal of Modern Applied Statistical Methods
November 2018, Vol. 17, No. 2, eP2701
doi: 10.22237/jmasm/1555340224

בס"ד

ISSN 1538 − 9472

EMERGING SCHOLARS

Evaluating the Efficacy of Conditional Analysis of Variance under Heterogeneity and Non-Normality

Yan Wang
University of Massachusetts
Lowell, MA

Thanh Pham
University of South Florida
Tampa, FL

Diep Nguyen
University of South Florida
Tampa, FL

Eun Sook Kim
University of South Florida
Tampa, FL

Yi-Hsin Chen
University of South Florida
Tampa, FL

Jeffrey Kromrey
University of South Florida
Tampa, FL

Zhiyao Yi
University of South Florida
Tampa, FL

Yue Yin
University of South Florida
Tampa, FL

A simulation study was conducted to examine the efficacy of conditional analysis of variance (ANOVA) methods where the initial homogeneity of variance screening leads to the choice between the ANOVA *F* test and robust ANOVA methods. Type I error control and statistical power were investigated under various conditions.

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

Introduction

The analysis of variance (ANOVA) *F* test is a commonly-used method to test the equality of population means in psychology (e.g., Ames, Wilson, Barnett, Njoh, & Ottomanelli, 2017; Mas et al., 2016; Molina & Musich, 2016; Walsh et al., 2017). A critical assumption of ANOVA is homogeneity of variance (HOV), that is, that the compared populations have equal variances. Given the importance of the HOV assumption in testing mean differences (Zimmerman, 2004), a conditional procedure has been a common practice in the *t* test, which is a special case of

doi: 10.22237/jmasm/1555340224 | Accepted: November 19, 2017; Published: April 16, 2019.
Correspondence: Yan Wang, Yan_Wang1@uml.edu

ANOVA with two independent sample means. That is, if the HOV assumption is satisfied, the regular *t* test is conducted; if violated, an alternative test such as the Satterthwaite approximate *t* test, which is robust to the violation of the HOV assumption, is conducted. The conditional testing procedure has also been recommended for ANOVA when two or more population means are compared (e.g., Keselman et al., 1998; Lix, Keselman, & Keselman, 1996). Specifically, the ANOVA *F* test is conducted if variances are homogeneous and, otherwise, robust ANOVA methods such as the Brown-Forsythe test (Brown & Forsythe, 1974) and the Wilcox test (Wilcox, 1988, 1989) can be employed.

The selection of a conditional testing procedure involves both the choice of tests to be used (both the test of variances and the test of means) and the selection of an alpha level for the test of variances. Simulation studies have evaluated the performance of HOV testing methods (e.g., Lee, Katz, & Restori, 2010; Wang et al., 2017) and robust ANOVA approaches (e.g., Fan & Hancock, 2012; Nguyen et al., 2016), based on which recommendations have been made regarding the selection of optimal tests. Yet, those recommendations might not be applicable to the conditional ANOVA procedure because they were made assuming the test of variances and the test of means were conducted separately. In conditional ANOVA, however, a combination of an HOV test and an ANOVA method is used, and the ANOVA results might be affected by the initial screening of variance heterogeneity. For example, the HOV test might not detect variance heterogeneity (i.e., lack of power) and thus the *F* test is conducted instead of the robust ANOVA methods (Olejnik, 1987); or the HOV test incorrectly shows variance heterogeneity (i.e., inflation of Type I error rates) so robust ANOVA methods are used instead of the *F* test. The selection of an alpha level for the HOV test is also important because of its influence on the power of this test, which would further impact the test of mean equality.

Olejnik (1987) examined the Type I error rates of the conditional *F* test under variance homogeneity and heterogeneity through Monte Carlo simulations. Note that this conditional *F* test referred to the procedure of conducting the *F* test for replications where researchers failed to reject the null hypothesis of equal variances based on the HOV test results, whereas no test of mean equality was conducted for replications that showed unequal variances. The author found that the conditional *F* test using O'Brien or Brown-Forsythe tests of HOV performed well in terms of the Type I error control with variance homogeneity, except that it became conservative for skewed and leptokurtic distributions. Under variance heterogeneity, both unconditional and conditional *F* tests had adequate Type I error control when sample sizes were relatively large (i.e., average 20 per group). When

sample sizes were small and unequal, both tests were liberal if sample size and group variance were negatively correlated and conservative when they were positively correlated. Regarding the alpha level for the HOV test, Olejnik (1987) noted that increasing the alpha level from .05 to .10 improved the power of the HOV test, but power was still not acceptable with unequal sample sizes and/or skewed and leptokurtic distributions.

Although the study conducted by Olejnik (1987) shed some light upon the behaviors of the conditional *F* test, the efficacy of the conditional ANOVA procedure has not been systematically examined yet. First, it is not clear how the initial screening of variance heterogeneity might impact the ANOVA results when the choice of the *F* test and robust ANOVA tests depends upon the results of HOV. Second, among many possible combinations of the HOV test and the ANOVA method, it is not known which combination performs well under what circumstances. Third, it remains unclear what alpha level should be used for the HOV test that would lead to the optimal results for ANOVA in terms of adequate Type I error control and sufficient statistical power. Therefore, to better understand the performance of the conditional ANOVA procedure, a Monte Carlo simulation study was conducted with various combinations of the HOV and ANOVA tests and different alpha levels considered.

Specifically, this study investigated the Type I error rates and statistical power of four robust ANOVA approaches coupled with five HOV methods. The goal was to provide recommendations for applied researchers regarding the selection of an optimal combination of the HOV and ANOVA tests as well as an appropriate alpha level for the HOV test. The HOV and ANOVA methods considered in this study will be introduced in the following section. Selections of those particular methods were based on their superior performance in Type I error control and statistical power reported in the methodological literature (e.g., Fan & Hancock, 2012; Lee et al., 2010; Nguyen et al., 2016; Ramsey & Ramsey, 2007; Sharma & Kibria, 2013; Wang et al., 2017), which will be discussed shortly as well.

Statistical Tests Examined

A summary of the HOV and ANOVA tests, including the test statistics and equations, is presented below. Brief descriptions of each test are also provided.

HOV

(i) Levene (squared deviations):

$$Z_{ij} = \left(Y_{ij} - \bar{Y}_{.j}\right)^2$$

$$W = \frac{(N-k)\sum_{j=1}^{k} n_j \left(\bar{Z}_{.j} - \bar{Z}_{..}\right)^2}{(k-1)\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(Z_{ij} - \bar{Z}_{.j}\right)^2}$$

where Y_{ij} is the raw score of individual i in group j, $\bar{Y}_{.j}$ is the mean of the j^{th} group, $\bar{Z}_{.j}$ is the group mean of Z_{ij}, $\bar{Z}_{..}$ is the grand mean, N is the total sample size, n_j is the sample size of group j, and k is the number of groups.

(ii) Brown-Forsythe (BF_{HOV}):

$$Z_{ij} = \left|Y_{ij} - \tilde{Y}_{.j}\right|^2$$

$$W = \frac{(N-k)\sum_{j=1}^{k} n_j \left(\bar{Z}_{.j} - \bar{Z}_{..}\right)^2}{(k-1)\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(Z_{ij} - \bar{Z}_{.j}\right)^2}$$

where $\tilde{Y}_{.j}$ is the median of the j^{th} group (note that the bootstrap version of BF_{HOV} was also evaluated).

(iii) O'Brien:

$$\mathrm{r}_{ij}(w) = \frac{\left(w + n_j - 2\right) n_j \left(Y_{ij} - \bar{Y}_{.j}\right)^2 - w s_j^2 \left(n_j - 1\right)}{\left(n_j - 1\right)\left(n_j - 2\right)},$$

where s_j^2 is the within-group unbiased estimate of variance for group j and w ($0 \le w \le l$) is a weighing factor.

(iv) Ramsey conditional test

$$b_2 = m_4 / m_2^2$$

where $m_r = \Sigma(Y_{ij} - \bar{Y}_{.j})^r / n_j$.

ANOVA

(i) *F* test

$$F = \frac{\sum_{j=1}^{k} n_j \left(\bar{Y}_{.j} - \bar{Y}_{..}\right)^2 \Big/ (k-1)}{\sum_{j=1}^{k} \left(n_j - 1\right) s_j^2 \Big/ \left(N-k\right)},$$

where $\bar{Y}_{..}$ is the grand mean of the raw scores.

(ii) Brown-Forsythe (BF) test:

$$F^* = \frac{\sum_{j=1}^{k} n_j \left(\bar{Y}_{.j} - \bar{Y}_{..}\right)^2}{\sum_{j=1}^{k} \left(1 - n_j / N\right) s_j^2}.$$

(iii) Structured means modeling (SMM) with maximum likelihood (ML) estimation:

$$T_{\mathrm{ML}} = \left(N-1\right) F_{\mathrm{ML}},$$

where FML is the maximum likelihood fit function.

(iv) SMM with Bartlett's correction to the ML test statistic (SMM-Bartlett):

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

where p is the number of observed variables ($p = 1$ in ANOVA), m is the number of latent constructs ($m = 0$ in ANOVA), and q is the number of parameters estimated across all groups.

(v) Wilcox:

$$D_j = n_j / s_j^2, \quad W_s = \sum D_j$$
$$Y^* = \sum D_j Y_j^* / W_S, \quad H_m = \sum D_j \left(Y_j^* - Y^*\right)^2$$

where $Y_j^* = Y_{n_j j} / n_j + \sum_{i=1}^{n_j - 1} \left(1 - 1/n_j\right) Y_{ij} \Big/ \left(n_j + 1\right)$.

Statistical Methods for Testing Homogeneity of Variance

Levene with Squared Deviations Test (Levene) Levene (1960) proposed to transform the dependent variable values into either the absolute values of deviations from group means (residuals) or squared residuals. These transformed values will then be used in the ANOVA model as values of the new dependent variable. Thus, a test of variances is transformed into a test of means. This study only examines the Levene test with squared residuals, because it had better Type I error control than the test with absolute residuals (Wang et al., 2017). The obtained W test statistic is compared to the F critical value (F_{crit}) with degrees of freedom ($k - 1$) and ($N - k$) for the numerator and denominator, respectively. The null hypothesis that the group variances are equal is rejected if $W > F_{\text{crit}}$.

Brown-Forsythe Test (BF_{HOV}) This test (Brown & Forsythe, 1974) differs from the Levene test in that it uses the group median instead of the group mean to calculate absolute deviations. The obtained statistic W is computed using the same formula as that in the Levene test.

Bootstrap Brown-Forsythe Test (Bootstrap BF_{HOV}) Boos and Brownie (2004) recommended a bootstrap approach for testing variances based on the BF_{HOV} test. The test draws bootstrap samples from residuals (i.e., deviations from group medians) in the original sample. The residuals are pooled across groups for the bootstrapping, rather than drawing a separate bootstrap sample from each of the groups. In each bootstrap sample, a test statistic for variances is computed and the p-value for the bootstrap test is obtained as the proportion of bootstrap samples with a statistic's value that is greater than that observed in the original data.

O'Brien Test (OB) O'Brien (1979) proposed a method that transforms original scores and then uses these scores in ANOVA or the Welch test as the new dependent variable. The transformation he proposed include a weight (w) to account for the possible departure from kurtosis = 0 in the distribution. The weight ranges between 0 and 1 and it is suggested to set $w = .5$ as default (O'Brien, 1981). The mean of the transformed values for a particular group equals the corresponding group variance, that is,

$$\bar{r}_j = \frac{\sum r_{ij}}{n_j} = S_j^2 .$$

Ramsey Conditional Test: BF_{HOV} or OB (Ramsey) Ramsey (1994) proposed a conditional procedure where the selection between the BF_{HOV} and the OB methods depends upon a test of kurtosis. The kurtosis value for each group (b_{2j}) is compared to critical values obtained from a table provided by Ramsey and Ramsey (1993). A score of −1, 0, or 1 is recorded depending on the test being significantly platykurtic, nonsignificant, or significantly leptokurtic, respectively. A total score, S, across groups is then calculated and used to identify the population as platykurtic if $S \leq -1$, mesokurtic if $S = 0$, or leptokurtic if $S \geq 1$. The OB method will be implemented if the data are platykurtic (i.e., $S \leq -1$), and the BF_{HOV} method will be applied if the data are mesokurtic or leptokurtic (i.e., $S \geq 0$).

Statistical Methods for Testing Mean Equality

ANOVA F Test The ANOVA F test has commonly been used to test the equality of group means. The F statistic follows the F distribution with $(k - 1)$ and $(N - k)$ degrees of freedom. The F test is known to be sensitive to violations of the HOV assumption, especially when sample sizes are unequal across groups.

Brown-Forsythe Test (BF) The Brown-Forsythe test (Brown & Forsythe, 1974) is a modification of the F test. It has been recommended when the HOV assumption is violated and sample sizes are unequal. The test statistic, F^*, has an F distribution with $(k - 1)$ and f degrees of freedom, where f is defined by the Satterthwaite approximation

$$\frac{1}{f} = \sum_j \frac{c_j^2}{n_j - 1} \tag{1}$$

$$c_j = \frac{\left(1 - n_j / N\right) s_j^2}{\sum_{j=1}^{k} \left(1 - n_j / N\right) s_j^2} \tag{2}$$

Structured Means Modeling Approach with Maximum Likelihood Estimation (SMM-ML) Originating from the framework of structural equation modeling, the SMM approach can be applied to test the mean equality of the measured variable (Fan & Hancock, 2012). That is, the dependent variable y can be expressed as $\mathbf{y} = \mathbf{v}_j + \boldsymbol{\delta}$, where $\mathbf{v}_j$ is a $p \times 1$ vector of intercept values (or means) for group j, $\boldsymbol{\delta}$ is a $p \times 1$ vector of normal errors, and p is the number of observed variables ($p = 1$ in ANOVA). The null hypothesis is tested by constraining means to be equal across groups while still allowing for variances of $\boldsymbol{\delta}$ to be

heterogeneous. In other words, the assumption of homogeneity of variance is relaxed with the SMM approach. Estimation within SMM is commonly handled by maximum likelihood. The test statistic T_{ML} follows the χ^2 distribution with degrees of freedom $kp(p + 3) / 2 - q$, where q is the number of parameters estimated across all groups.

SMM with Bartlett's Correction to the ML Test Statistic (SMM-Bartlett) Bartlett (1950) suggested a correction to the ML test statistic in order to accommodate non-normality. The test statistic with correction, T_{BC}, is expected to follow the χ^2 distribution more closely than T_{ML}.

Wilcox Test In the Wilcox method (Wilcox, 1988, 1989), the null hypothesis is rejected when the test statistic H_m exceeds the $(1 - \alpha)$ quantile of the χ^2 distribution with $(k - 1)$ degrees of freedom. In this study, the Wilcox test was conducted after grand mean centering in each sample, because poor Type I error control has been observed if the population grand mean differed from zero (Hsiung, Olejnik, & Huberty, 1994).

Literature Review on the Performance of the Included HOV and ANOVA Tests

Based on simulation studies that have evaluated the performance of HOV testing methods (e.g., Lee et al., 2010; Wang et al., 2017), several patterns have been observed. For example, the Type I error rate inflation under non-normal distributions was evidenced in the Levene test (Wang et al., 2017). The Levene test was inferior to OB and Ramsey, which performed well in terms of Type I error control across a wide range of shapes (Lee et al., 2010; Ramsey & Ramsey, 2007; Sharma & Kibria, 2013; Wang et al., 2017). BF_{HOV} and Bootstrap BF_{HOV} had adequate Type I error control across all shapes except for the extremely leptokurtic distribution (e.g., kurtosis = 25) where they became conservative. When the group size was small (e.g., 5), OB outperformed the other tests in maintaining good Type I error control (Wang et al., 2017). Inconsistent findings regarding the statistical power of the HOV tests have been found in the literature. For instance, Parra-Frutos (2012) observed that the power of the BF_{HOV} test was low (below .60) for small sample sizes (e.g., 5 per group) and decreased when coupled with unbalanced samples; on the other hand, its statistical power increased with larger samples, both balanced and unbalanced. Wang et al. (2017) found that BF_{HOV}, as well as Bootstrap BF_{HOV} and Ramsey, outperformed other tests in power regardless of the sample sizes, with power estimates reaching .80 when the group size was 20.

Ramsey and Ramsey (2007) observed that the Ramsey test had substantially higher power than the BF_{HOV} test (approximately .30 higher) only when the distribution was extremely leptokurtic.

For the ANOVA tests, it has long been known that the conventional *F* test is sensitive to heterogeneous variances, especially when sample sizes are unequal across groups (Harwell, Rubinstein, Hayes, & Olds, 1992; Lix et al., 1996; Rogan & Keselman, 1977). Alternative robust ANOVA tests that are based on SMM, such as SMM-ML or SMM-Bartlett, have been shown to provide adequate Type I error control across a wide range of distribution shapes, sample sizes, and variance heterogeneity patterns (Fan & Hancock, 2012; Nguyen et al., 2016). Inconsistent findings have been observed in terms of the Type I error control of the BF test. Fan and Hancock (2012) found the BF test had inflated Type I error rates under heterogeneous variances regardless of sample sizes being equal or unequal across groups and the inflation was very severe with moderate or large sample sizes. Lix et al. (1996) also cautioned the use of the BF test with heterogeneous variances regardless of the equal or unequal sample sizes. By contrast, Nguyen et al. (2016) found the robustness of the BF test to variance heterogeneity. That is, the test controlled Type I error rates well across various heterogeneous variance patterns and sample sizes. They also noticed the adequate Type I error control of the Wilcox test when average sample size per group increased from 5 to 10 and 20. There was no substantial difference in the statistical power of the SMM-ML, SMM-Bartlett, and BF tests.

As discussed earlier, those studies examined the performance of HOV or ANOVA methods separately, whereas combinations of both methods in the conditional ANOVA procedure have not been investigated systematically and comprehensively. Thus, this study compared the efficacy of various combinations of HOV and ANOVA methods across a wide range of alpha levels for the HOV test. The combinations included five HOV tests (i.e., Levene, BF_{HOV}, Bootstrap BF_{HOV}, OB, and Ramsey) coupled with four robust ANOVA approaches (i.e., BF, SMM-ML, SMM-Bartlett, and Wilcox), which created 20 conditional ANOVA procedures. In these conditional procedures, the conventional *F* test is conducted when the HOV assumption is met; otherwise, one of the robust ANOVA methods is used.

Methods

In this simulation study, the design factors included: number of groups (4 and 8), average number of observations per group (or cell size; 5, 10, and 20), sample size

pattern (4 patterns, see Table 1), variance pattern (7 patterns, see Table 2), mean pattern (4 patterns), maximum group variance ratio (1, 4, 8, and 16), Cohen's *f* effect size (0, .10, .25, and .4), and population shape (γ_1 and γ_2 were [0.00, 0.00], [1.00, 3.00], [1.50, 5.00], [2.00, 6.00], [0.00, 25.00], and [0.00, −1.00], where γ_1 and γ_2 represent skewness and kurtosis, respectively). Non-normal populations were generated by implementing Fleishman's transformation (Fleishman, 1978). Mean patterns included: (1) equal population means; (2) progressive, with all population means equally spaced; (3) one extreme, where one mean differed from the others; and (4) split, where half the group means were different from the other half. Eleven alpha levels were considered for the tests of variances: .01, and .05 to .50 with an incremental increase of .05. Thus, this factorial design had a total of 300,960 conditions (27,360 data conditions × 11 alpha levels for tests of variances).

Continuous data for this study were generated using a random number generator, RANNOR in the SAS/IML statistical software, using a different seed value for each execution of the program. For each condition, 5,000 samples were generated, which provides a maximum standard error of an observed proportion (e.g., Type I error rate estimate) of 0.003 and a 95% confidence interval no wider than ± .006 (Robey & Barcikowski, 1992).

Table 1. Sample size patterns

		Sample sizes											
K	**Group**	**Progressive *N***			**Equal *N***			**Split *N***			**One extreme**		
8	1	2	3	8	5	10	20	2	5	10	4	8	16
	2	3	5	10	5	10	20	2	5	10	4	8	16
	3	4	7	14	5	10	20	2	5	10	4	8	16
	4	5	9	18	5	10	20	2	5	10	4	8	16
	5	5	11	22	5	10	20	8	15	30	4	8	16
	6	6	13	26	5	10	20	8	15	30	4	8	16
	7	7	15	30	5	10	20	8	15	30	4	8	16
	8	8	17	32	5	10	20	8	15	30	12	24	48
	Average *N*	5	10	20	5	10	20	5	10	20	5	10	20
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 var. ratio		1:4	1:8	1:16	1:4	1:8	1:16	1:4	1:8	1:16	1:1
K=8	Group 1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	2	1.43	2.00	3.14	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	3	1.86	3.00	5.28	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	4	2.29	4.00	7.42	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	5	2.72	5.00	9.56	4.00	8.00	16.00	1.00	1.00	1.00	1.00
	6	3.15	6.00	11.70	4.00	8.00	16.00	1.00	1.00	1.00	1.00
	7	3.58	7.00	13.84	4.00	8.00	16.00	1.00	1.00	1.00	1.00
	8	4.00	8.00	16.00	4.00	8.00	16.00	4.00	8.00	16.00	1.00
K=4	Group 1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	2	2.00	3.30	6.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	3	3.00	5.70	11.00	4.00	8.00	16.00	1.00	1.00	1.00	1.00
	4	4.00	8.00	16.00	4.00	8.00	16.00	4.00	8.00	16.00	1.00

		Population variances								
		Progressive inv.			Split inv.			One extreme inv.		
Max var. ratio		4:1	8:1	16:1	4:1	8:1	16:1	4:1	8:1	16:1
K=8	Group 1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	2	1.43	2.00	3.14	1.00	1.00	1.00	1.00	1.00	1.00
	3	1.86	3.00	5.28	1.00	1.00	1.00	1.00	1.00	1.00
	4	2.29	4.00	7.42	1.00	1.00	1.00	1.00	1.00	1.00
	5	2.72	5.00	9.56	4.00	8.00	16.00	1.00	1.00	1.00
	6	3.15	6.00	11.70	4.00	8.00	16.00	1.00	1.00	1.00
	7	3.58	7.00	13.84	4.00	8.00	16.00	1.00	1.00	1.00
	8	4.00	8.00	16.00	4.00	8.00	16.00	4.00	8.00	16.00
K=4	Group 1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
	2	2.00	3.30	6.00	1.00	1.00	1.00	1.00	1.00	1.00
	3	3.00	5.70	11.00	4.00	8.00	16.00	1.00	1.00	1.00
	4	4.00	8.00	16.00	4.00	8.00	16.00	4.00	8.00	16.00

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

Type I error rates and statistical power of the conditional ANOVA tests were evaluated as the simulation outcomes. The unconditional ANOVA tests were also evaluated, serving as a reference for the conditional tests. The Type I error rate was defined as the proportion of replications where the null hypothesis of equal means was rejected when there was no mean difference, regardless of the ANOVA test being conducted. That is, although for each condition, replications followed either

the traditional *F* test or a certain robust ANOVA test based on the HOV test results of equal or unequal variances, respectively, Type I error rates were calculated by taking together the replications that rejected the null hypothesis for both tests. Statistical power was defined likewise. For Type I error rates, the robustness of conditional ANOVA tests using Bradley's (1978) liberal criterion was investigated. This criterion is set at .5α around a nominal alpha. For instance, a test is considered robust when the Type I error rate falls between .025 (= .5 × .05) and .075 (= 1.5 × .05) at alpha level .05. Finally, eta-square analyses were conducted to explore the impact of design factors on variability of the estimated Type I error rates and power. Cohen's (1992) moderate effect size of .0588 was set as a cutoff value for eta-square analyses.

Results

Type I Error Rates under Homogeneous Variances

The overall distributions of Type I error rates for conditional ANOVA tests under the homogeneous variances conditions were investigated using boxplots. Figure 1 shows the distributions of average Type I error rates for the conditional BF test across five HOV tests at 11 alpha levels of HOV tests and unconditional ANOVA tests. As the alpha level of the HOV test increased from .01 to .50, Type I error rates of the conditional test deviated more from the nominal alpha level. This might be because, with the increase of statistical power of the HOV test, the robust ANOVA tests were more frequently selected over the ANOVA *F* test. That is, the average percentage of replications selecting the robust ANOVA tests across conditions increased from 1.45% to 48.56% as the alpha level of the HOV test increased from .01 to .50. Because the Type I error control of the robust ANOVA tests was inferior to that of the *F* test, increasing the alpha level of the HOV test would lead to less adequate Type I error control for the test of means under variance homogeneity. This pattern was also observed from a series of boxplots like Figure 1 for other conditional ANOVA tests. Similarly, the proportion of conditions meeting Bradley's criterion decreased from 1 to close to that of the corresponding unconditional test as the alpha of HOV tests increased.

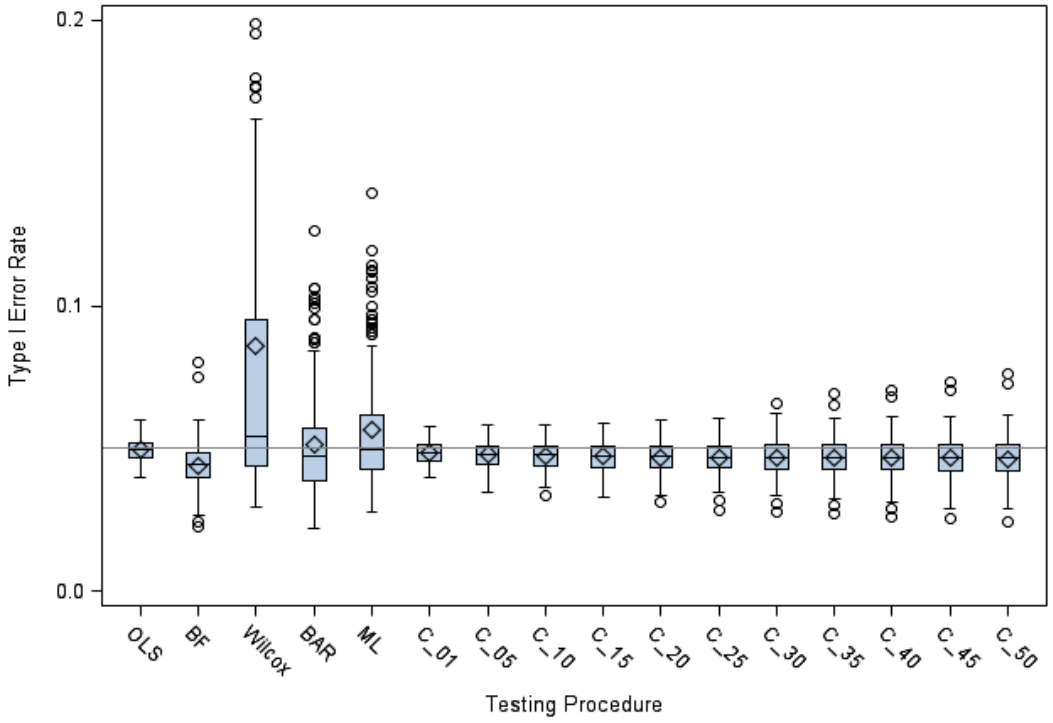

Figure 1. Distribution of Type I error rates for unconditional tests and conditional Brown-Forsythe robust ANOVA test with combinations of HOV tests at different alpha levels denoted by C_01 to C_50; OLS is the ANOVA *F* test with ordinary least squares, BAR is SMM-Bartlett, and ML is SMM-ML

Eta-square analyses revealed that cell size by test of means (η^2 = .119), test of means (η^2 = .102), shape (η^2 = .085), and cell size (η^2 = .063) had substantial impact on the Type I error rates of conditional ANOVA procedures. When sample size increased to 20, Type I error control notably improved across the tests of means, particularly for Wilcox. Conditional tests using BF, SMM-Bartlett, and SMM-ML as tests of means had adequate Type I error control with normal data. When data were non-normal, the BF controlled Type I error rates better than SMM-Bartlett and SMM-ML which showed inflated Type I error rates. Regardless of the distribution shape, Wilcox tended to have inflated Type I error rates. Although there was no significant difference in Type I error control among tests of means paired with different HOV tests, conditional tests using Levene and OB outperformed those using BF_{HOV}, Ramsey, and Bootstrap BF_{HOV} for all tests of means and across all simulation conditions. For example, the proportions of conditions meeting Bradley's criterion were .60 and .55 for conditional SMM-Bartlett with Levene and

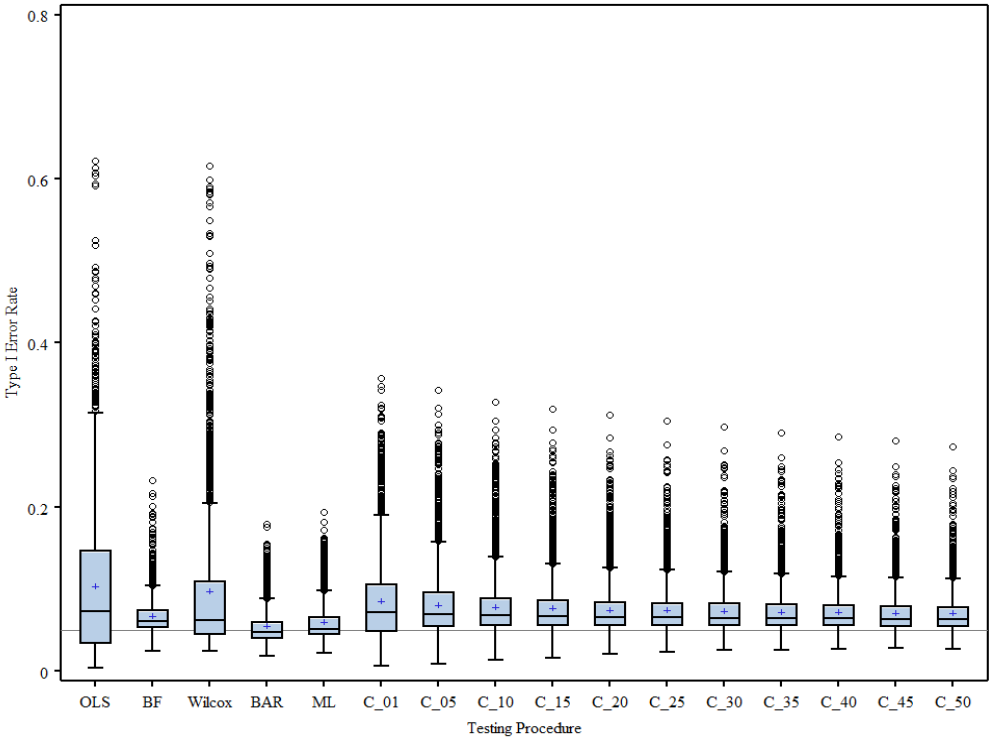

Figure 2. Distribution of Type I error rates under variance heterogeneity for unconditional tests and conditional Brown-Forsythe robust ANOVA test with the Brown-Forsythe test for homogeneity of variance (HOV); C_01 to C_50 denote the alpha levels of the HOV test, from .01 to .50

OB, respectively, as opposed to .50, .52, and .50 with BF_{HOV}, Ramsey, and Bootstrap BF_{HOV}, when the cell size was 10 and the HOV alpha level was .40.

Type I Error Rates under Heterogeneous Variances

Figure 2 presents the overall distributions of Type I error rates for 5 unconditional tests of means and the conditional BF test (with BF_{HOV} as test of variances) at 11 alpha levels under the heterogeneous variances conditions. Observing a series of boxplots like Figure 2 for other conditional procedures revealed that the performance of the conditional tests became closer to their unconditional counterparts as the alpha level of HOV tests increased. SMM-Bartlett performed slightly better than SMM-ML, followed by BF and Wilcox, and the ANOVA *F* test had the worst Type I error control. As the alpha level of the HOV tests increased

from .01 to .50, the average percentage of replications selecting the ANOVA *F* test across conditions decreased from 64.16% to 11.95%. Therefore, a larger alpha level of the HOV tests was associated with more adequate Type I error control. Conditional tests using Levene and OB as the HOV tests prior to testing mean equality had better Type I error control than those using BF_{HOV}, Ramsey, and Bootstrap BF_{HOV}, which is consistent with the finding under homogeneity of variance. These patterns were also evidenced when the proportions of conditions meeting Bradley's liberal criterion were examined. In addition, as can be seen from Figure 3, among the conditional tests, SMM-Bartlett, SMM-ML, and BF, paired with Levene and OB had higher proportions of conditions meeting Bradley's criterion across different alpha levels. The BF test of means paired with Levene seemed to excel above the rest based on the largest proportion of replications that met Bradley's criterion across all alpha levels of HOV.

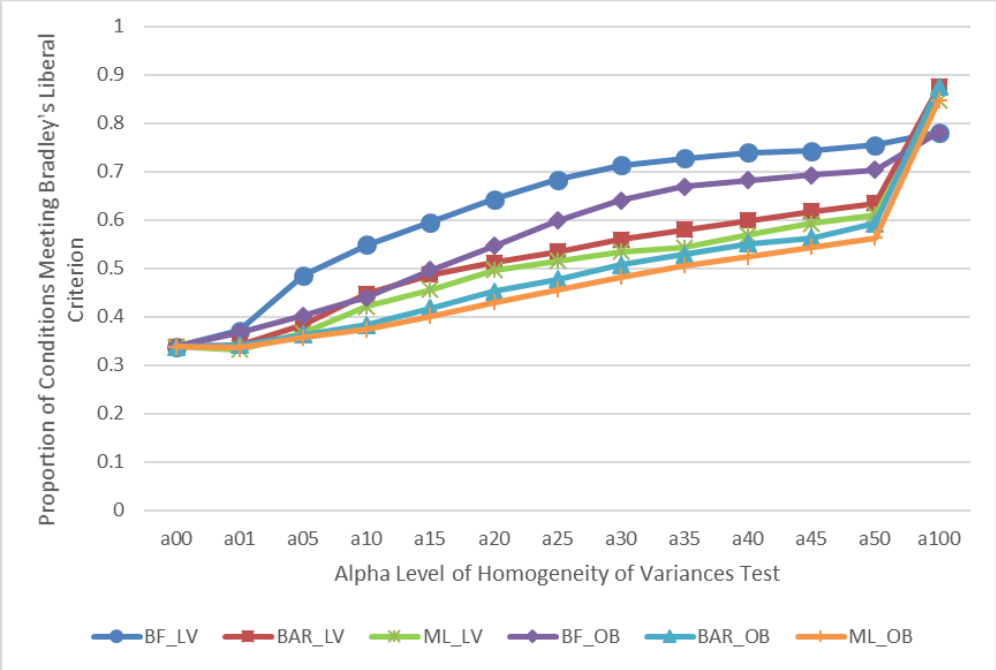

Figure 3. Proportion of conditions meeting Bradley's criterion with cell size 10 under variance heterogeneity; note that a00 represents the ANOVA *F* test and a100 represents the unconditional test of the corresponding conditional test; BF_LV, BAR_LV, and ML_LV refer to the Brown-Forsythe test, SMM-Bartlett, and SMM-ML, each paired with Levene test of homogeneity of variance, respectively; BF_OB, BAR_OB, and ML_OB refer to BF, SMM-Bartlett, and SMM-ML, each paired with O'Brien test of homogeneity of variance, respectively

Table 3. Proportions of conditions that met Bradley's liberal criterion by test, cell size, and variance pattern under variance heterogeneity

		Variance pattern					
Cell size	**Test**	**2**	**3**	**4**	**5**	**6**	**7**
5	BF_LV	698	843	862	196	601	622
	BAR_LV	665	688	743	251	378	383
	ML_LV	643	713	767	217	317	309
	BF_OB	684	813	867	142	439	460
	BAR_OB	638	673	753	170	285	305
	ML_OB	614	691	770	150	239	249
	OLS	563	500	632	56	90	236
	BF	715	951	1000	299	882	875
	BAR	771	778	792	708	688	674
	ML	771	799	785	708	660	653
10	BF_LV	546	896	963	249	857	880
	BAR_LV	872	891	907	544	723	677
	ML_LV	866	902	908	505	684	625
	BF_OB	551	895	947	247	828	857
	BAR_OB	860	881	888	529	708	658
	ML_OB	857	889	899	489	667	607
	OLS	549	458	597	63	111	250
	BF	472	875	1000	361	979	1000
	BAR	924	917	903	840	840	826
	ML	889	903	882	813	806	799
20	BF_LV	481	771	990	311	799	953
	BAR_LV	926	941	951	797	824	795
	ML_LV	916	932	949	782	816	778
	BF_OB	482	776	985	315	792	946
	BAR_OB	926	941	948	799	822	793
	ML_OB	917	934	948	780	814	775
	OLS	583	521	590	69	125	250
	BF	458	660	1000	368	847	1000
	BAR	938	917	910	889	840	826
	ML	910	910	910	875	840	819

Note: OLS is the ANOVA *F* test with ordinary least squares; BF is the Brown-Forsythe test; BAR is SMM-Bartlett; ML is SMM-ML; BF_LV, BAR_LV, and ML_LV refer to BF, BAR, and ML each paired with Levene test of homogeneity of variance, respectively; BF_OB, BAR_OB, and ML_OB refer to BF, BAR, and ML each paired with O'Brien test of homogeneity of variance, respectively; variance pattern 2 is one extreme, 3 is split, 4 is progressive, 5 is one extreme inversely, 6 is split inversely, and 7 is progressive inversely; the value of proportion for each cell should be divided by 1000

Eta-square analyses showed that variance pattern (η^2 = .163), cell size (η^2 = .113), cell size by variance pattern (η^2 = .073), variance pattern by cell size pattern (η^2 = .071), and cell size by test of means (η^2 = .063) had substantial impact

on Type I error rates under variance heterogeneity. Table 3 presents the Bradley results by test, cell size, and variance pattern. Note that only a few selected conditional tests are presented, including BF, SMM-Bartlett, and SMM-ML, each paired with Levene and OB, due to their better performance in Type I error control. As shown in Table 4, when cell size was 5, the conditional BF seemed to have better control of Type I error rates than the rest across all variance patterns, except when the pattern was one extreme inversely where none of the conditional tests meets Bradley's liberal criterion. As cell size increased to 10, the advantage of the conditional BF was only present for split inversely and progressive inversely patterns, whereas with cell size 20, the conditional BF was inferior to the conditional SMM-Bartlett and the conditional SMM-ML across all variance patterns. Put another way, increasing the cell size improved the Type I error control substantially for SMM-Bartlett and SMM-ML, but BF seemed to be least affected in terms of Type I error rates by cell size.

Table 4. Proportions of conditions that met Bradley's liberal criterion by test, cell size pattern, and variance pattern under variance heterogeneity

		Variance pattern					
Cell size pattern	**Test**	**2**	**3**	**4**	**5**	**6**	**7**
Progressive	BF_LV	602	887	968	163	739	763
	BAR_LV	889	927	918	391	629	579
	ML_LV	853	927	934	369	586	529
	BF_OB	617	889	949	146	639	662
	BAR_OB	893	932	907	363	556	508
	ML_OB	864	932	920	345	524	474
	OLS	972	750	593	0	0	0
	BF	565	843	1000	250	907	935
	BAR	870	861	907	833	778	778
	ML	843	852	907	806	750	769
Equal	BF_LV	389	871	1000	417	864	994
	BAR_LV	782	835	860	791	827	871
	ML_LV	746	810	833	748	795	840
	BF_OB	375	847	1000	399	838	996
	BAR_OB	731	821	864	738	806	884
	ML_OB	700	794	838	701	774	853
	OLS	241	444	981	250	435	981
	BF	426	898	1000	454	898	991
	BAR	907	833	833	889	852	824
	ML	889	843	824	889	833	843

Table 4 (continued).

		Variance pattern					
Cell size pattern	**Test**	**2**	**3**	**4**	**5**	**6**	**7**
Split	BF_LV	656	718	855	141	603	644
	BAR_LV	833	697	822	373	461	440
	ML_LV	817	752	854	348	430	397
	BF_OB	662	693	868	125	542	568
	BAR_OB	836	664	812	368	453	420
	ML_OB	821	704	852	344	426	382
	OLS	954	19	241	0	0	0
	BF	583	796	1000	250	852	907
	BAR	824	917	889	750	750	750
	ML	759	926	880	731	713	704
One extreme	BF_LV	654	870	929	288	803	871
	BAR_LV	780	901	867	567	650	582
	ML_LV	818	908	878	540	613	518
	BF_OB	635	884	915	269	727	791
	BAR_OB	773	911	869	529	606	529
	ML_OB	800	923	880	501	569	465
	OLS	93	759	611	0	0	0
	BF	620	778	1000	417	954	1000
	BAR	907	870	843	778	778	750
	ML	935	861	824	769	778	713

Note: OLS is the ANOVA *F* test with ordinary least squares; BF is the Brown-Forsythe test; BAR is SMM-Bartlett; ML is SMM-ML; BF_LV, BAR_LV, and ML_LV refer to BF, BAR, and ML each paired with Levene test of homogeneity of variance, respectively; BF_OB, BAR_OB, and ML_OB refer to BF, BAR, and ML each paired with O'Brien test of homogeneity of variance, respectively; variance pattern 2 is one extreme, 3 is split, 4 is progressive, 5 is one extreme inversely, 6 is split inversely, and 7 is progressive inversely; the value of proportion for each cell should be divided by 1000

In addition, the Type I error rates and proportions of conditions meeting Bradley's liberal criterion (see Table 4) were examined by test, cell size pattern, and variance pattern. Taken together, several major trends emerged. When cell sizes were equal, the conditional BF controlled Type I error rates more adequately with split, progressive, split inversely, and progressive inversely patterns than the conditional SMM-Bartlett and SMM-ML tests. When cell sizes were unequal, SMM-Bartlett and SMM-ML seemed to have good Type I error control consistently across all heterogeneous patterns, while BF outperformed them only with progressive, split inversely, and progressive inversely variance patterns. Type I error rates were inflated noticeably across all conditional tests under one extreme inversely, split inversely, and progressive inversely variance patterns. This was expected because, with these three patterns, smaller cell sizes were paired with larger variances. Despite this, the conditional BF paired with Levene seemed to

have a relatively large proportion (above .700) meeting Bradley's criterion under split inversely and progressive inversely patterns, except when the cell size pattern was split.

Statistical Power Analysis

This section presents the analyses of statistical power among conditional and unconditional ANOVA tests. Based on the performance of conditional ANOVA tests in controlling for Type I error rates, six conditional tests were selected that had adequate Type I error control to include in the power analyses. These conditional tests are the combinations of BF, SMM-Bartlett, and SMM-ML with Levene and OB. The power of the ANOVA *F* test was analyzed for the homogeneous conditions but not for the heterogeneous conditions due to the adequate control of Type I error in the first scenario but not the second one. In addition, there were eleven alpha levels examined for each conditional test, resulting in 70 (11 × 6 conditional plus 4 unconditional) tests for homogeneous conditions and 69 (11 × 6 conditional plus 3 unconditional) tests for heterogeneous conditions.

We excluded the conditions that did not have all tests satisfying the Bradley criterion for homogeneous and heterogeneous conditions separately. Thus, among 144 homogeneous conditions, 29 conditions (or 20.14%) were excluded from the statistical power analysis. Generally, those excluded conditions involved different levels of non-normal distributions for cell size of 5 and extremely non-normal conditions (particularly, skewness = 2 and kurtosis = 6) for cell sizes of 10 and 20. Regarding heterogeneous conditions, 772 out of 2,592 (29.78%) null conditions met the Bradley criterion for all 69 tests. These 772 conditions were distributed relatively equally among population shapes (from 130 to 170 conditions for each shape), except for the shape of with skewness = 2, kurtosis = 6 that had a smaller number of conditions (only 49 conditions) included in the power analysis. Among these 772 conditions, a majority (549 conditions) had one extreme, split, or progressive variance patterns with a small variance ratio (1:4). These Type I error conditions in which Type I error was adequately controlled across tests were then matched with non-null conditions to define the conditions used for power analysis. As a result, 1,035 homogeneous and 6,948 heterogeneous conditions were selected to use in power analyses.

Table 5. Statistical power for conditional ANOVA tests under homogeneous and heterogeneous variances conditions at different alpha levels

		Alpha level										
Variances	**Test**	**0.01**	**0.05**	**0.10**	**0.15**	**0.20**	**0.25**	**0.30**	**0.35**	**0.40**	**0.45**	**0.50**
Homogenous	BF_LV	302	299	297	296	294	294	293	293	292	292	292
	BAR_LV	304	304	304	304	303	303	302	301	300	299	297
	ML_LV	305	305	306	306	306	306	306	306	305	305	304
	BF_OB	304	301	299	298	296	295	294	293	292	292	291
	BAR_OB	305	305	304	304	304	303	303	302	301	300	300
	ML_OB	305	305	305	306	306	306	306	306	306	305	304
	OLS	306										
	BF	292										
	BAR	279										
	ML	289										
Heterogenous	BF_LV	299	303	305	307	308	309	310	311	312	313	314
	BAR_LV	298	302	305	306	308	309	310	310	311	312	313
	ML_LV	305	310	312	313	315	316	316	317	318	318	318
	BF_OB	305	309	311	313	314	315	316	317	318	318	319
	BAR_OB	308	314	317	319	321	322	324	325	326	326	327
	ML_OB	307	312	315	318	320	321	323	324	325	326	327
	BF	317										
	BAR	319										
	ML	328										

Note: OLS is the ANOVA *F* test with ordinary least squares; BF is the Brown-Forsythe test; BAR is SMM-Bartlett; ML is SMM-ML; BF_LV, BAR_LV, and ML_LV refer to BF, BAR, and ML each paired with Levene test of homogeneity of variance, respectively; BF_OB, BAR_OB, and ML_OB refer to BF, BAR, and ML each paired with O'Brien test of homogeneity of variance, respectively; variance pattern 2 is one extreme, 3 is split, 4 is progressive, 5 is one extreme inversely, 6 is split inversely, and 7 is progressive inversely; the value of proportion for each cell should be divided by 1000

The distributions of statistical power for each conditional test under homogeneous and heterogeneous variances were examined. In general, there were no substantial differences in power estimates across conditional tests for homogeneous or heterogeneous variances conditions. Note that eta-square analyses for statistical power estimates were not conducted due to the unbalanced designs. Instead, the summaries of estimated power by alpha level for each test are presented for homogeneous and heterogeneous variances conditions (see Table 5). Overall, as the alpha level increased from .01 to .50, the power of the conditional tests decreased gradually (and slightly) under homogeneous variances conditions. This was because the robust ANOVA test was selected more frequently than the *F* test (i.e., the average percentage of replications selecting the robust test across conditions increased from 1.51% to 48.96%) and the robust test had slightly lower power than the *F* test. The opposite scenario was observed with heterogeneous variances. That is, when the alpha level increased, the power became greater for all conditional tests and was very close to that of the unconditional tests with alpha

level .50. For conditional tests, the average percentage of replications selecting the *F* test over the robust test decreased from 63.34% to 11.68% as the alpha level for the HOV test increased from .01 to .50. Among the six conditional tests, SMM-Bartlett and SMM-ML paired with OB tended to have higher power than the rest.

Discussion and Conclusions

Testing the HOV assumption has been recommended as a critical procedure prior to testing mean equality. If the assumption appears to be satisfied, the ANOVA *F* test is recommended; otherwise, alternative ANOVA methods, i.e., robust ANOVA methods, can be applied. To our knowledge, this simulation study was the first study to comprehensively examine the efficacy of this conditional ANOVA procedure, aiming to select optimal combinations of the HOV and ANOVA tests and identify an appropriate alpha level for the HOV test. Evidence from this study indicates that overall SMM-Bartlett, BF, and SMM-ML coupled with Levene and OB are the best performing conditional ANOVA methods. Particularly, Levene and OB provided notably superior Type I error control in the conditional tests than the two BF tests and Ramsey's test. This might occur because, when there was no group mean difference, Levene had the highest power in detecting heterogeneous variances and OB had the most adequate Type I error control (Lee et al., 2010; Ramsey & Ramsey, 2007; Sharma & Kibria, 2013; Wang et al., 2017). Therefore, both tests could lead to the correct decisions in selecting a test of mean (i.e., a robust ANOVA test and the regular *F* test, respectively). Between the Levene and OB tests, the latter resulted in conditional tests with more statistical power although the power advantages were small. The superior performance of OB in the conditional ANOVA is consistent with what Olejnik (1987) found in their simulation study.

The selection of an alpha level for the test of variances is important because of its influence on the power of this test (Olejnik, 1987). Larger alpha levels allow the test of variances to steer researchers away from the ANOVA *F* test under conditions in which it is likely to perform poorly in terms of Type I error control. Concomitantly, larger alpha levels for this test also steer researchers away from the ANOVA *F* test more often under conditions of variance homogeneity, conditions in which it is the most powerful test of means. In other words, with large alpha levels, the HOV test incorrectly rejects the null hypothesis of equal variances (i.e., inflation of Type I error rates) so the robust ANOVA methods are selected. With small alpha levels, the HOV test might not detect variance heterogeneity (i.e., lack of power) and thus the *F* test is mistakenly conducted. Although it is possible that these two incorrect actions might lead to the correct conclusion regarding the mean

equality, they are not encouraged. Instead, alpha levels should be carefully selected so they can serve as a reasonable compromise between these competing effects. Alpha levels are suggested near the middle of the range examined in this study. That is, alpha levels between .20 and .30 in conditional tests provide adequate Type I error control in heterogeneous variance conditions, while providing nearly as much power as the unconditional robust tests.

In addition, the choice among SMM-Bartlett, BF, and SMM-ML in the conditional ANOVA procedure appears to be dependent upon the sample sizes in the study. With the smallest samples examined in this simulation (average $n_j = 5$), the BF test of means, coupled with Levene and OB, provided the best Type I error control. The robustness of the BF test to small sample sizes is also recognized in Nguyen et al. (2016). Conversely, as sample size increased, the SMM-ML and SMM-Bartlett tests used in SMM were superior to the BF test of means. Further, these SMM tests provided more statistical power than the BF test under both homogeneous and heterogeneous conditions, when these tests were paired with Levene and OB.

To conclude, ANOVA is a popular method used to compare the means of several groups. The sensitivity of ANOVA to violations of the homogeneity of variance assumption is well known, which calls for a conditional procedure where the choice of the *F* test and robust ANOVA methods depends upon the test of variances. Despite this, the efficacy of such a conditional testing procedure has not been thoroughly investigated. The current study systematically examined tests of variance homogeneity coupled with tests of means for one-factor models in terms of Type I error control and statistical power. Results of the study contribute to the literature by evaluating the performance of such conditional testing procedures for testing group means under a wide variety of conditions. Based on the results, the Levene and O'Brien tests with an alpha level between .20 and .30 are recommended for applied researchers to test the equality of variances. If the test of variances fails to reject the null hypothesis, applied researchers can choose the regular ANOVA *F* test to compare the mean equality among groups. If the test of variances shows unequal variances, applied researchers can choose the BF test when average cell size is around 5 and the SMM-ML and SMM-Bartlett tests when average cell size is larger than 5.

References

Ames, H., Wilson, C., Barnett, S., Njoh, E., & Ottomanelli, L. (2017). Does functional motor incomplete (AIS D) spinal cord injury confer unanticipated challenges? *Rehabilitation Psychology, 62*(3), 401-406. doi: 10.1037/rep0000146

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

Boos, D. D., & Brownie, C. (2004). Comparing variances and other measures of dispersion. *Statistical Science, 19*(4), 571-578. doi: 10.1214/088342304000000503

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31*(2), 144-152. doi: 10.1111/j.2044-8317.1978.tb00581.x

Brown, M. B., & Forsythe, A. B. (1974). Robust tests for the equality of variances. *Journal of the American Statistical Association, 69*(346), 364-367. doi: 10.1080/01621459.1974.10482955

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

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: 10.3102/10769986028001045

Hsiung, T.-H., 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

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B., … Levin, J. R. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses.

Review of Educational Research, 68(3), 350-386. doi: 10.3102/00346543068003350

Lee, H. B., Katz, G. S., & Restori, A. F. (2010). A Monte Carlo study of seven homogeneity of variance tests. *Journal of Mathematics and Statistics, 6*(3), 359-366. doi: 10.3844/jmssp.2010.359.366

Levene, H. (1960). Robust tests for equality of variances. In I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow, & H. B. Mann (Eds.), *Contributions to probability and statistics: Essays in honor of Harold Hotelling* (pp. 278-292). Palo Alto, CA: Stanford University Press.

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

Mas, J. M., Baqués, N., Balcells-Balcells, A., Dalmau, M., Giné, C., Gràcia, M., & Vilaseca, R. (2016). Family quality of life for families in early intervention in Spain. *Journal of Early Intervention, 38*(1), 59-74. doi: 10.1177/1053815116636885

Molina, M. F., & Musich, F. M. (2016). Perception of parenting style by children with ADHD and its relation with inattention, hyperactivity/impulsivity and externalizing symptoms. *Journal of Child and Family Studies, 25*(5), 1656-1671. doi: 10.1007/s10826-015-0316-2

Nguyen, D. T., Kim, E. S., Wang, Y., Pham, T. V., Kromrey, J., & Chen, Y.-H. (2016, April). Testing mean equality under heterogeneity and non-normality: An empirical comparison of tests for one-factor ANOVA models. Paper presented at the meeting of American Educational Research Association, Washington, D.C.

O'Brien, R. G. (1979). A general ANOVA method for robust tests of additive models for variances. *Journal of the American Statistical Association, 74*(368), 877-880. doi: 10.1080/01621459.1979.10481047

O'Brien, R. G. (1981). A simple test for variance effects in experimental designs. *Psychological Bulletin, 89*(3), 570-574. doi: 10.1037/0033-2909.89.3.570

Olejnik, S. (1987). Conditional ANOVA for mean differences when population variances are unknown. *The Journal of Experimental Education, 55*(3), 141-148. doi: 10.1080/00220973.1987.10806447

Parra-Frutos, I. (2012). Testing homogeneity of variances with unequal sample sizes. *Computational Statistics, 28*(3), 1269-1297. doi: 10.1007/s00180-012-0353-x

Ramsey, P. H. (1994). Testing variances in psychological and educational research. *Journal of Educational and Behavioral Statistics, 19*(1), 23-42. doi: 10.3102/10769986019001023

Ramsey, P. H., & Ramsey, P. P. (1993). Updated version of the critical values of the standardized fourth moment. *Journal of Statistical Computation and Simulation, 44*(3-4), 231-241. doi: 10.1080/00949659308811460

Ramsey, P. H., & Ramsey, P. P. (2007). Testing variability in the two-sample case. *Communications in Statistics – Simulation and Computation, 36*(2), 233-248. doi: 10.1080/03610910601158310

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. D. & 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

Sharma, D., & Kibria, B. M. G. (2013). On some test statistics for testing homogeneity of variances: A comparative study. *Journal of Statistical Computation and Simulation, 83*(10), 1944-1963. doi: 10.1080/00949655.2012.675336

Walsh, A. S. J., Wesley, K. L., Tan, S. Y., Lynn, C., O'Leary, K., Wang, Y., … Rodriguez, C. A. (2017). Screening for depression among youth with HIV in an integrated care setting. *AIDS Care, 29*(7), 851-857. doi: 10.1080/09540121.2017.1281878

Wang, Y., Rodriguez de Gil, P., Chen, Y.-H., Kromrey, J. D., Kim, E. S., Nguyen, D. T., … Romano, J. (2017). Comparing the performance of approaches for testing the homogeneity of variance assumption in one-factor ANOVA models. *Educational and Psychological Measurement, 77*(2), 305-329. doi: 10.1177/0013164416645162

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

Zimmerman, D. W. (2004). A note on preliminary tests of equality of variances. *British Journal of Mathematical and Statistical Psychology, 57*(1), 173-181. doi: 10.1348/000711004849222