Conditional and Unconditional ANOVA Tests: An Empirical Comparison of Type I Error Control and Statistical Power under Variance Heterogeneity and Non-Normality

Yan Wang[1], Zhiyao Yi[1], Thanh Pham[1], Diep Nguyen [1], Yi-Hsin Chen [1], Eun Sook Kim[1], Jeffrey Kromrey[1], Yue Yin[1]
[1] University of South Florida, 4202 E Fowler Ave, Tampa, FL 33620

Abstract
The analysis of variance (ANOVA) *F* test is a commonly used method to test the mean equality among two or more populations. A critical assumption of ANOVA is homogeneity of variance (HOV), that is, the compared groups have equal population variances. Although it is encouraged to test HOV as part of the regular ANOVA procedure, the efficacy of the initial HOV screening that leads to the choice between the ANOVA *F* test and robust ANOVA methods (namely, conditional ANOVA) has not been investigated systematically. This simulation study examined the efficacy of conditional ANOVA methods under various research conditions. Results suggested that under a small sample size (e.g., 5 per group) the combination of the Brown-Forsythe test of means with the Levene or O'Brien test of variances is the best choice; with large sample sizes, structured means modeling with maximum likelihood or the Bartlett's correction coupled with Levene or O'Brien are the best combinations; and alpha levels between .20 and .30 for the test of variances are most appropriate.

Keywords: Analysis of variance, Homogeneity of variance, Non-normality, Type I error control, Statistical power

1. Introduction

The analysis of variance (ANOVA) *F* test is a commonly used method to test the equality of group 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, 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 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 group means are compared (e.g., 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 methods. 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 the 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 that the *F* test was conducted 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. Therefore, to better understand the performance of the conditional ANOVA procedure with various combinations of the HOV and ANOVA tests and different alpha levels, a Monte Carlo simulation study was conducted.

Specifically, this study investigated the Type I error rates and statistical power of four robust ANOVA approaches coupled with five HOV methods, under a wide range of alpha levels for the HOV method. 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.

2. Statistical Tests Examined

Table 1 presents a summary of the HOV and ANOVA tests, including the test statistics and equations. Brief descriptions of each test are provided in this section.

2.1 Statistical Methods for Testing Homogeneity of Variance

2.1.1 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's 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 (*Fcrit*) with degrees of freedom $(k-1)$ and $(N-k)$ for numerator and denominator, respectively. The null hypothesis that the group variances are equal is rejected if $W > Fcrit$.

2.1.2 *Brown-Forsythe Test (BF_{HOV})*

This test (Brown & Forsythe, 1974) differs from the Levene's 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's test. The Brown-Forsythe test is more robust than the Levene's test with skewed distributions.

2.1.3 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.

2.1.4 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 is the weighted average of a modified Levene's squared deviations. The weighted average, $r_{ij}(w)$ is a modification of Levene's squared deviations from the group mean (w = 0), and a jackknife pseudo value of ${S_j}^2$ $(w = 1)$. 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, $\overline{r_j} = \frac{\sum r_{ij}}{n_j} = {S_j}^2$.

2.1.5 Ramsey Conditional Test: Brown-Forsythe or O'Brien (Ramsey)

Ramsey (1994) proposed a conditional procedure based on the Brown-Forsythe method and the O'Brien method. He suggested the appropriate test between the two methods

should be selected conditional on 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 then is calculated and used to identify the population as platykurtic if $S \leq -1$, mesokurtic if $S = 0$, or leptokurtic if $S \geq 1$. The O'Brien method will be implemented if the data are platykurtic and the Brown-Forsythe method will be applied if the data are mesokurtic or leptokurtic.

2.2 Statistical Methods for Testing Mean Equality

2.2.1 The ANOVA F Test

The ANOVA *F* test has been commonly 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 the violations of the HOV assumption, especially when sample sizes are unequal across groups.

2.2.2 Brown-Forsythe (BF) Test

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:

$$\frac{1}{f} = \text{ and } c_i = \frac{(1-n_j/N)s_j^2}{\sum_{j=1}^{k}(1-n_j/N)s_j{}^2}.$$

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

Originated from the framework of structural equation modeling (SEM), 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 $y = v_j + \delta$, where v_j is a *p* x 1 vector of intercept values (or means) for group *j*, δ is a *p* x 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 δ 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.

2.2.4 SMM with Bartlett's Correction to the ML Test Statistic (SMM with Bartlett Correction or 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}.

2.2.5 Wilcox Test

The Wilcox method (Wilcox, 1988) was contrasted with James's second-order (James, 1951) method. The modification of the Wilcox's procedure was proposed by Wilcox (1989). The null hypothesis is rejected when the test statistic H_m exceeds the $(1 - \alpha)$ quantile of the chi-square 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).

3. 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 nonnormal 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 for small sample sizes 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. Ramsey and Ramsey (2007) observed that the Ramsey test had higher power than the BF_{HOV} test.

For the ANOVA tests, it has been long 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 with ML or 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 well controlled Type I error rates 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 with ML, 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. Thus this study compared the efficacy of 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 with ML, SMM with Bartlett correction, and Wilcox), which created 20 conditional ANOVA tests.

4. Method

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 2), variance pattern (7 patterns, see Table 3), 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 ($\gamma 1$ and $\gamma 2$ 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 $\gamma 1$ and $\gamma 2$ 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. We considered 11 alpha levels 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 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 .003, and a 95% confidence interval no wider than ± .006 (Robey & Barcikowski, 1992).

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, we also investigated the robustness of conditional ANOVA tests 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 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.

5. Results

5.1 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 distribution of Type I error rates for the conditional BF test (BF paired with all 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, whereas the Type I error control of the robust ANOVA tests was inferior to that of the *F* test. Therefore, 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 the Bradley's criterion decreased from 1 close to that of the corresponding unconditional test as the alpha of HOV tests increased.

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, Bartlett, and ML as tests of means had adequate Type I error control with normal data. When data were nonnormal, the BF controlled Type I error rates better than Bartlett and 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 the Type I error control of tests of means paired with different HOV tests, conditional tests using Levene and OB outperformed those using BF_{HOV}, Ramsey, and bootstrap BF_{HOV} across all simulation conditions. For example, the proportions of conditions meeting the Bradley's criterion were .60 and .55 for conditional Bartlett with Levene and 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.

5.2 Type I Error Rates under Heterogeneous Variances

The series of boxplots in 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. As displayed in Figure 2, the performance of the conditional tests became closer to their unconditional counterparts as the alpha level of HOV tests increased. Bartlett performed slightly better than ML, followed by BF, and Wilcox had the worst 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 the Bradley's liberal criterion were examined. In addition, as can be seen from Figure 3, among the conditional tests, Bartlett, 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 the rest based on the largest proportion of replications that met the Bradley's criterion across all alpha levels of HOV.

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 4 presents the Bradley results by test, cell size, and variance pattern. Note that only a few selected conditional tests are presented, including BF, Bartlett, and 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 Bartlett and the conditional ML across all variance patterns. Put it another way, increasing the cell size improved the Type I error control substantially

for Bartlett and ML, but BF seemed to be least affected in terms of Type I error rates by cell size.

In addition, we examined the Type I error rates and proportions of conditions meeting the Bradley's liberal criterion (see Table 5) 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 Bartlett and ML tests. When cell sizes were unequal, Bartlett and 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 the Bradley's criterion under split inversely and progressive inversely patterns, except when the cell size pattern was split.

5.3 Statistical Power Analyses

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, we selected 6 conditional tests that had adequate Type I error control to include in the power analyses. These conditional tests are the combinations of BF, Bartlett, and 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 nonnormal distributions for cell size of 5 and extremely nonnormal 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, we selected 1,035 homogeneous and 6,948 heterogeneous conditions to use in power analyses.

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 6). 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, 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 of .50. Among the six conditional tests, Bartlett and ML paired with OB tended to have higher power than the rest.

6. 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 Bartlett, BF, and 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. Between the Levene and OB tests, the latter resulted in conditional tests with more statistical power although the power advantages were small.

In addition, the choice among Bartlett, BF, and 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. Conversely, as sample size increased the ML and 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.

The selection of an alpha level for the test of variances is important because of its influence on the power of this test. 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. Alpha levels near the middle of the range examined in this study appear to be a reasonable compromise between these competing effects. 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.

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 well 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.

Table 1

Statistics of Methods for Testing Homogeneity of Variance (HOV) and Mean Equality

Test	Test Statistic	Notation
	HOV	
Levene (Squared Deviations)	$Z_{ij} = (Y_{ij} - \bar{Y}_{.j})^2$ $W = \frac{(N-k)\sum_{j=1}^{k} n_j(\bar{Z}_{.j} - \bar{Z}_{..})^2}{(k-1)\sum_{j=1}^{k}\sum_{i=1}^{n_j}(Z_{ij} - \bar{Z}_{.j})^2}$	Y_{ij} = raw score of individual *i* in group *j*; $\bar{Y}_{.j}$ = mean of the j^{th} group; $\bar{Z}_{.j}$ = group mean of Z_{ij}; $\bar{Z}_{..}$ = grand mean; N = total sample size; n_j = group *j* sample size; k = number of groups.
Brown-Forsythe (BF_{HOV})[a]	$Z_{ij} = \lvert Y_{ij} - \tilde{Y}_{.j} \rvert$ $W = \frac{(N-k)\sum_{j=1}^{k} n_j(\bar{Z}_{.j} - \bar{Z}_{..})^2}{(k-1)\sum_{j=1}^{k}\sum_{i=1}^{n_j}(Z_{ij} - \bar{Z}_{.j})^2}$	$\tilde{Y}_{.j}$ = median of the j^{th} group.
O'Brien	$r_{ij}(w) = \frac{(w + n_j - 2)n_j(Y_{ij} - \bar{Y}_{.j})^2 - w s_j{}^2(n_j - 1)}{(n_j - 1)(n_j - 2)}$	$s_j{}^2$ = within-group unbiased estimate of variance for group *j*; w ($0 \le w \le 1$) = weighting factor.
Ramsey Conditional Test	$b_2 = m_4/m_2{}^2$, where $m_r = \Sigma(Y_{ij} - \bar{Y}_{.j})^r / n_j$ $b_{2j} = \frac{\frac{\sum(Y_{ij} - \bar{Y}_{.j})^4}{n_j}}{\left[\frac{\sum(Y_{ij} - \bar{Y}_{.j})^2}{n_j}\right]^2}$	
	ANOVA	
F test	$F = \frac{\sum_{j=1}^{k} n_j(\bar{Y}_{.j} - \bar{Y}_{..})^2/(k-1)}{\sum_{j=1}^{k}(n_j - 1)s_j{}^2/(N-k)}$	

Brown-Forsythe (BF) Test	$F^* = \frac{\sum_{j=1}^{k} n_j\left(\bar{Y}_{.j} - \bar{Y}_{..}\right)^2}{\sum_{j=1}^{k}\left(1 - n_j/N\right)s_j{}^2}$	
SMM with ML	$T_{\mathrm{ML}} = (N-1)F_{\mathrm{ML}}$	F_{ML} is the ML fit function.
SMM with Bartlett's Correction to the ML Test Statistic (Bartlett)	$T_{\mathrm{BC}} = (N - p/3 - 2m/3 - 11/6)\, F_{\mathrm{ML}}$	p = number of observed variables (p = 1 in ANOVA); m= the number of latent constructs (m = 0 in ANOVA); q = number of parameters estimated across all groups.
Wilcox	$D_j = n_j/s_j{}^2$, $W_s = \sum D_j$, $Y^* = \sum D_j Y_j^*/W_s$, where $Y_j^* = Y_{n_j j}/n_j + \sum_{i=1}^{n_j-1}\left(1 - \frac{1}{n_j}\right)Y_{ij}/(n_j+1)$, $H_m = \sum D_j (Y_j^* - Y^*)^2$.	

Note. [a]The bootstrap version of the BF test was also evaluated. ANOVA = analysis of variance; SMM = structured means modeling; ML = maximum likelihood.

Table 2

Sample Size Patterns

Sample Sizes												
	Progressive *N*			Equal *N*			Split *N*			One Extreme		
K=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
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 3

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=8										
1	1	1	1	1	1	1	1	1	1	1
2	1.43	2	3.14	1	1	1	1	1	1	1
3	1.86	3	5.28	1	1	1	1	1	1	1
4	2.29	4	7.42	1	1	1	1	1	1	1
5	2.72	5	9.56	4	8	16	1	1	1	1
6	3.15	6	11.70	4	8	16	1	1	1	1
7	3.58	7	13.84	4	8	16	1	1	1	1
8	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	3.30	6	1	1	1	1	1	1	1
3	3	5.70	11	4	8	16	1	1	1	1
4	4	8	16	4	8	16	4	8	16	1

Population Variances									
	Progressive Inversely			Split Inversely			One Extreme Inversely		
Max Variance Ratio	4:1	8:1	16:1	4:1	8:1	16:1	4:1	8:1	16:1
K=8									
1	4	8	16	4	8	16	4	8	16
2	3.58	7	13.84	4	8	16	1	1	1
3	3.15	6	11.70	4	8	16	1	1	1
4	2.72	5	9.56	4	8	16	1	1	1
5	2.29	4	7.42	1	1	1	1	1	1
6	1.86	3	5.28	1	1	1	1	1	1
7	1.43	2	3.14	1	1	1	1	1	1
8	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 4

Proportions of Conditions That Met the Bradley's Liberal Criterion by Test, Cell Size, and Variance Pattern under Variance Heterogeneity

Test	Cell Size 5						Cell Size 10					
	Variance Patterns						Variance Patterns					
	2	3	4	5	6	7	2	3	4	5	6	7
BF_LV	69	84		19	60		54	89		24	85	
	8	3	862	6	1	622	6	6	963	9	7	880
BAR_L	66	68		25	37		87	89		54	72	
V	5	8	743	1	8	383	2	1	907	4	3	677
ML_LV	64	71		21	31		86	90		50	68	
	3	3	767	7	7	309	6	2	908	5	4	625
BF_OB	68	81		14	43		55	89		24	82	
	4	3	867	2	9	460	1	5	947	7	8	857
BAR_O	63	67		17	28		86	88		52	70	
B	8	3	753	0	5	305	0	1	888	9	8	658
ML_OB	61	69		15	23		85	88		48	66	
	4	1	770	0	9	249	7	9	899	9	7	607
OLS	56	50		05	09		54	45		06	11	
	3	0	632	6	0	236	9	8	597	3	1	250
BF	71	95	100	29	88		47	87	100	36	97	100
	5	1	0	9	2	875	2	5	0	1	9	0
BAR	77	77		70	68		92	91		84	84	
	1	8	792	8	8	674	4	7	903	0	0	826
ML	77	79		70	66		88	90		81	80	
	1	9	785	8	0	653	9	3	882	3	6	799

Test	Cell Size 20					
	Variance Patterns					
	2	3	4	5	6	7
BF_LV	48	77		31	79	
	1	1	990	1	9	953
BAR_L	92	94		79	82	
V	6	1	951	7	4	795
ML_LV	91	93		78	81	
	6	2	949	2	6	778
BF_OB	48	77		31	79	
	2	6	985	5	2	946
BAR_O	92	94		79	82	
B	6	1	948	9	2	793
ML_OB	91	93		78	81	
	7	4	948	0	4	775
OLS	58	52		06	12	
	3	1	590	9	5	250
BF	45	66	100	36	84	100
	8	0	0	8	7	0
BAR	93	91		88	84	
	8	7	910	9	0	826
ML	91	91		87	84	
	0	0	910	5	0	819

Note. OLS = the ANOVA *F* test with ordinary least squares; BF = the Brown-Forsythe test; BAR = structured means modeling (SMM) with Bartlett's correction to the maximum likelihood (ML) test statistic; ML = SMM with ML estimation; 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 patterns 2 = one extreme, 3 = split, 4 = progressive, 5 = one extreme inversely, 6 = split inversely, and 7 = progressive inversely. The value of proportion for each cell should be divided by 1000.

Table 5

Proportions of Conditions That Met the Bradley's Liberal Criterion by Test, Cell Size Pattern and Variance Pattern under Variance Heterogeneity

Test	Progressive Cell Sizes						Equal Cell Sizes					
	Variance Patterns						Variance Patterns					
	2	3	4	5	6	7	2	3	4	5	6	7
BF_LV	60	88		16	73	76	38	87	100	41	86	
	2	7	968	3	9	3	9	1	0	7	4	994
BAR_LV	88	92		39	62	57	78	83		79	82	
	9	7	918	1	9	9	2	5	860	1	7	871
ML_LV	85	92		36	58	52	74	81		74	79	
	3	7	934	9	6	9	6	0	833	8	5	840
BF_OB	61	88		14	63	66	37	84	100	39	83	
	7	9	949	6	9	2	5	7	0	9	8	996
BAR_O	89	93		36	55	50	73	82		73	80	
B	3	2	907	3	6	8	1	1	864	8	6	884
ML_OB	86	93		34	52	47	70	79		70	77	
	4	2	920	5	4	4	0	4	838	1	4	853
OLS	97	75		00	00	00	24	44		25	43	
	2	0	593	0	0	0	1	4	981	0	5	981
BF	56	84	100	25	90	93	42	89	100	45	89	
	5	3	0	0	7	5	6	8	0	4	8	991
BAR	87	86		83	77	77	90	83		88	85	
	0	1	907	3	8	8	7	3	833	9	2	824
ML	84	85		80	75	76	88	84		88	83	
	3	2	907	6	0	9	9	3	824	9	3	843
Test	Split Cell Sizes						One Extreme Cell Sizes					
	Variance Patterns						Variance Patterns					
	2	3	4	5	6	7	2	3	4	5	6	7
BF_LV	65	71		14	60	64	65	87		28	80	
	6	8	855	1	3	4	4	0	929	8	3	871
BAR_LV	83	69		37	46	44	78	90		56	65	
	3	7	822	3	1	0	0	1	867	7	0	582
ML_LV	81	75		34	43	39	81	90		54	61	
	7	2	854	8	0	7	8	8	878	0	3	518
BF_OB	66	69		12	54	56	63	88		26	72	
	2	3	868	5	2	8	5	4	915	9	7	791
BAR_O	83	66		36	45	42	77	91		52	60	
B	6	4	812	8	3	0	3	1	869	9	6	529
ML_OB	82	70		34	42	38	80	92		50	56	
	1	4	852	4	6	2	0	3	880	1	9	465

OLS	95	01		00	00	00	09	75		00	00	
	4	9	241	0	0	0	3	9	611	0	0	000
BF	58	79	100	25	85	90	62	77	100	41	95	100
	3	6	0	0	2	7	0	8	0	7	4	0
BAR	82	91		75	75	75	90	87		77	77	
	4	7	889	0	0	0	7	0	843	8	8	750
ML	75	92		73	71	70	93	86		76	77	
	9	6	880	1	3	4	5	1	824	9	8	713

Note. OLS = the ANOVA *F* test with ordinary least squares; BF = the Brown-Forsythe test; BAR = structured means modeling (SMM) with Bartlett's correction to the maximum likelihood (ML) test statistic; ML = SMM with ML estimation; 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 patterns 2 = one extreme, 3 = split, 4 = progressive, 5 = one extreme inversely, 6 = split inversely, and 7 = progressive inversely. The value of proportion for each cell should be divided by 1000.

Table 6

Statistical Power for Conditional ANOVA Tests under Homogeneous and Heterogeneous Variances Conditions at Different Alpha Levels

Test	α=01	α=05	α=10	α=15	α=20	α=25	α=30	α=35	α=40	α=45	α=50
Homogeneous Variances											
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										
Heterogeneous Variances											
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 = the ANOVA *F* test with ordinary least squares; BF = the Brown-Forsythe test; BAR = structured means modeling (SMM) with Bartlett's correction to the maximum

likelihood (ML) test statistic; ML = SMM with ML estimation. 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. The value of statistical power for each cell should be divided by 1000, and the value for each alpha level should be divided by 100.

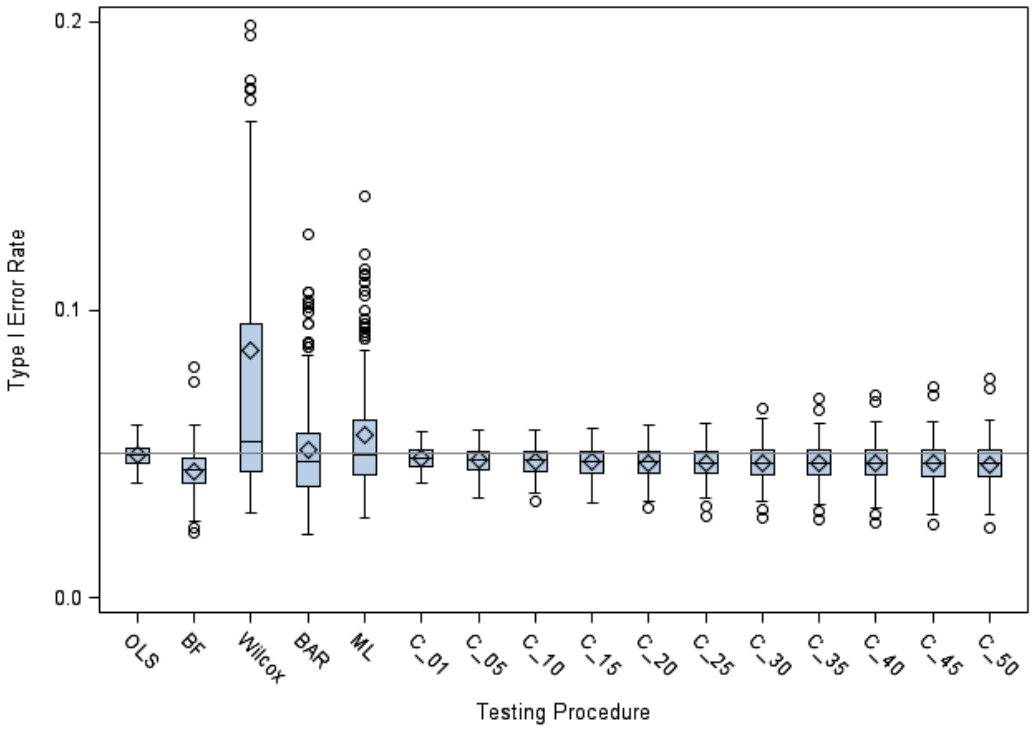

Figure 1 : Distribution of Type I error rates for unconditional tests and conditional Brown-Forsythe robust ANOVA test with combinations of HOV tests denoted by C_01 to C_50. OLS = the ANOVA *F* test with ordinary least squares; BF = the Brown-Forsythe test; BAR = structured means modeling (SMM) with Bartlett's correction to the maximum likelihood (ML) test statistic; ML = SMM with ML estimation.

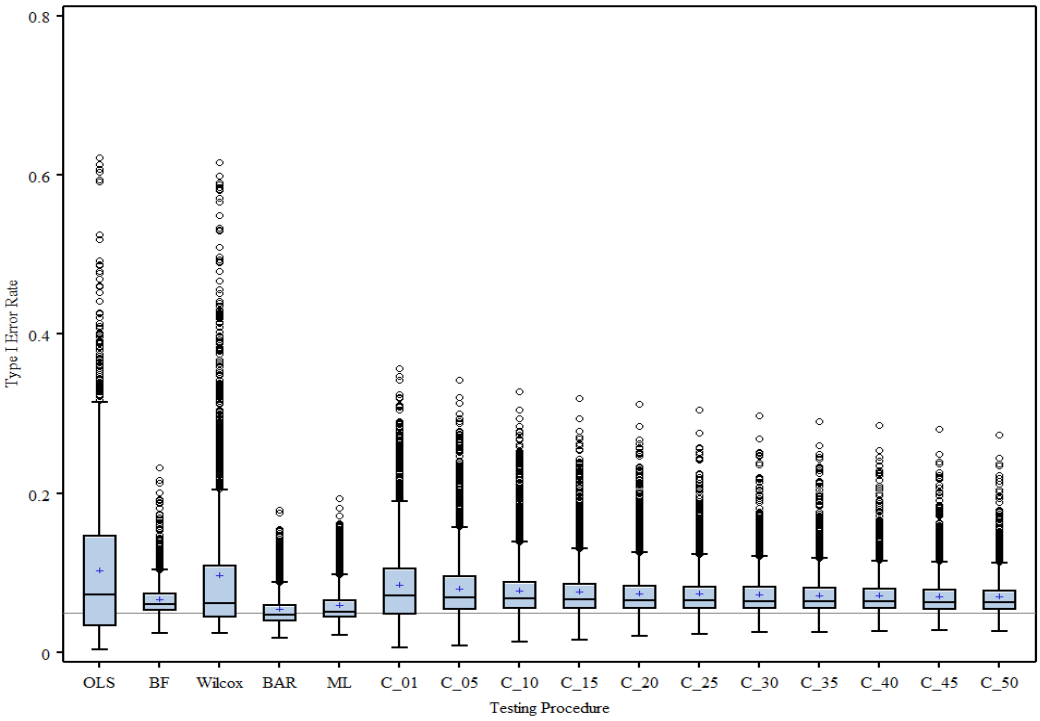

Figure 2 : Distribution of Type I error rates under variance heterogeneity for unconditional tests and conditional Brown-Forsythe robust ANOVA test with the Brown-Forsythe test for homogeneity of variance (HOV). OLS = the ANOVA *F* test with ordinary least squares; BF = the Brown-Forsythe test; Bartlett = structured means modeling (SMM) with Bartlett's correction to the maximum likelihood (ML) test statistic; ML = SMM with ML estimation. C_01 to C_50 denote the alpha level of the HOV test was .01 to .50.

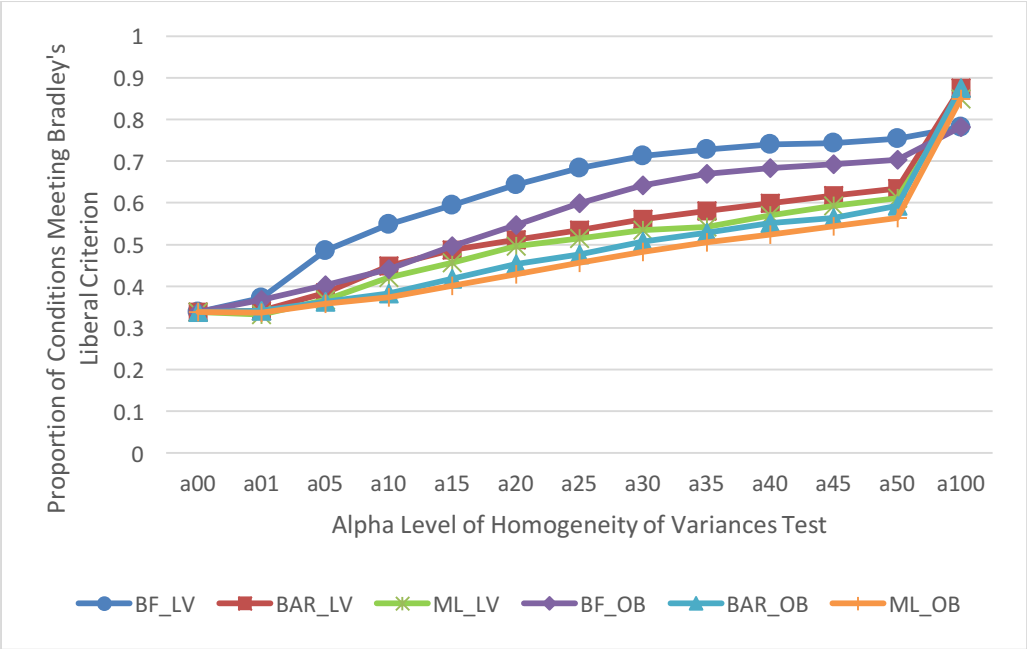

Figure 3 : Proportion of conditions meeting Bradley's criterion with cell size 10 under variance heterogeneity. Note that a00 represents the ANOVA *F* test and a100 represents

the unconditional test of the corresponding conditional test. BF_LV, BAR_LV, and ML_LV refer to the Brown-Forsythe test, structured means modeling (SMM) with Bartlett's correction to the maximum likelihood (ML) test statistic, and SMM with ML estimation 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.

References

Ames, H., Wilson, C., Barnett, S., Njoh, E., & Ottomanelli, L. 2017. Rehabilitation Psychology. Advance online publication.

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

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

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

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

Cohen, J. 1992. A power primer. *Psychological bulletin, 112*(1), pp. 155-159.

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), pp. 137-156.

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

Glass, G. V., Peckham, P. D., & Sanders, J. R. 1972. Consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance. *Review of Educational Research, 42*(3), pp. 237-288.

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), pp. 315-339.

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

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), pp. 350-386.

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), pp. 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*. Palo Alto, CA: Stanford University Press, pp. 278–292.

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), pp. 579-619.

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), pp. 59-74.

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), pp. 1656-1671.

Nguyen, D. T., Kim, E. S., Wang, Y., Pham, T. V., Kromrey, J., & Chen, Y.-H. 2016. *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), pp. 877-880.

O'Brien, R. G. 1981. A simple test for variance effects in experimental designs. *Psychological Bulletin, 89*(3), pp. 570-574.

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

Parra-Frutos, I. 2012. Testing homogeneity of variances with unequal sample sizes. *Computational Statistics, 28*(3), pp. 1269-1297.

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

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), pp. 231-241.

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

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), pp. 283-288.

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), pp. 493-498.

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), pp. 1944-1963.

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), pp. 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), pp. 269-278.

Walsh, A. St. 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), pp. 851-857.

Wang, Y., Rodriguez de Gil, P., Chen, Y.-H., Kromrey, J. D., Kim, E. S., Nguyen, D. T., Pham, 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), p. 305-329.

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