Comparing the Performance of Approaches for Testing the Homogeneity of Variance Assumption in One-Factor ANOVA Models

Yan Wang, Aarti P. Bellara, Thanh V. Pham, Diep T. Nguyen, Patricia Rodriguez de Gil, Yi-Hsin Chen, Harold Holmes, Tyler Hicks, Isaac Li, Eun Sook Kim, Jeanine Romano, & Jeffrey D. Kromrey
University of South Florida, 4202 E. Fowler Avenue, Tampa, FL 33620

Abstract
The validity of the results of an ANOVA test is largely dependent on satisfying the homogeneity of variance, normality, and independence assumptions. Violations of these assumptions lead to distorted Type I error rates. Various tests to check the homogeneity of variance assumption for non-normal data have been proposed in the literature, yet there is no consensus as to which test is most appropriate. A simulation study was conducted to explore the Type I error rates and statistical power of fourteen approaches for testing the homogeneity of variance assumption in one-way ANOVA models. Seven factors were manipulated in the study: number of groups, average number of observations per group, pattern of sample sizes in groups, pattern of population variances, maximum variance ratio, population distribution shape, and nominal alpha level for the test of variances. Results from this study delineate the performance of the tests under a wide variety of conditions, providing researchers with information to guide the selection of a valid test for assessing the tenability of this critical assumption.

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

1. Introduction

In an ANOVA procedure, the assumption of homogeneity of variance (HOV) is that treatment variances are equal. That is,

$$H_0 : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$$

Moderate deviations from the assumption of equal variances may not seriously affect the results in ANOVA (Glass, Peckham, & Sanders, 1972). Because the ANOVA procedure may be robust to small deviations from the HOV assumption, researchers may only need to be concerned about large deviations from the HOV assumption. However, the tests used to evaluate HOV are sensitive to departures of normality, for which researchers should turn to alternative tests when the assumption of normality is not met. In this study, we assemble fourteen HOV tests. Some of these methods are prevailing and available in statistical analytical software packages (e.g., Statistical Analysis System or SAS). And many are alternatives of the HOV testing approaches that have been proposed in literature but are not included in the existing software packages. Table 1 presents all the HOV methods evaluated in this study, followed by the test statistic and mathematical equation of each method. These fourteen approaches are elaborated in detail below.

Table 1: Alternative Homogeneity of Variance Tests Statistics and Distributions.

HOV Test	*Test Statistic and Distribution*	*Notation*
Bartlett	$\chi^2 = \frac{(N-k)log\left[\frac{\sum_{j=1}^{k}(n_j-1)S_j^2}{(N-k)}\right]-\sum_{j=1}^{k}(n_j-1)\log(S_j^2)}{1+\frac{\left(\sum_{j=1}^{k}\frac{1}{n_j-1}\right)-\frac{1}{(N-k)}}{3(k-1)}}$,	N = total sample size; N_j = group j sample size; k = number of groups; S_j^2 = group j variance.
Levene (Absolute and Squared)	$Z_{ij} = \left\|Y_{ij} - \bar{Y}_{.j}\right\|$ *and* $Z_{ij} = (Y_{ij} - \bar{Y}_{.j})^2$, $W = \frac{(N-k)\sum_{j=1}^{k} n_j(\bar{Z}_{.j}-\bar{Z}_{..})^2}{(k-1)\sum_{j=1}^{k}\sum_{i=1}^{n_j}(Z_{ij}-\bar{Z}_{.j})^2}$	Y_{ij} = raw score; $\bar{Y}_{.j}$ = mean of the j^{th} group; $\bar{Z}_{.j}$ = group mean of Z_{ij}; $\bar{Z}_{..}$ = grand mean.
Brown-Forsythe (BF)[a]	$z_{ij} = \left\|Y_{ij} - \tilde{Y}_{ji}\right\|$, $W = \frac{(N-k)\sum_{j=1}^{k} n_j(\bar{Z}_{.j} - \bar{Z}_{..})^2}{(k-1)\sum_{j=1}^{k}\sum_{i=1}^{n_j}(Z_{ij} - \bar{Z}_{.j})^2}$	$\tilde{Y}_j$ = median of group j; z_{ij} = transformed value of $Y_{ij;}$ $\bar{Z}_{.j}$ = group mean of Z_{ij}; $\bar{Z}_{..}$ = grand mean.
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 sample j; w ($0 \le w \le 1$) = weighting factor.
Ramsey	$b_2 = m_4/m_2^2$, $b_{2j} = \frac{\frac{\sum(Y_{ij}-\bar{Y}_{.j})^4}{n_j}}{\left[\frac{\sum(Y_{ij}-\bar{Y}_{.j})^2}{n_j}\right]^2}$.	$m_r = \Sigma(Y_{ij} - \bar{Y}_{.j})^r / n_j$.
Cochran's C[b]	$C = \frac{s_{max}^2}{\sum s_j^2}$, $Critical\ C = \frac{1}{1+\frac{F_{\frac{\alpha}{k},\ n-1,(k-1)(n-1)}}{k-1}}$.	n = number of observations in each group (for the balanced design); F = critical value of F at α/k with df = n-1, (k-1)(n-1).
G test	$G = \frac{v_{max}{s_{max}}^2}{\sum v_j {s_j}^2}$, $Critical\ G = \frac{1}{1+\frac{v_{pool}/(v_{max}-1)}{F_{\frac{\alpha}{k},\ \overline{n_j}-1,(k-1)(\overline{n_j}-1)}}}$.	v_{pool} = pooled degrees of freedom; v_{max} = degrees of freedom for the group within the largest variance; $\overline{n_j}$ = mean number of observations in each group; $F_{\frac{\alpha}{k}}$ = critical value of F at α/k with df = $\overline{n_j} - 1$, $(k-1)(\overline{n_j}-1)$.
F_{max}[b]	$F_{max} = \frac{s_{max}^2}{s_{min}^2}$	s_{max}= largest group variance; s_{min}= smallest group variance.

Z-variance	$z = \sqrt{2\chi^2} - \sqrt{2(df) - 1}$, $\chi_{n-1}{}^2 = \frac{(n-1)S^2}{\sigma^2}$, $Z_j = \sqrt{\frac{c(n_j-1)S_j{}^2}{MS_w}} - \sqrt{c(n_j - 1) - 1}$, $F = \frac{\sum_{j=1}^{k} Z_j{}^2}{k-1}$.	s^2 = sample variance estimate; σ^2 = true population variance; $c = 2 + 1/n_j$; MS_w= pooled within-cells mean square across all groups (or cells in a more complex factorial design).
Modified *Z*-variance	$c = 2.0\left(\frac{2.9+.2/n_j}{K}\right)^{1.6(n_j-1.8K+14.7)/n_j}$, $K = \frac{\sum Z_{ij}^4}{n_j-2}$, $Z_{ij} = \frac{Y_{ij}-\bar{Y}_{.j}}{\sqrt{\frac{n_j-1}{n_j}S_j^2}}$.	K = mean of the kurtosis indices from all groups.

Note: [a]The bootstrap version of the BF test was also evaluated. [b]With arithmetic mean and harmonic mean for the group size under unbalanced design.

1.1 Methods Available in SAS

Bartlett Test. Bartlett (1937) proposed a special use of the chi-square test for testing the HOV assumption, under which the null hypothesis of equal variances will be rejected if the Bartlett's χ^2 is greater than the critical χ^2 value with *df = k-1*. However, Snedecor and Cochran (1989) found that the Bartlett's test is sensitive to non-normal distributions and instead recommended alternative testing approaches: Levene (Absolute and Squared), Brown-Forsythe (BF), and O'Brien tests.

Levene Test. Levene (1960) proposed the use of absolute residual values or squared residuals, which transforms the test of variances into a test of means that is relatively robust to the normality assumption. The *W* statistics of the absolute residual values and squared residuals are compared to the *F* critical value with *N-k* and *k*-1 as *df* in the numerator and denominator, respectively.

Brown-Forsythe (BF) Test. Brown and Forsythe (1974) proposed the Brown-Forsythe (BF) test that follows the idea of Levene's test but uses the group median instead of the group mean in the calculation of the absolute residual values. It is expected to be more robust than Levene's test when the population distribution is skewed.

O'Brien Conditional Test. O'Brien (1979) proposed a test that transforms original scores so they represent sample variances. The mean of the transformed values per group, $\overline{r_j} = \frac{\sum r_{ij}}{n_j} = s_j{}^2$, will equal the variance computed for that group. The weighted average, $r_{ij}(w)$, is a modification of Levene's squared difference from the group mean (w = 0), and a jackknife pseudo value of $s_j{}^2$ ($w = 1$). O'Brien (1981) suggested setting w = .5 as default.

These aforementioned approaches that are available in the existing software have been well-examined. Snedecor and Cochran (1989) stated that the Bartlett's test is sensitive to

violations of the normality assumption. Therefore, statisticians do not recommend its use for testing the HOV assumption when data depart from normality and recommend instead alternative testing approaches that are not sensitive to departures from normality, namely, Levene, BF, and O'Brien tests.

However, simulation studies (e.g., Algina, Olejnick, & Ocanto, 1989; Conover, Johnson, & Johnson, 1981; Lee, Katz, & Restory, 2010; Olejnick & Algina, 1987) have showed differences among these tests. The O'Brien test provided Type I error rates near the nominal alpha in unbalanced samples but with platykurtic distributions it was more sensitive to variance differences than the BF test. When sample sizes were equal, O'Brien had a power advantage used with platykurtic distributions and had a slight power advantage when used with mesokurtic distributions regardless of whether the distributions were skewed or not. On the other hand, BF had a power advantage used with leptokurtic distributions regardless of the skewness. When sample sizes were unequal, results were different because the relative power of these tests depended on the direction of the relationship between the population variances and sample sizes. For example, the power of the O'Brien test was improved when used with skewed-platykurtic and symmetric-platykurtic distributions. The power of BF was also enhanced when the relationship between population variance and sample size was direct (i.e. larger samples come from populations with larger variances) and with leptokurtic or mesokurtic distributions. With other distributions, the tests had similar power.

1.2 Methods Not Available in SAS

Bootstrap Brown-Forsythe Test. Boos and Brownie (2004) as well as Lim and Loh (1996) recommended using the median version of Levene's test statistic (i.e., the BF statistic), then obtaining the *p*-value via the bootstrap, which provided more power than the *F* distribution version.

Ramsey Conditional Test: Brown-Forsythe or O'Brien. Ramsey's (1994) conditional procedure is based on using BF or OB method, conditional on a test of kurtosis. Kurtosis (b_2) for each of the *k* groups is computed by using Pearson's traditional measure; the kurtosis value for each group is then compared to critical values obtained from a table provided by Ramsey and Ramsey (1993). The O'Brien test will be applied if the data are platykurtic and the BF test will be used if the data are mesokurtic or leptokurtic.

Cochran's C test. Cochran's *C* test (Cochran, 1941) is a ratio of the largest group variance to the sum of sample variances. If the obtained value exceeds the critical value, the null hypothesis of variance homogeneity is rejected. For an unbalanced design one could use either an arithmetic mean of n_j or the harmonic mean of n_j to calculate degrees of freedom. Both were included in our study.

G Test. The G test is a ratio of the product of the largest variance and its degrees of freedom to the sum of the products of each variance and its degrees of freedom. If the obtained value exceeds the critical value, the null hypothesis of variance homogeneity is rejected.

F-max Test. Hartley (1950) developed the Hartley's or F_{max} test for comparing three or more group variances, which is a ratio of the largest group variance to the smallest group variance and requires independent random samples of the same size from normally distributed populations (Ott & Longnecker, 2010). The value of F_{max} is compared to a

critical value from the table containing the test sampling distribution. Similar to the Cochran's C test, this study included the use of both the arithmetic and harmonic means of n_j for an unbalanced design.

Z-variance Test. Overall and Woodward (1974) proposed the *Z*-variance test based on Fisher and Yates' formula (1963). It transforms the chi-square statistics with large *df* into standard normal deviate *z*-scores. It performed very well with normally distributed data but produced too many Type I errors when samples were from leptokurtic or skewed distributions.

Modified Z-variance Test. To improve the performance of the *Z*-variance test when sample distributions are leptokurtic or skewed, Overall and Woodward (1976) developed the modified *Z*-variance approach to testing the HOV by implementing a *c* value based on sample size, skewness, and kurtosis.

1.3. Research Purpose

Considering the ongoing controversy on testing the homogeneity of variance and the minimal consensus among studies as to which test is appropriate for a particular analysis, two over-arching research questions guide this area of inquiry: (a) Does it make sense to statistically test the homogeneity of variance assumption? (b) What method should we use for testing the homogeneity of variance assumption? To our knowledge, there are no studies that examine all the HOV tests proposed in this study simultaneously. Thus, the goal of the current paper is to conduct a comprehensive examination of fourteen approaches for testing the homogeneity of variance assumption in one-way ANOVA models under diverse data conditions. The specific questions are as follows: (a) Which HOV tests possess adequate control of the Type I error and competitive power? (b) Which design factors have significant impacts on Type I error control and statistical power among the HOV tests?

2. Method

This study was conducted using a simulation approach. The six design factors manipulated in this study included: (a) number of groups (k = 4 and 6), (b) average number of observations per group (n = 5, 10, and 20), (c) sample size pattern (equal, one extreme, split, and progressive), (d) variance pattern (equal, one extreme, split, progressive, one extreme inversely, split inversely, and progressive inversely), (e) maximum group variance (4, 8, and 16), and (f) population distribution ($\gamma1 = 0.00$ and $\gamma2 = 0.00$, $\gamma1 = 1.00$ and $\gamma2 = 3.00$, $\gamma1 = 1.50$ and $\gamma2 = 5.00$, $\gamma1 = 2.00$ and $\gamma2 = 6.00$, $\gamma1 = 0.00$ and $\gamma2 = 25.00$, and $\gamma1 = 0.00$ and $\gamma2 = -1.00$, where $\gamma1$ and $\gamma2$ represent skewness and kurtosis, respectively). Non-normal populations were generated by implementing the Fleishman's transformation (Fleishman, 1978). Table 2 shows four sample size patterns and Table 3 presents seven variance patterns. In addition to six designed factors, we used six alpha levels for testing the homogeneity assumption (α = .01, .05, .10, .15, .20, and .25). Thus, this crossed mixed factorial design had a total of 16,416 conditions in this study. For each condition, 5,000 replications were generated.

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 simulation program. For each condition in the simulation, 5,000 samples were generated. The use of 5,000 replications provides a maximum standard error of an

observed proportion (e.g., Type I error rate estimate) of .00158, and a 95% confidence interval no wider than ± .003 (Robey & Barcikowski, 1992).

We examined Type I error and statistical power as the simulation outcomes. Type I error was examined when the variances across groups were equal (i.e., equal variance pattern); otherwise, power was computed. For Type I error, we further investigated the robustness using the Bradley's liberal criterion. The liberal criterion for the robustness is set at .5α around the 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) using the alpha level of .05. Finally, eta-square analyses were conducted to explore the significant impacts of design factors on variability in the estimated Type I error. The Cohen's (1988) moderate effect size of .0588 was set as a cutoff value for eta-square analyses.

Table 2: Sample Size Pattern.

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

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

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

(Cont'd)

	Population Variances								
	Progressive Inversely			*Split Inversely*			*One Extreme Inversely*		
Max Variance Ratio	4:1	8:1	16:1	4:1	8:1	16:1	4:1	8:1	16:1
K=6									
1	4	8	16	4	8	16	4	8	16
2	3.4	6.6	13	4	8	16	1	1	1
3	2.8	5.2	10	4	8	16	1	1	1
4	2.2	3.8	7	1	1	1	1	1	1
5	1.6	2.4	4	1	1	1	1	1	1
6	1	1	1	1	1	1	1	1	1
K=4									
1	4	8	16	4	8	16	4	8	16
2	3	5.7	11	4	8	16	1	1	1
3	2	3.3	6	1	1	1	1	1	1
4	1	1	1	1	1	1	1	1	1

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

3. Results

3.1 Type I Error Rate Estimates

Boxplots were first checked to examine the distributions of Type I error rate estimates for the fourteen HOV tests across all simulation conditions at each nominal alpha level. Figure 1 presents a set of the boxplots at the nominal level of .05. The results for different nominal levels (i.e., .01, .10, .15, .20, and .25) are consistent with those at .05. As shown in Figure 1, the Levene test with squared deviations, the BF test, the Bootstrap BF test, the O'Brien test, and the Ramsey conditional test were the five best testing approaches that controlled Type I error adequately.

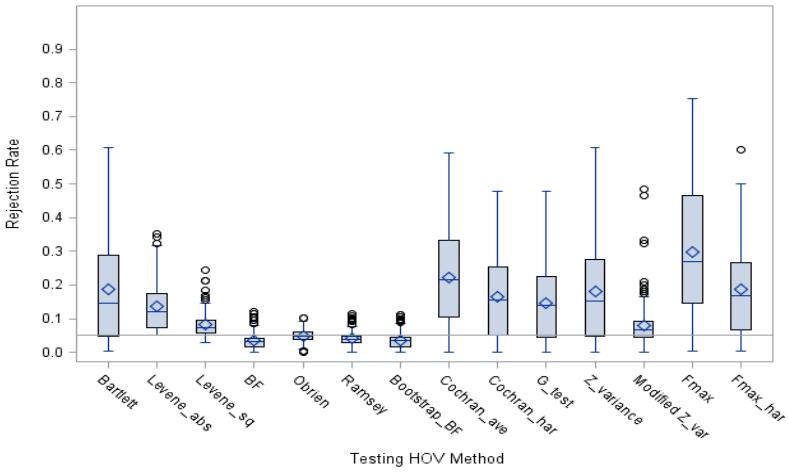

Figure 1: Distributions of estimated Type I error rates across all simulation conditions at .05. The horizontal line is the reference line at the nominal level of .05.

3.2 A Closer Look at Type I Error Control: Bradley's Liberal Criterion

The performance of the HOV tests in terms of the Type I error control was further examined using Bradley's liberal criterion of robustness, which is set at .5α around the nominal alpha. For each nominal alpha level investigated, the value of proportion meeting Bradley's Liberal Criterion is the proportion of Type I error rates (across all simulation conditions with 5,000 replications simulated for each condition) that fell within the range of $\alpha \pm .5\alpha$. Figure 2 shows the overall proportion meeting Bradley's criterion for the fourteen HOV tests at each nominal alpha level. Consistent with the results shown in Figure 1, the O'Brien test, the Ramsey conditional test, the Bootstrap BF test, the BF test, and the Levene test with squared deviations had the larger proportions of conditions meeting Bradley's criterion across all nominal alpha levels, compared with other tests. For instance, the proportions of those five tests that met the Bradley's criterion were .77, .72, .58, .58, and .52 respectively at the nominal level of .05.

As the nominal alpha level increased, the proportions meeting Bradley's criterion for the fourteen HOV tests all increased. It is worthwhile noting that the modified Z-variance test had a considerable increase when the nominal alpha level increased and the Levene test with absolute deviations also had a significant increase when the nominal level was at .15 or above. At the nominal level of .25, all the HOV tests had at least 50% of conditions meeting the Bradley's criterion.

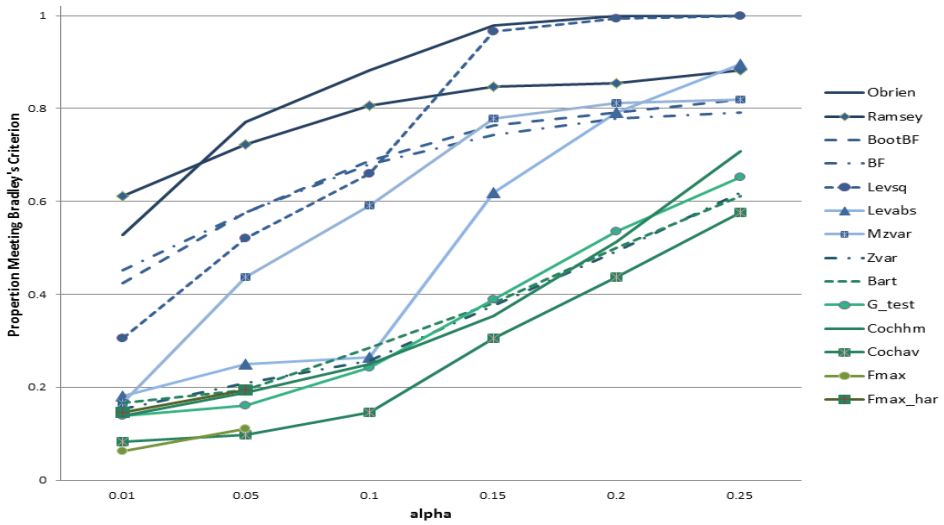

Figure 2: Overall proportion conditions of the fourteen HOV tests meeting Bradley's liberal criterion.

3.3 Impact of Design Factors on the Type I Error Control

Additionally, we conducted factorial ANOVA analyses to estimate the generalized eta square effect size (η_G^2). Cohen's (1988) moderate effect size of .0588 was used a cutoff value to indicate the significant impact of design factors on the Type I error estimates. Among all design factors, population shape and average cell size were two factors that had the largest impacts on Type I error estimates across all the HOV tests.

Figure 3 displays the proportions of conditions that met the Bradley's liberal criterion for each test across all population distribution shapes. The O'Brien and Ramsey conditional tests tended to have adequate Type I error control (i.e., the high proportions meeting the Bradley's criterion) across all population distribution shapes. For the BF, Bootstrap BF, Levene with squared deviations, and modified Z-variance tests, the Type I error control was not adequate under some population shapes. For instance, the BF and Bootstrap BF tests did not perform well in terms of the Type I error control when population distributions were leptokurtic (skewness = 0, kurtosis = 25). The Levene test with squared deviations had a poor control of Type I error under skewed distributions (skewness = 5 or 6). When data were normally distributed, however, the Bartlett, Cochran with the harmonic mean, and Z-variance tests controlled Type I error extremely well so that the proportions meeting Bradley's criterion were all 1.00.

Figure 4 summarizes the performance of the fourteen HOV tests in terms of Type I error estimates under different population distribution shapes. It was constructed based on the results of Bradley's liberal criterion of robustness and the box plots which showed the distributions of Type I error estimates of those tests under different population shapes. The shaded areas indicated adequate Type I error control, i.e. the proportions meeting the Bradley's liberal criterion were equal to and greater than .50. If the proportions were lower than .50, the test might be either liberal or conservative, depending on whether the distributions of Type I error rates were mostly above or below the reference line, as shown in the box plots. Consistent with the findings presented in Figure 3, the O'Brien test and the Ramsey conditional test showed adequate Type I error control across all

distribution shapes. The BF and Bootstrap BF tests were conservative under leptokurtic distributions (i.e., Shape 5: skewness = 0, kurtosis = 25), while the Levene test with squared deviations was liberal when distributions were skewed (i.e., Shapes 3 and 4). Under the normal distribution, almost all HOV tests had an adequate Type I error control except for the Cochran's C test with the arithmetic means and the F-max test.
Under the skewed and leptokurtic distributions, the HOV tests that did not have the adequate Type I error control had liberal Type I errors, whereas under the platykurtic distributions (i.e., Shape 6: skewness = 0, kurtosis = -1) they tended to have conservative Type I errors.

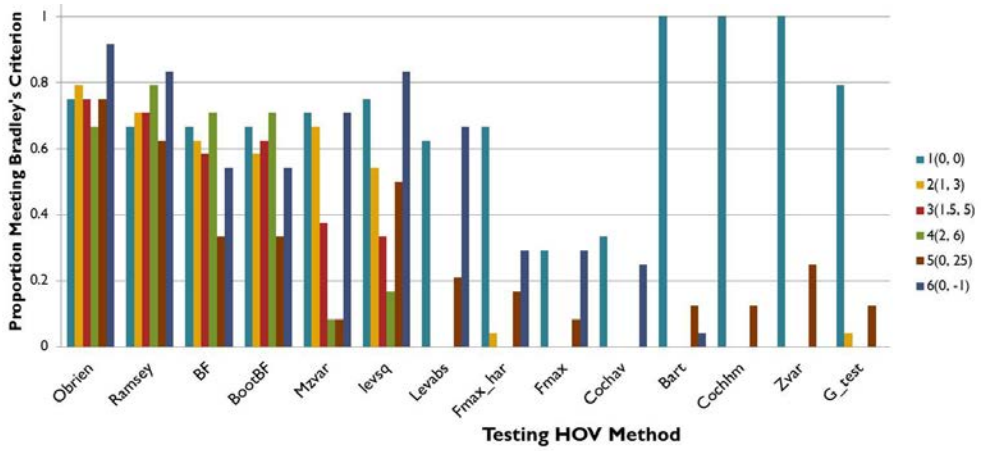

Figure 3: Proportions meeting the Bradley's liberal criterion for the fourteen HOV tests by population shape (alpha=.05).

	Population Shape (skewness, kurtosis)					
ANOVA HOV Test	1(0,0)	2(1,3)	3(1.5,5)	4(2,6)	5(0,25)	6(0,-1)
Barlett		L	L	L	L	C
Leveneabs		L	L	L	L	
Levenesq			L	L		
BF					C	
Obrien						
Ramsey						
Bootbf					C	
Cochav	L	L	L	L	L	C
Cochhm		L	L	L	L	C
G_test		L	L	L	L	C
Zvar		L	L	L	L	C
Mzvar			L	L	L	
Fmax	L	L	L	L	L	L
Fmax_har		L	L	L	L	C

Figure 4: The performance of HOV tests under different population distribution shapes.

In addition to population distribution shape, average cell size was the other factor that had a significant impact on Type I error estimates across the fourteen HOV tests. Figure 5 presents the proportions of conditions meeting Bradley's criterion for all HOV tests by

average cell size. It can be noticed that the O'Brien test tended to have adequate Type I error control regardless of average cell sizes. For the Ramsey conditional test, BF, Bootstrap BF, Levene with squared deviations, and Levene with absolute deviations tests, the proportions meeting Bradley's criterion increased as average cell size increased. And increasing average cell size did not seem to improve the Type I error control for the rest of the HOV tests because these tests had a poor control of Type I error regardless of average cell sizes.

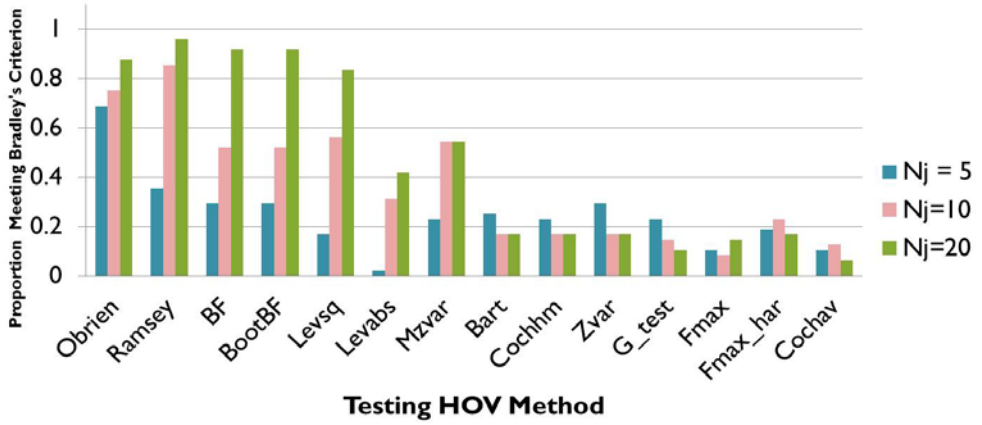

Figure 5: Proportion meeting Bradley's liberal criterion of HOV tests by average cell size (alpha=.05).

3.4 Statistical Power Estimates

Statistical power was estimated only for conditions in which Type I error was adequately controlled. Thus, the Levene with squared deviation, BF, O'Brien, Ramsey, and Bootstrap BF tests were included in the power analysis, as they had overall adequate Type I error control across conditions. As Figure 6 shows, the power differences among these five tests were very subtle; while the O'Brien test tended to have slightly less power and the Bootstrap BF test had slightly greater power than others.

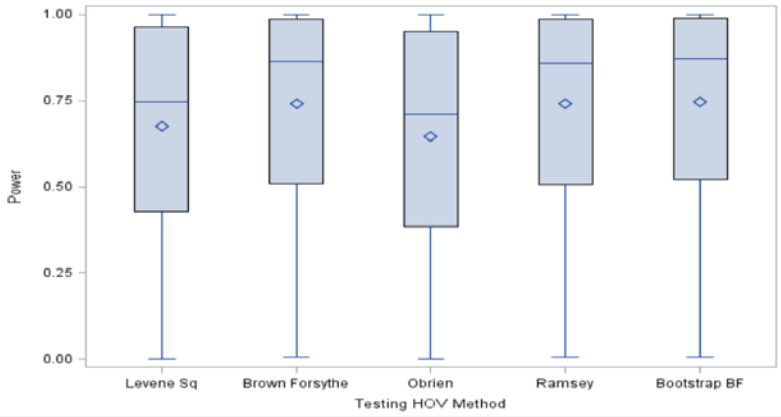

Figure 6: Distributions of estimated statistical power of HOV tests with overall adequate Type I error control.

3.5 Type I Error rates and Statistical Power under the Normal Distribution

The proportions of conditions meeting the Bradley's criterion under the normal

distribution for 12 out of 14 HOV tests (see Figures 3 and 4 for these 12 tests), ranged from .63 to 1, except for the Cochran's C test with the arithmetic mean (Bradley proportion = .33) and the F-max test with the arithmetic mean (Bradley proportion = .27). It seemed that under the normal distribution almost all the HOV tests examined in this study showed adequate Type I error control (i.e., the Bradley's proportions were greater than .50). As shown in Figure 3, the Bartlett, Cochran's C with the harmonic mean, and Z-variance tests had the perfect Type I error control (Bradley's proportions = 1.00).

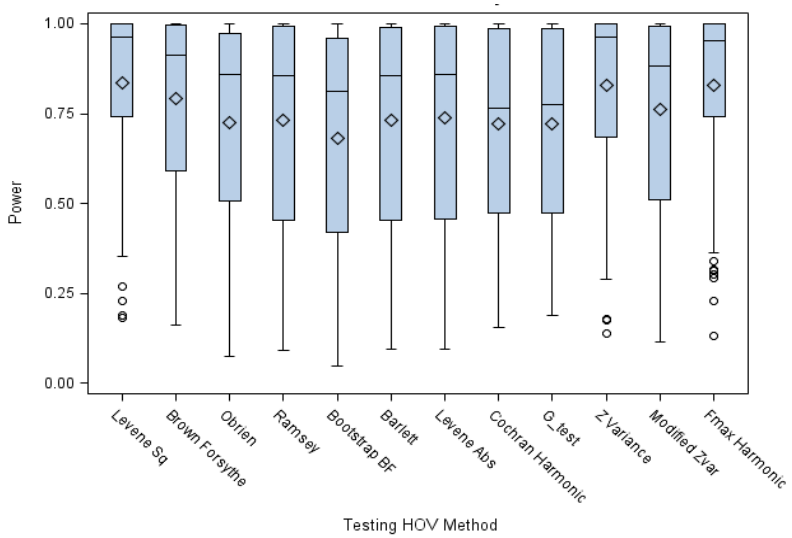

Figure 7: Distributions of estimated statistical power of HOV tests under normal distributions at the nominal level of .05.

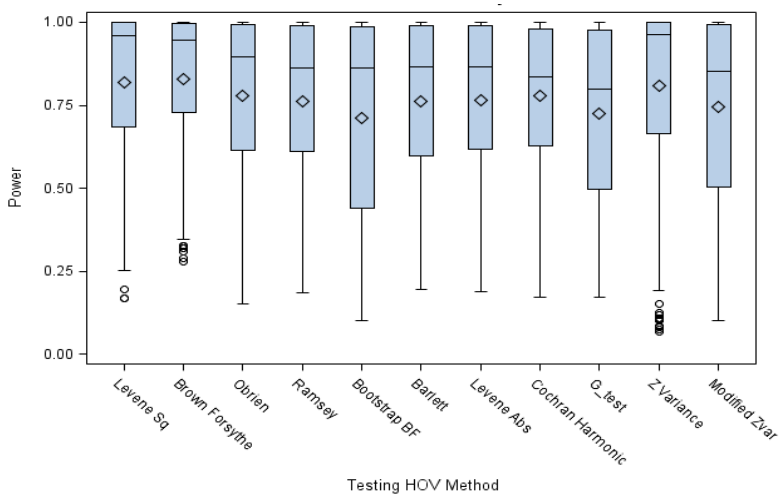

Figure 8: Distributions of estimated statistical power of HOV tests under normal distributions at the nominal level of .25.

Statistical power was estimated for these 12 HOV tests under normal distributions. As Figure 7 shows, they all had acceptable power estimates, but the Levene with squared deviations, Z-variance, and F-max with harmonic mean tests tended to have slightly greater power than other tests across all nominal levels. When the nominal level was .10, .15, .20, or .25, a similar pattern occurred and the power differences among them were

very subtle, as is shown in Figure 8 using the nominal level of .25 as an example. It is worth noting that the BF test tended to have increased power to become one of the best performers at the nominal level of .25.

4. Conclusions and Recommendations

Overall, five HOV tests maintain adequate Type I error control better than the others across all the conditions, including the Ramsey conditional test, the O'Brien test, the Brown-Forsythe (BF) test, the Bootstrap Brown-Forsythe test, and the Levene test with squared deviations. The power for each of these five tests is acceptable and the power differences are subtle. The O'Brien test tends to have slightly less power than the others and the Bootstrap BF tends to have slightly greater power. Across all the nominal levels investigated, the results in terms of Type I error rates and statistical power are consistent.

Among six design factors, population shape and average cell size are two factors that significantly affect Type I error rate control. As for the impact of population shape, the Ramsey and O'Brien tests are the only two tests that maintained adequate Type I error control across all the population shapes. Among the five best tests that maintain adequate Type I error control, the BF and Bootstrap BF tests have conservative Type I error rates if the distribution shape is extremely leptokurtic (i.e., kurtosis = 25). In contrast, the Levene test with squared deviations tends to have liberal Type I error rates when the shape is skewed (skewness = 1.5 or above). Fortunately, there are 12 out of 14 tests that maintain adequate Type I error control if the population shape is normal, except for the F-max test and the Cochran test with the arithmetic mean. Interestingly, the Bartlett test, the Cochran test with the harmonic mean, and the Z-variance test maintain Type I error control extremely well when the population shape is normal. As for power for these tests that adequately control Type I error when the population shape is normal, the Levene test with squared deviations, the F-max test with the harmonic mean, and the Z-variance have larger power. Based on the results of the Type I error control and power, the Z-variance test seems to be the best choice when the distribution is normal.

Average cell size has significant impacts on Type I error control for the five HOV tests, including the Ramsey conditional test, the BF test, the Bootstrap BF test, the Levene test with squared deviations, and modified Z-variance test. With average cell size of 5, these five tests do not maintain adequate Type I error control. Increasing average cell size improves Type I error control. The O'Brien test maintains adequate Type I error control regardless of average cell sizes; that is, the O'Brien test shows adequate control of the Type I error although the average sample size is as small as 5. Average cell size has no impact on the rest of the HOV testing methods because these methods show poor control of the Type I error even though the sample size is large.

There are two types of uses for these HOV tests that are examined in this study: (a) substantive hypotheses about population variances (e.g., do educational enrichment programs increase heterogeneity of student achievement scores?) and (b) testing for the tenability of homogeneity of variance in consideration of a subsequent test of mean differences (Bryk & Raudenbush, 1988). Different nominal alpha levels may be indicated for these uses. This study indicates that as the nominal alpha level increases, the proportion of conditions meeting Bradley's criterion increases for all of the tests.

Based on the findings from this study, there is a caveat for researchers and practitioners: Choosing a HOV test with care, especially under the data conditions of small samples and/or non-normal distributions. Since it is difficult to assess population distribution shape based on samples (Ghasemi & Zahediasl, 2012), we make some recommendations for selecting appropriate HOV testing methods based on average sample size: (a) if average sample size is less than 10 ($n_j < 10$), the O'Brien test is the best choice because of Type I error control; (b) if average sample size is between 10 and 20 ($10 \leq n_j < 20$), the Ramsey conditional test is the best choice because of Type I error control and greater power; and (c) if average sample size is greater than 20 ($n_j \geq 20$), the Bootstrap BF test is the best because of greater power.

References

Algina, Olejnik, & Ocanto (1989). Type I error rates and power estimates for selected two-sample tests of scale. *Journal of Educational and Behavioral Statistics, 14*, 373-384.

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Statistical Society, Series A, 160*, 268–282.

Boos, D. D. & Brownie, C. (1989). Bootstrap methods for testing homogeneity of variances. *Technometrics, 3*, 69-82.

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

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

Bryk, A. S. & Raudenbush, S. W. (1988). Heterogeneity of variance in experimental studies: A challenge to conventional interpretations. *Psychological Bulletin, 104*, 396-404.

Cochran, W. G. (1941). The distribution of the largest of a set of estimated variances as a fraction of their total. *Annals of Human Genetics, 11*, 47-52.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Academic Press

Conover, W. J., Johnson, M. E., & Johnson, M. M. (1981). A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. *Technornetrics, 23*, 351-361.

Fisher, R. A., & Yates, F. (1963). *Statistical tables for biological agricultural and medical research.* New York: Hafner.

Ghasemi, A. & Zahediasl, S. (2012). Normality tests for statistical analysis: A guide for non-statisticians. *International Journal of Endocrinology and Metabolism, 10*, 486-489.

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

Gupta, A. K., Harrar, S. W., & Pardo, L. (2007). On testing homogeneity of variances for nonnormal models using entropy. *Statistical Methods and Applications, 16*, 245-261.

Hartley, H. O. (1950). The maximum f-ratio as a short-cut test for heterogeneity of variance. *Biometrika, 37*, 308-312.

Howell, D. C. (2002). *Statistical methods for psychology* (5th edition). Pacific Grove, CA: Duxbury/Thomson Learning.

Huck, S. W. & McLean R. A. (1975). Using a repeated measures ANOVA to analyze the data from a pretest-posttest design: A potentially confusing task. *Psychological Bulletin, 82*, 511-518.

Katz, G. S., Restori, A. F., & Lee, H. B. (2009). A Monte Carlo study comparing the Levene test to other homogeneity of variance tests. *North American Journal of Psychology, 11,* 511-522.

Keselman, H. J., Wilcox, R. R., & Lix, L. M. (2003). A generally robust approach to hypothesis testing in independent and correlated groups designs. *Psychophysiology, 40*, 586-596.

R.U.E. 't Lam. (2010). Scrutiny of variance results for outliers: Cochran's test optimized. *Analytica Chimica Acta, 659,* 68–84.

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*, 359-366.

Lemeshko, B. Y., Lemeshko, S. G., & Gorbunova, A. A. (2010). Application and power of criteria for testing the homogeneity of variances: Part I. *Parametric criteria. Measurement Techniques, 53,* 237-247.

Levene, H. (1960). Robust tests for the equality of variance. In Contributions to Probability and Statistics (I. Olkin, ed.) 278–292. Stanford Univ. Press, Palo Alto, CA.

Lim, T. S. & Loh, W. Y. (1996). A comparison of tests of equality of variances. *Computational Statistics and Data Analysis, 22*, 287-301.

Montgomery, D. C. (2001). *Design and analysis of experiments* (5th ed.). New York: Wiley.

O'Brien, R. G. (1979). A general ANOVA method for robust test of additive models for variance. *Journal of the American Statistical Association, 74*, 877-880.

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

Olejnik, S. F. & Algina, J. (1987). Type I error rates and power estimates of selected parametric and nonparametric tests of scale. *Journal of Educational Statistics, 12*, 45-61.

Ott, L. R. & Longnecker, M. T. (2010). *An introduction to statistical methods and data analysis.* Brooks/Cole, Cengage Learning, Belmont, CA.

Overall, J. E. & Woodward, J. A. (1974). A simple test for heterogeneity of variance in complex factorial designs. *Psychometrika, 39*, 311-318.

Overall, J. E. & Woodward, J. A. (1976). *A robust and powerful test for heterogeneity of variance.* University of Texas Medical Branch Psychometric Laboratory.

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

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

Rodriguez, P., Nguyen, D. T., Kim, E. S., Kellermann, A. P., Bellara, A., Chen, Y.-H., & Kromrey, J. D. (2013). Testing two population means: Another look at alternative approaches. *Proceedings of the American Educational Research Association.*

Snedecor, G. W. & Cochran, W. G. (1989). *Statistical methods* (8th Edition). Iowa State University Press.