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

Educational and Psychological
Measurement
2017, Vol. 77(2) 305–329
© The Author(s) 2016
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0013164416645162
journals.sagepub.com/home/epm

Yan Wang¹, Patricia Rodríguez de Gil¹,
Yi-Hsin Chen¹, Jeffrey D. Kromrey¹,
Eun Sook Kim¹, Thanh Pham¹, Diep Nguyen¹,
and Jeanine L. Romano²

Abstract

Various tests to check the homogeneity of variance assumption have been proposed in the literature, yet there is no consensus as to their robustness when the assumption of normality does not hold. This simulation study evaluated the performance of 14 tests for the homogeneity of variance assumption in one-way ANOVA models in terms of Type I error control and statistical power. Seven factors were manipulated: 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. Overall, the Ramsey conditional, O'Brien, Brown–Forsythe, Bootstrap Brown–Forsythe, and Levene with squared deviations tests maintained adequate Type I error control, performing better than the others across all the conditions. The power for each of these five tests was acceptable and the power differences were subtle. Guidelines for selecting a valid test for assessing the tenability of this critical assumption are provided based on average cell size.

Keywords

homogeneity of variance, analysis of variance, nonnormality, Type I error control, statistical power

¹University of South Florida, Tampa, FL, USA

²American Board of Psychology, Chapel Hill, NC, USA

Corresponding Author:

Yan Wang, Educational Measurement and Research, University of South Florida, 4202 E. Fowler Avenue, EDU105, Tampa, FL 33620, USA.

Email: yanw@mail.usf.edu

Although advances in computer hardware and statistical computing have led to the development of complex analytical procedures, the analysis of variance (ANOVA) remains the most frequently used statistical analysis in applied research (Jan & Shieh, 2014). Whenever researchers are interested in comparing more than two treatments or levels of a factor, one-factor ANOVA is an appropriate, reasonable statistical method. It has received considerable attention in the literature, across applications in social, biological, and behavioral sciences. One-factor ANOVA has been examined in the context of career and technical education (Rojewski, Lee, & Gemici, 2012), sample size (Jan & Shieh, 2014), four-group experiment (Katz, Restori, & Lee, 2009), non-normal one-way ANOVA (Tonda & Wakaki, 2003), and the asymptotic expansion of the null distribution of one-way ANOVA test (Yanagihara, 2000). One-way ANOVA has been implemented in recent dissertations, for example, in the context of the collaborative technologies in remote lab delivery systems (Ashby, 2009), and cross-training and role assignments in collaborative learning groups (Beck-Jones, 2004).

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

$$H_0 : \sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2,$$

where k denotes the number of groups compared in a study. Moderate deviations from the assumption of equal variances may not seriously affect the results in ANOVA (Glass, Peckham, & Sanders, 1972), but researchers may be concerned about large deviations from the HOV assumption. Thus, testing HOV is always a crucial step in ANOVA analyses. However, the conventional test (e.g., Bartlett test) used to evaluate HOV is sensitive to departures from normality, in which case researchers should turn to alternative tests. Many alternative tests of HOV have been proposed in the literature. In this study, we collected 14 HOV tests and evaluated their performance in one-way ANOVA models under several experimental conditions across normal and non-normal populations. Some of these methods are prevailing and available in statistical software packages (e.g., Statistical Analysis System or SAS). Many are alternatives of the HOV testing approaches that have been proposed in the 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. A brief description of each method is also presented below.

Fourteen HOV Methods

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 test is sensitive to nonnormal distributions

Table 1. Alternative Homogeneity of Variance Tests Statistics.

HOV test	Test statistic	Notation
Barlett	$\chi^2 = \frac{(N-k) \log \left[\frac{\sum_{j=1}^k \frac{(\eta_j-1) s_j^2}{(N-k)}}{\frac{(\sum_{j=1}^k \frac{1}{\eta_j-1}) - \frac{1}{3(k-1)}}{(N-k)}} \right]}{1 + \frac{(\sum_{j=1}^k \frac{1}{\eta_j-1}) - \frac{1}{3(k-1)}}{3(k-1)}}$	<p>N = Total sample size η_j = Group j sample size k = Number of groups S_j^2 = Group j variance</p>
Levene (absolute and squared deviations)	$Z_{ij} = Y_{ij} - \bar{Y}_j \text{ and } Z_{ij} = (Y_{ij} - \bar{Y}_j)^2$ $W = \frac{(N-k) \sum_{j=1}^k \eta_j (\bar{Z}_j - \bar{Z}_{..})^2}{(k-1) \sum_{j=1}^k \sum_{i=1}^{\eta_j} (Z_{ij} - \bar{Z}_j)^2}$	<p>Y_{ij} = Raw score $\bar{Y}_j$ = Mean of the jth group $\bar{Z}_j$ = Group mean of Z_{ij} $\bar{Z}_{..}$ = Grand mean</p>
Brown–Forsythe ^a	$z_{ij} = Y_{ij} - \bar{Y}_{ji} $ $W = \frac{(N-k) \sum_{j=1}^k \eta_j (\bar{Z}_j - \bar{Z}_{..})^2}{(k-1) \sum_{j=1}^k \sum_{i=1}^{\eta_j} (Z_{ij} - \bar{Z}_j)^2}$	<p>$\bar{Y}_j$ = Median of group j Z_{ij} = Transformed value of Y_{ij} $\bar{Z}_j$ = Group mean of Z_{ij} $\bar{Z}_{..}$ = Grand mean</p>
O'Brien	$r_{ij}(w) = \frac{(w + \eta_j - 2) \eta_j (y_{ij} - \bar{y}_j)^2 - w s_j^2 (\eta_j - 1)}{(\eta_j - 1)(\eta_j - 2)}$	<p>s_j^2 = within-group unbiased estimate of variance for sample j w ($0 \leq w \leq 1$) = weighting factor $m_j = \sum (Y_{ij} - \bar{y}_j) / \eta_j$</p>
Ramsey	$b_2 = m_4 / m_2^2$ $b_{2j} = \frac{\sum (\hat{y}_i - \bar{y}_j)^4}{\frac{\eta_j}{\left[\frac{\sum (\hat{y}_i - \bar{y}_j)^2}{\eta_j} \right]^2}}$	
Cochran's C^b	$C = \frac{s_{\max}^2}{\sum s_j^2}$ $\text{Critical } C = \frac{1}{1 + \frac{F_{\frac{\alpha}{k}, n-1, (k-1)(n-1)}}{k-1}}$	<p>n = Number of observations in each group (for the balanced design) F = Critical value of F at α/k with $df = n - 1, (k - 1)(n - 1)$</p>

(continued)

Table 1. (continued)

HOV test	Test statistic	Notation
G test	$G = \frac{v_{\max} s_{\max}^2}{\sum_{j=1}^k y_j s_j^2}$ $\text{Critical } G = \frac{1}{1 + \frac{v_{\text{pool}} / (v_{\max} - 1)}{F_{\frac{\alpha}{k}, \bar{n}_j - 1, (k-1)(\bar{n}_j - 1)}}$	$v_{\text{pool}} =$ Pooled degrees of freedom $v_{\max} =$ Degrees of freedom for the group within the largest variance $\bar{n}_j =$ Mean number of observations in each group $F_{\frac{\alpha}{k}} =$ Critical value of F at α/k with $df = \bar{n}_j - 1, (k - 1)(\bar{n}_j - 1)$ $s_{\max} =$ Largest group variance $s_{\min} =$ Smallest group variance $S^2 =$ Sample variance estimate $\sigma^2 =$ True population variance $c = 2 + 1/\eta_j$ $MS_w =$ Pooled within-cells mean square across all groups (or cells in a more complex factorial design) $K =$ Mean of the kurtosis indices from all groups
$F_{\max}^b$	$F_{\max} = \frac{s_{\max}^2}{s_{\min}^2}$	
Z-variance	$z = \sqrt{2\chi^2 - \sqrt{2(df)} - 1}$ $\chi_{n-1}^2 = \frac{(n-1)s^2}{\sigma^2}$ $Z_j = \sqrt{\frac{c(\eta_j - 1)s_j^2}{MS_w}} - \sqrt{c(\eta_j - 1) - 1}$ $F = \frac{\sum_{j=1}^k Z_j^2}{k-1}$	
Modified Z-variance	$c = 2.0 \left(\frac{2.9 + 2(\eta_j)}{K} \right)^{1.6(\eta_j - 1.8K + 14.7)/\eta_j}$ $K = \frac{\sum_{j=1}^k Z_j^4}{\eta_j - 2}$ $Z_{ij} = \frac{Y_{ij} - \bar{Y}_j}{\sqrt{\frac{\eta_j - 1}{\eta_j} S_j^2}}$	

^aThe bootstrap version of the BF test was also evaluated. ^bWith arithmetic mean and harmonic mean for the group size under unbalanced design.

and instead recommended alternative testing approaches: Levene (absolute and squared), Brown–Forsythe (BF), and O’Brien tests.

Levene Test With Absolute or Squared Deviations

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

Brown–Forsythe Test

Brown and Forsythe (1974) proposed the BF test that follows the idea of the Levene 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 the Levene test when the population distribution is skewed.

Bootstrap Brown–Forsythe Test

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

Cochran’s C Test With Arithmetic or Harmonic Means

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

$F_{\max}$ Test With Arithmetic or Harmonic Means

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

G Test

The *G* test is a ratio of the product of the largest group variance and its degrees of freedom to the sum of the products of each group variance and its degrees of freedom (‘t Lam, 2010). If the obtained value exceeds the critical value, the null hypothesis of variance homogeneity is rejected.

O’Brien Test

O’Brien (1979) proposed a test that transforms original scores so they represent sample variances. The mean of the transformed values per group,

$$\bar{r}_j = \frac{\sum r_{ij}}{n_j} = s_j^2,$$

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 jack-knife pseudo value of $s_j^2(w = 1)$. O’Brien (1981) suggested setting $w = .5$ as default.

Ramsey Conditional Test: BF or O’Brien

Ramsey’s (1994) conditional procedure is based on the use of either BF or O’Brien 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 with 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.

Z-Variance Test

Overall and Woodward (1974) proposed the *Z*-variance test based on Fisher and Yates’s formula (1963). It transforms the chi-square statistics with large degrees of freedom into standard normal deviate *z*-scores. It performed very well with normally distributed data but produced inflated 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 test HOV by implementing a *c* value based on sample size, skewness, and kurtosis.

Previous Simulation Studies on the Performance of HOV Tests

Among the 14 HOV tests included in this study, some have been examined by simulation studies in terms of their Type I error control and statistical power under various experimental conditions. For instance, simulation studies (e.g., Bhandary & Dai, 2009; Lemeshko & Mirkin, 2004; Parra-Frutos, 2012; Sharma & Kibria, 2013) have consistently reported that the Bartlett test controlled Type I error rates adequately when the population distribution was normal, but it was very sensitive to departures of normality. Type I error rates were inflated substantially due to nonnormality. This pattern was also identified for the Cochran's C test. The Bartlett test was especially recommended with normal distributions, as its performance was independent of sample sizes. Type I error control of the Levene test with absolute deviations depended on the sample size (Bhandary & Dai, 2009). It required a larger sample size (at least 30 per group) to ensure a good control of Type I error rates, while with a smaller number of observations per group (e.g., 5, 10, and 20), Type I error rate was inflated. Compared with the Bartlett test, the Levene test was less sensitive to departures of normality (Gorbunova & Lemeshko, 2012). But in this case, both the Bartlett and Levene tests were inferior to the O'Brien test that performed very well in terms of Type I error control across all population distributions, except for small sample size coupled with skewed distributions (Lee, Katz, & Restori, 2010).

BF was reported to provide robust Type I error control across a wide range of shapes and sample sizes (Ramsey & Ramsey, 2007; Sharma & Kibria, 2013). Different findings were presented by Parra-Frutos (2012) that BF performed well in large sample sizes but was too conservative with small sample sizes (i.e., group size 5) across distributions. Ramsey's conditional test was also found to be slightly conservative across various sample sizes and population distributions (Ramsey & Ramsey, 2007).

$F_{\max}$ test was reported to have adequate Type I error control with normal distributions when the group size was relatively small (i.e., 10), while with a larger group size (i.e., 30), Type I error rates were slightly lower than the nominal alpha level. It was extremely conservative with platykurtic distributions, whereas with leptokurtic and skewed distributions Type I error rates were inflated substantially (Lee et al., 2010). The performance of the Z-variance test was very similar to that of the $F_{\max}$ test in terms of robustness. The modified Z-variance test had very adequate Type I error control across all distribution shapes, indicating that it was superior to the original Z-variance test. In contrast, Bhandary and Dai (2009) noticed that group size did not have an impact on the Type I error control of $F_{\max}$ under normal distributions and balanced designs. They also identified a slight inflation of Type I error rates as the differences in group sizes increased.

In terms of statistical power, Bartlett, Z-variance, Cochran's C , $F_{\max}$, and O'Brien achieved high statistical power with the presence of normal or platykurtic distributions, with Cochran's C having the highest power (Gorbunova & Lemeshko, 2012). Modified Z-variance had lower power than the Z-variance test (Lee et al., 2010).

Although with normal or platykurtic distributions Levene had a lower power than those tests, it had higher power when distributions were leptokurtic and skewed. Power of the BF test was low for small sample sizes and decreased when coupled with unbalanced samples; on the other hand, statistical power increased with larger samples, both balanced and unbalanced (Parra-Frutos, 2012). An improvement in its power can be obtained by combining BF with the O'Brien procedure to form Ramsey's conditional test that showed consistent and significant advantage in power than BF (Ramsey & Ramsey, 2007). When the normality assumption was met or with the platykurtic distributions, Cochran's *C* test had the highest power. Cochran's *C* was the best method when the variance of one group was much larger than the other groups. But with different variance patterns, the power of the Cochran's *C* test decreased significantly (Gorbunova & Lemeshko, 2012).

In summary, although simulation studies have examined the performance of a few widely used HOV testing procedures in terms of Type I error control and statistical power, a few questions remain unaddressed. First, simulation studies have reported inconsistent results regarding the performance of HOV procedures. Thus, a comprehensive simulation study that incorporates various experimental conditions is essential to more nuanced understanding of those tests' performance. Based on this detailed understanding, implications can be drawn to inform applied researchers in the selection of appropriate HOV tests. Second, the extant literature in this field is focused on the comparisons of commonly used HOV procedures that are available in statistical packages, such as Bartlett, O'Brien, BF, and Levene. The performance of the modified versions of those methods (e.g., bootstrap BF) and other methods (e.g., Ramsey's conditional test, *G* test) remains unexplored. Therefore, this study evaluated the performance of the 14 HOV tests through Monte Carlo simulations that take various design factors into consideration.

Research Purpose

Considering the ongoing controversy on testing HOV and the minimal consensus among studies as to which test is appropriate for a particular analysis, the purpose of this study was to compare approaches with testing variance homogeneity based on sample data. The goal was to conduct a comprehensive examination of the 14 approaches for testing the HOV assumption in one-way ANOVA models under various 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 impact on Type I error control and statistical power among the HOV tests?

Method

Manipulated Factors

This study was conducted using a simulation approach. The six design factors manipulated in this study included the following: (a) number of groups ($k = 4$ and 6),

Table 2. Simulation Study Sample Size Patterns.

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

Note. K = Number of groups; Progressive N = Progressive increase of sample size; Split N = Half of groups has the same sample size; One extreme = One group has a larger sample size than the other groups.

(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 shape ($\gamma_1 = 0.00$ and $\gamma_2 = 0.00$, $\gamma_1 = 1.00$ and $\gamma_2 = 3.00$, $\gamma_1 = 1.50$ and $\gamma_2 = 5.00$, $\gamma_1 = 2.00$ and $\gamma_2 = 6.00$, $\gamma_1 = 0.00$ and $\gamma_2 = 25.00$, and $\gamma_1 = 0.00$ and $\gamma_2 = -1.00$, where γ_1 and γ_2 represent skewness and kurtosis, respectively). Nonnormal populations were generated by implementing the Fleishman's transformation (Fleishman, 1978). Table 2 shows the four sample size patterns and Table 3 presents the seven variance patterns investigated. In addition to six design factors, we used six alpha levels for testing the HOV assumption ($\alpha = .01, .05, .10, .15, .20,$ and $.25$). Thus, this study yielded a total of 16,416 simulation conditions for the HOV assumption tests in one-factor ANOVA models with fixed effects.

Data Generation

For each condition, 5,000 replications 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 $\pm .003$ (Robey & Barcikowski, 1992). 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. Of note is that a SAS macro for

Table 3. Simulation Study Variance Patterns.

Max variance ratio	Population variances								
	Progressive			Split			One extreme		
	1:4	1:8	1:16	1:4	1:8	1:16	1:4	1:8	1:16
K = 6									
1	1	1	1	1	1	1	1	1	1
2	1.6	2.4	4	1	1	1	1	1	1
3	2.2	3.8	7	1	1	1	1	1	1
4	2.8	5.2	10	4	8	16	1	1	1
5	3.4	6.6	13	4	8	16	1	1	1
6	4	8	16	4	8	16	4	8	16
K = 4									
1	1	1	1	1	1	1	1	1	1
2	2	3.3	6	1	1	1	1	1	1
3	3	5.7	11	4	8	16	1	1	1
4	4	8	16	4	8	16	4	8	16
Max variance ratio	Progressive inversely			Split inversely			One extreme inversely		
	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. Progressive = Each K population variance increases progressively across groups; Progressive inversely = K variance patterns increase across but inversely by group order.

testing the HOV assumption in one-factor ANOVA using the 14 approaches is provided elsewhere (Nguyen et al., 2014) and also available on request.

Statistical Analysis

We examined Type I error rates 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

Figure 1. Distributions of estimated Type I error rates across all simulation conditions at .05.

Note. The horizontal line is the reference line at the nominal level of .05. Levene_abs = Levene test with absolute deviations; Levene_sq = Levene test with squared deviations; BF = Brown–Forsythe; O'Brien = O'Brien; Ramsey = Ramsey conditional test; Cochran_ave = Cochran's *C* test with arithmetic mean; Cochran_har = Cochran's *C* test with harmonic mean; Modified_Z_var = Modified *Z*-variance; Fmax = $F_{\max}$ test with arithmetic mean; Fmax_har = $F_{\max}$ test with harmonic mean.

investigated the robustness using Bradley's liberal criterion (Bradley, 1978). The liberal criterion for the robustness is set at $\pm .5\alpha$ around the nominal alpha (i.e., $\alpha \pm .5\alpha$). For instance, using the alpha level of .05, a test is considered robust when the Type I error rate falls between .025 (i.e., $.05 - .5 \times .05$) and .075 (i.e., $.05 + .5 \times .05$). Finally, factorial ANOVA analyses with the generalized eta-square effect size were conducted to explore the impact of design factors on variability in the estimated Type I error rates and statistical power. Cohen's (1988) moderate effect size of .0588 was set as a cutoff value to indicate the significant level.

Results

Type I Error Rates

Boxplots were first checked to examine the distributions of Type I error rate estimates for the 14 HOV tests across all simulation conditions at each nominal alpha level. Figure 1 presents a set of boxplots based on Type I error rates across all conditions at

Figure 2. Overall proportion conditions of the 14 homogeneity of variance tests meeting Bradley's liberal criterion.

Note. O'Brien = O'Brien; Ramsey = Ramsey conditional test; BootBF = Bootstrap Brown–Forsythe; BF = Brown–Forsythe; Levsq = Levene test with squared deviations; Levabs = Levene test with absolute deviations; Mzvar = Modified Z-variance; Zvar = Z-variance; Bart = Bartlett; Cochhm = Cochran's C test with harmonic mean; Cochav = Cochran's C test with arithmetic mean; Fmax = $F_{\max}$ test with arithmetic mean; Fmax_har = $F_{\max}$ test with harmonic mean.

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 performing approaches that controlled Type I error adequately across all simulation conditions.

A Closer Look at Type I Error Control Using 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 $\pm .5\alpha$ around the nominal alpha level. That is, for each nominal alpha level investigated, the proportion meeting Bradley's liberal criterion is the proportion of simulation conditions with Type I error rates that fell within the range of $\alpha \pm .5\alpha$. Figure 2 shows the overall proportions of simulation conditions meeting Bradley's criterion for the 14 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 simulation conditions meeting Bradley's criterion across all nominal alpha levels compared with other tests. The proportions of simulation conditions that met Bradley's criterion for these five tests 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 14 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 Bradley's criterion.

Impact of Design Factors on the Type I Error Control

Additionally, we conducted factorial ANOVA analyses with the generalized eta-square (η_G^2) effect size to explore the impact of design factors on Type I error rates, using Cohen's (1988) moderate effect size of .0588 as a cutoff value to indicate the significant level. Among all design factors, population shape and average cell size had the greatest impact on Type I error rate across all the HOV tests, $\eta_G^2 > .0588$.

Figure 3 displays the proportions of simulation conditions that met Bradley's liberal criterion for each test by all population distribution shapes. The O'Brien and Ramsey conditional tests tended to have adequate Type I error control across all population distribution shapes (i.e., all the proportions of conditions meeting Bradley's criterion $> .60$). 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 Type I error control when population distributions were extremely leptokurtic (i.e., Shape 5: skewness = 0, kurtosis = 25; the proportions of conditions meeting Bradley's criterion $< .40$). The Levene test with squared deviations had a poor control of Type I error under skewed distributions (i.e., Shape 3: skewness = 1.5, kurtosis = 5; Bradley's proportion $< .40$; Shape 4: skewness = 2, kurtosis = 6; Bradley's proportion $< .20$). When data were normally distributed, however, the Bartlett, Cochran with the harmonic mean, and Z-variance tests controlled Type I error extremely well such that the proportions of simulation conditions meeting Bradley's criterion were all 100%.

Figure 4 summarizes the performance of the 14 HOV tests regarding 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 boxplots that showed the distributions of Type I error estimates of these 14 tests under different population shapes. The shaded areas indicated adequate Type I error control; that is, the proportions of simulation conditions meeting Bradley's liberal criterion were equal to or greater than 50%. If the proportions of conditions meeting Bradley's liberal criterion

Figure 3. Proportions meeting Bradley's liberal criterion for the 14 homogeneity of variance tests by population shape ($\alpha = .05$).
Note. Six population shapes are shown on the right. Two numbers within the parentheses represent skewness and kurtosis, respectively. O'Brien = O'Brien; Ramsey = Ramsey conditional test; BF = Brown–Forsythe; BootBF = Bootstrap Brown–Forsythe; Mzvar = Modified Z-variance; levsq = Levene test with squared deviations; Levabs = Levene test with absolute deviations; Fmax_har = $F_{\max}$ test with harmonic mean; Fmax = $F_{\max}$ test with arithmetic mean; Cochav = Cochran's C test with arithmetic mean; Bart = Bartlett; Cochhm = Cochran's C test with harmonic mean; Zvar = Z-variance.

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 boxplots. 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 extremely 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., Shape 3: skewness = 1.5, kurtosis = 5; Shape 4: skewness = 2, kurtosis = 6). 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 (i.e., Shape 2: skewness = 1, kurtosis = 3; Shape 3: skewness = 1.5, kurtosis = 5; Shape 4: skewness = 2, kurtosis = 6), the HOV tests that did not have the adequate Type I error control were liberal, whereas under the slightly platykurtic distributions (i.e., Shape 6: skewness = 0, kurtosis = -1), they tended to have conservative Type I errors.

In addition to population distribution shape, average cell size was the other factor that had significant impact on Type I error rates across the 14 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 (i.e., all Bradley's

	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 homogeneity of variance tests under different population distribution shapes.

Note. L = Liberal; C = Conservative. Shaded areas indicate adequate Type I error control (i.e., the proportions meeting Bradley’s liberal criterion were equal to or greater than .50).

proportions > .60). For the Ramsey conditional test, BF, Bootstrap BF, and Levene with squared deviations, the proportions of conditions meeting Bradley’s criterion increased as average cell size increased. With average cell size of 20, these five tests had the proportions of conditions meeting Bradley’s criterion greater than .80. However, increasing average cell size did not improve the performance of the rest of the HOV tests because these tests had poor control of Type I error across all three average cell sizes (i.e., 5, 10, and 20).

Statistical Power and Impact of Design Factors

Statistical power was obtained only for tests with adequate Type I error control under most simulation conditions. Thus, only five HOV tests (i.e., the Levene with squared

Figure 5. Proportion of simulation conditions meeting Bradley's liberal criterion of homogeneity of variance tests by average cell size ($\alpha = .05$).
Note. N_j = average number of observations per group.

deviations, BF, O'Brien, Ramsey, and Bootstrap BF tests) were included in the power analysis. Figure 6 shows the distributions of statistical power estimates of these five tests under conditions with adequate Type I error control. As seen in Figure 6, the power differences among these five tests were very subtle; the O'Brien test tended to have slightly less power and the Ramsey, Bootstrap BF, and BF tests had slightly greater power than the others.

Based on factorial ANOVA analyses with the generalized eta-square (η_G^2) effect size on statistical power, four design factors, including maximum group variance, variance pattern, population distribution, and average cell size, had significant impacts on the power estimates of these five HOV tests, $\eta_G^2 > .0588$. Table 4 presents the power estimates by maximum group variance and variance pattern for each test. Power estimates of all five tests increased substantially as the maximum group variance increased from 4 to 8 or 16, compared with the minimum group variance as 1. Consistent with the findings shown in Figure 6, the O'Brien test had slightly lower power across various maximum group variances, whereas the Ramsey conditional, Bootstrap BF, and BF tests had slightly higher power, and these three tests reached statistical power around .52, .78, and .91 for the maximum group variances of 4, 8, and 16, respectively. Furthermore, power estimates of all five methods were also affected by variance pattern. As seen in Table 4, greater power was evidenced under variance patterns of one extreme, one extreme inversely, split, and split inversely (ranging from .76 to .87), except for Levene with squared deviation and O'Brien with the split variance pattern (.64 and .60, respectively). In other words, when one group or half of the groups had extreme maximum variances, it was more likely for the five HOV tests to detect heterogeneity of variances. With progressive and progressive inversely variance patterns, however, the power was relatively low, ranging from .40 to .58.

Figure 6. Distributions of estimated statistical power of five homogeneity of variance tests under conditions with adequate Type I error control.

Table 5 presents the power estimates of five HOV tests by population shape and average cell size. When population distributions were nonskewed (i.e., normal: skewness = 0, kurtosis = 0; extremely leptokurtic: skewness = 0, kurtosis = 25; and slightly platykurtic: skewness = 0, kurtosis = -1), greater power was observed across all five tests (ranging from .70 to .89) compared with that under skewed distributions (ranging from .48 to .69). Interestingly, power estimates were greatest under the slightly platykurtic distribution for five HOV tests that all reached .85. As population distributions were skewed, power decreased marginally. Under skewed and extremely leptokurtic distributions, the Ramsey conditional, BF, and Bootstrap BF tests tended to have slightly greater power than O'Brien and Levene with squared deviation. With the normal and slightly platykurtic distributions, power of all five tests seemed equivalent. Finally, average cell size also showed significant impact on power estimates. Power increased substantially from around .53 to .76 as average cell size increased from 10 to 20. It should be noted that average cell size of 5 was not included in the power analysis due to poor Type I error control.

Type I Error Rates and Statistical Power Under the Normal Distribution

The proportions of conditions meeting 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

Table 4. Statistical Power Estimates by Maximum Group Variance and Variance Pattern.

	Maximum group variance			Variance pattern					
	4	8	16	One extreme	Split	Progressive	One extreme inversely	Split inversely	Progressive inversely
HOV test									
Levene sq	.52	.71	.81	.78	.64	.45	.82	.85	.55
Brown-Forsythe	.52	.78	.92	.86	.81	.57	.77	.87	.56
O'Brien	.49	.68	.79	.76	.60	.40	.80	.83	.53
Ramsey	.53	.78	.91	.86	.81	.57	.79	.87	.57
Bootstrap BF	.53	.79	.92	.87	.82	.58	.78	.87	.57

Note. Levene sq = Levene test with squared deviations; Bootstrap BF = Bootstrap Brown–Forsythe; Progressive = Each K population variance increases progressively across groups; Progressive inversely = K variance patterns increase across but inversely by group order.

Table 5. Statistical Power Estimates by Population Shape and Average Cell Size.

HOV test	Population shape (skewness, kurtosis)						Average cell size	
	1 (0, 0)	2 (1, 3)	3 (1.5, 5)	4 (2, 6)	5 (0, 25)	6 (0, −1)	10	20
Levene sq	.74	.57	.54	.52	.77	.89	.53	.73
Brown–Forsythe	.76	.68	.68	.66	.85	.85	.55	.79
O'Brien	.70	.54	.50	.48	.76	.87	.46	.70
Ramsey	.76	.67	.68	.66	.84	.86	.54	.80
Bootstrap BF	.76	.68	.69	.68	.85	.85	.56	.80

Note. Levene sq = Levene test with squared deviations; Bootstrap BF = Bootstrap Brown–Forsythe.

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

.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., Bradley's proportions 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 (i.e., Bradley's proportions of 1).

Statistical power was estimated for these 12 HOV tests across all conditions under normal distributions at various nominal levels. Figure 7 shows the distributions of estimated statistical power of the 12 HOV tests under normal distributions at the nominal level of .05. As shown in Figure 7, 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 and become one of the best performing procedures at the nominal level of .25.

Summary 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

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

BF test, the Bootstrap BF test, and the Levene test with squared deviations. The robustness of the Ramsey conditional, O’Brien, and BF tests confirmed findings of previous studies (Lee et al., 2010; Ramsey & Ramsey, 2007; Sharma & Kibria, 2013). The differences among the power distributions for these five robust HOV tests are subtle. The O’Brien test tends to have slightly less power, which is consistent with the finding of the Lee et al. (2010) study, and the Ramsey conditional, Bootstrap BF, and BF tests tend to have slightly greater power. Bartlett, Cochran, and Z-variance can be recommended due to its 100% Type I error control and reasonable power under normality. Particularly, the power of Z-variance is impressive except some outliers. Across all the nominal levels investigated, the results in terms of Type I error rates and statistical power are consistent.

Among the six design factors, population shape and average cell size are the two factors that significantly affect Type I error control across all HOV procedures. As for the impact of population shape, the Ramsey conditional 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's C test with the arithmetic mean. The Bartlett, Cochran's C with the harmonic mean, and Z -variance tests maintain Type I error control extremely well when the population shape is normal, consistent with findings of prior simulation studies (Bhandary & Dai, 2009; Lee et al., 2010; Lemeshko & Mirkin, 2004; Parra-Frutos, 2012; Sharma & Kibria, 2013). As for power of 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 greater power. The advantage of Z -variance in power with normal distributions was also identified by Gorbunova and Lemeshko (2012). Based on the results of the Type I error control and power, the Z -variance test seems to be the best choice when the population distribution is normal.

Average cell size has significant impact 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. The prior finding regarding the negative impact of small group size coupled with skewed distributions on the Type I error control of the O'Brien test was not observed in this study (Lee et al., 2010). This inconsistency might occur because distinctive evaluation criteria regarding the robustness of HOV tests were employed across studies. Lee et al. (2010) considered .061 (group size 10, skewness .58) and .060 (group size 10, skewness .63) as inflated Type I error rates of the O'Brien test, while based on Bradley's liberal criterion (Bradley, 1978) applied in this study, Type I error control was adequate as they fell within the range of .025 to .075. 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 despite relatively large group sizes.

Four design factors (i.e., maximum group variance, variance pattern, population shape, and average cell size) show significant impacts on statistical power for the five HOV tests mentioned above. Because the minimum group variance is set to 1, increasing maximum group variance represents the increase of the differences between group variances. As expected, it is more likely for the five HOV tests to detect heterogeneity of variances when maximum group variance increases. When maximum group variance is 8, the power of the five HOV tests is acceptable (.68 or above). As maximum group variance is increased to 16, the Ramsey conditional, Bootstrap BF, and BF tests have power over .90. Variance pattern also shows significant impact on power. Power is greater when the variance(s) of one group (one extreme or one extreme inversely) or a half of groups (split or split inversely) differ(s) substantially from variances of other groups. As for population distribution, skewed distributions tend to have lower power of the HOV tests compared with nonskewed distributions (i.e., normal, extremely leptokurtic, and slightly platykurtic distributions

Table 6. Recommendations for Selecting HOV Tests Based on Average Cell Size.

Average cell size	Recommended HOV test	Rationale
$n_j < 10$	O'Brien	Adequate Type I error control (Bradley's proportion $> .60$)
$10 \leq n_j < 20$	Ramsey conditional test	Adequate Type I error control (Bradley's proportion $> .80$)
$n_j \geq 20$	Ramsey conditional test/ Bootstrap Brown–Forsythe/ Brown–Forsythe	Adequate Type I error control (Bradley's proportion around .90) and greater statistical power (power around .80)

Note. n_j = average cell size.

in this study). However, increasing the skewness of population distributions from 1, 1.5 to 2 seems to have a slight or negligible decrease of power. With nonskewed distributions, kurtosis seems to have positive impact on power of the five HOV tests. Compared with normal distributions, the five HOV tests have higher power under extremely leptokurtic and slightly platykurtic distributions. Especially with slightly platykurtic distributions, power estimates of all five methods reach the highest. As for average cell size, increasing larger average cell size leads to the increase of power. Power for the five best HOV tests with average cell size of 20 is .70 or higher. It is anticipated that average cell size greater than 20 will obtain power higher than what we have in this study for these five HOV tests.

There are two types of uses for these HOV tests: (a) testing substantive hypotheses about population variances (e.g., do educational enrichment programs increase heterogeneity of student achievement scores?) and (b) testing for the tenability of HOV 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 the tests.

Based on the findings from this study, there is a caveat for researchers and practitioners: Choose an HOV test with care, especially under the data conditions of small samples and/or nonnormal 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 (Table 6): (a) if average cell size is less than 10 ($n_j < 10$), the O'Brien test is the best choice because of adequate Type I error control; (b) if average cell size is between 10 and 20 ($10 \leq n_j < 20$), the Ramsey conditional test is the best choice because of adequate Type I error control; and (c) if average cell size is greater than or equal to 20 ($n_j \geq 20$), the Ramsey conditional test, the Bootstrap BF test, or the BF test are the best because of adequate Type I error control and greater power.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

- Ashby, J. E. (2009). *Using collaborative technologies in remote lab delivery systems for topics in automation*. Retrieved from ProQuest Dissertations and Theses Global database. (Order No. 3402678)
- Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London, Series A*, 160, 268-282. Retrieved from <http://www.jstor.org/stable/96803>
- Beck-Jones, J. (2004). *The effect of cross-training and role assignment in cooperative learning groups on task performance, knowledge of accounting concepts, teamwork behavior, and acquisition of interpositional knowledge*. Retrieved from ProQuest Dissertations and Theses Global database. (Order No. 3098363)
- Bhandary, M., & Dai, H. (2009). An alternative test for the equality of variances for several populations when the underlying distributions are normal. *Communications in Statistics: Simulation and Computation*, 38, 109-117. doi:10.1080/03610910802431011
- Boos, D. D., & Brownie, C. (2004). Comparing variances and other measures of dispersion. *Statistical Science*, 19, 571-578. doi:10.1214/088342304000000503
- Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, 31, 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, 364-367. doi:10.2307/2285659
- Bryk, A. S., & Raudenbush, S. W. (1988). Heterogeneity of variance in experimental studies: A challenge to conventional interpretations. *Psychological Bulletin*, 104, 396-404. doi:10.1037/0033-2909.104.3.396
- Cochran, W. G. (1941). The distribution of the largest of a set of estimated variances as a fraction of their total. *Annals of Eugenics*, 11, 47-52. doi:10.1111/j.1469-1809.1941.tb02271.x
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York, NY: Academic Press.
- Fisher, R. A., & Yates, F. (1963). *Statistical tables for biological, agricultural and medical research*. New York, NY: Hafner.
- Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika*, 43, 521-532. doi:10.1007/BF02293811
- Ghasemi, A., & Zahediasl, S. (2012). Normality tests for statistical analysis: A guide for non-statisticians. *International Journal of Endocrinology and Metabolism*, 10, 486-489. doi:10.5812/ijem.3505

- Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research*, 42, 237-288. doi:10.3102/00346543042003237
- Gorbunova, A. A., & Lemeshko, B. Y. (2012, June). *Application of parametric homogeneity of variances tests under violation of classical assumption*. Paper presented at the proceedings of the 2nd Stochastic Modeling Techniques and Data Analysis International Conference, Chania, Crete, Greece.
- Hartley, H. O. (1950). The maximum F -ratio as a short-cut test for heterogeneity of variance. *Biometrika*, 37, 308-312. doi:10.2307/2332383
- Jan, S.-L., & Shieh, G. (2014). Determining sample size for precise control analysis with heterogeneous variances. *Journal of Educational and Behavioral Statistics*, 39, 91-116. doi:10.3102/1076998614523069
- 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.
- 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. doi:10.3844/jmssp.2010.359.366
- Lemeshko, B. Y., & Mirkin, E. P. (2004). Barlett and Cochran tests in measurements with probability laws different from normal. *Measurement Techniques*, 47, 960-968. doi:10.1007/s11018-005-0027-y
- Levene, H. (1960). *Robust tests for the equality of variance*. In I. Olkin (Ed.), *Contributions to probability and statistics* (pp. 278-292). Palo Alto, CA: Stanford University Press.
- Lim, T. S., & Loh, W. Y. (1996). A comparison of tests of equality of variances. *Computational Statistics and Data Analysis*, 22, 287-301. doi:10.1016/0167-9473(95)00054-2
- Nguyen, D. T., Pham, T. V., Rodriguez de Gil, P., Hicks, T., Wang, Y., Li, I., . . . Kromrey, J. D. (2014). ANOVA_HOV: A SAS® macro for testing homogeneity of variance in one-factor ANOVA models. In *SouthEast SAS Users Group 2014 Proceedings*. Cary, NC: SAS Institute. Retrieved from <http://analytics.ncsu.edu/sesug/2014/SD-05.pdf>
- O'Brien, R. G. (1979). A general ANOVA method for robust tests of additive models for variances. *Journal of the American Statistical Association*, 74, 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, 570-574. doi:10.1037/0033-2909.89.3.570
- Ott, R. L., & Longnecker, M. T. (2010). *An introduction to statistical methods and data analysis*. Belmont, CA: Cengage Learning.
- Overall, J. E., & Woodward, J. A. (1974). A simple test for heterogeneity of variance in complex factorial designs. *Psychometrika*, 39, 311-318. doi:10.1007/BF02291705
- Overall, J. E., & Woodward, J. A. (1976). *A robust and powerful test for heterogeneity of variance*. Galveston: University of Texas Medical Branch Psychometric Laboratory.
- Parra-Frutos, I. (2012). Testing homogeneity of variances with unequal sample sizes. *Computational Statistics*, 28, 1269-1297. doi:10.1007/s00180-012-0353-x
- Ramsey, P. H. (1994). Testing variances in psychological and educational research. *Journal of Educational Statistics*, 19, 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*, 83, 1387-1388. doi:10.1080/00949655.2012.678081
- Ramsey, P. H., & Ramsey, P. P. (2007). Testing variability in the two-sample case. *Communications in Statistics: Simulation and Computation*, 36, 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, 283-288. doi:10.1111/j.2044-8317.1992.tb00993.x
- Rojewski, J. W., Lee, I. H., & Gemicci, S. (2012). Use of *t*-test and ANOVA in career-technical education research. *Career and Technical Education Research*, 37, 263-275. doi: 10.5328/cter37.3.263
- 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, 1944-1963. doi:10.1080/00949655.2012.675336
- Snedecor, G. W., & Cochran, W. G. (1989). *Statistical methods* (8th ed.). Ames: Iowa State University Press.
- 't Lam, R. U. E. (2010). Scrutiny of variance results for outliers: Cochran's test optimized. *Analytica Chimica Acta*, 659, 68-84. doi:10.1016/j.aca.2009.11.032
- Tonda, T., & Wakaki, H. (2003). Asymptotic expansion of the null distribution of the likelihood ratio statistic for testing the equality of variance in a nonnormal one-way ANOVA model. *Hiroshima Mathematical Journal*, 33, 113-126. Retrieved from http://projecteuclid.org/download/pdf_1/euclid.hmj/1150997870
- Yanagihara, H. (2000). Asymptotic expansion of the null distribution of one-way ANOVA test statistic for heteroscedastic case under nonnormality. *Communication in Statistics: Theory and Methods*, 29, 463-476. doi:10.1080/03610920008832495