Comparison of Tests for Heteroscedasticity in Between-Subjects ANOVA Models

Mokshad P. Gaonkar **T. Mark Beasley**
University of Alabama at Birmingham

Several tests for heteroscedasticity in a two-group between-subject variances were compared with a simulation study. Two common rank-based procedures inflated test size with skewed error distributions. Nonparametric Levene test performed well but has notable limitations. Tests based on the absolute value of OLS residuals also inflated test size with skewed error distributions. Procedures based on squared OLS residuals performed better; however, the original Breusch-Pagan and Variance Function Regression are sensitive to even slight departures from the normality assumption. The Brown-Forsythe test based on taking the absolute value of median centered data performed the best; however, generalization to more complex analyses would not be straightforward.

Despite decades of research, there does not seem to be a consensus for a single procedure for testing for heteroscedasticity that works uniformly well across common data scenarios. In between-subjects ANOVA, testing for heteroscedasticity reduces to testing whether the *J* groups have identical variances with the following null hypothesis:

$$\mathrm{H}_0\text{: } \sigma_1^2 = \sigma_2^2 = . . . = \sigma_j^2 = . . . = \sigma_J^2 \ . \qquad (1)$$

Tests for heteroscedasticity (i.e., differences in variances across groups) may have two different goals: 1). Testing the homoscedasticity assumption and 2). Analyzing variability as an outcome of interest.

Testing the Homoscedasticity Assumption

Valid procedures for inference in linear models that estimate the mean of a response have been long established and are ubiquitous in countless fields of research. During the 1920's, the term ANalysis Of VAriance (ANOVA) was coined to describe a method for comparing mean responses among two or more groups of independent, normally distributed observations with a common variance (Muller, 2009). However, it cannot be safely assumed that groups of subjects are homogeneous or exchangeable. Hence, there is no basis to assume equality of variances when testing the null hypothesis of identical means among multiple groups, even in randomized experiments (Nordstokke et al., 2011). Furthermore, if this assumption is ignored, the results of statistical tests that use a pooled estimate of the variance (e.g., pooled *t*-test) can be greatly distorted, thus potentially leading to incorrect inferences. Of note is that nonparametric tests are also susceptible to issues with unequal variances when testing for shifts in location parameters (Zimmerman & Zumbo, 1993a; 1993b). Thus, switching to a nonparametric statistical approach to avoid the homoscedasticity assumption does not alleviate the problem of unequal variances. In fact, rank transformations have been shown to inherit the heteroscedasticity from the original untransformed variables (Zimmerman, 1996).

Some researchers advocate testing the homoscedasticity assumption to justify the use of tests that assume variance homogeneity in their primary analysis. In this case, the researcher would hope to find that the variances are homogeneous. This approach of conditional testing (i.e., preliminary testing of the assumptions in order to choose the appropriate analysis) is debatable. First, a non-significant result from a test of heteroscedasticity does not guarantee that the population variances are truly constant; this could be a Type 2 error. Also using this approach, the choice of the "appropriate" test is conditional upon the statistical properties (e.g., robustness; power) of the preliminary test (Zimmerman, 2004). Furthermore, it does not ensure that this non-significant heteroscedasticity will not affect inferences about means in procedures that assume homoscedasticity. It is also important to note that it is not necessary to use a preliminary test of variance homoscedasticity to justify the use of heteroscedastic procedures (e.g., Welch's (1951) heteroscedastic ANOVA) because these tests are generally effective regardless of whether variances are equal or unequal across groups. Although many researchers have suggested abandoning non-robust parametric procedures completely in favor of robust procedures that do not require the homogeneity of variances assumption (e.g., Wilcox, Charlin, & Thompson, 1986; Zimmerman, 2004), researchers in many disciplines still widely use traditional homoscedastic parametric procedures and feel the need to screen for the assumptions associated with these tests. For this reason alone, valid tests for heteroscedasticity are needed.

Heteroscedasticity as an Outcome

In a growing number of research disciplines (e.g., education, sociology, experimental biology), investigators are becoming increasingly interested in the properties of their data aside from central tendency (Parra-Frutos, 2009). Therefore, how the variability of an outcome is affected by an experimental treatment or other categorical factor may be of research interest. Thus, a more interesting reason for assessing differences among of variances is that the primary research question is concerned with whether the dispersion of the dependent variable is different across multiple groups. For example, Kowalski, et al. (2020) examined whether professional development programs affecting the variation in educational outcomes for science teachers. Western & Bloome (2009) examined factors affect the variance in residual inequalities among racial groups. Mattison et al. (2017) reported the effects of caloric restriction on the variability in health outcomes in rhesus monkeys.

Bryk and Raudenbush (1988) argue that the presence of heterogeneity of variance across groups can have important implications for the research conclusions. Specifically, the presence of heterogeneity of variances in an experimental study may indicate the presence of an interaction between subject characteristics and treatment group membership. In other words, heterogeneity of variances can indicate that individuals vary in their response to the treatment (assuming the treatment group was a fixed effect). This could be an important consideration for researchers, and valid tests for evaluating heterogeneity of variances would be important to evaluate within an experimental design.

Given these reasons for testing variance homogeneity (1), valid tests for assessing heteroscedasticity are relevant to many research questions of interest and crucial if an investigator feels the need to justify the use of pooled variance tests of mean differences.

Parametric Tests for Heteroscedasticity for Between-Subjects ANOVA Designs.

Several tests for differences in variances for basic between-subjects ANOVA designs have been developed. It has been consistently shown that Bartlett's (1937) test is sensitive to departures from normality (e.g., Snedecor & Cochran, 1989; Conover, Johnson, & Iman, 1981; Parra-Frutos, 2012). Furthermore, tests based on the ratio of variances (Hartley, 1950) perform poorly in several simulation studies because it requires independent random samples of the same size from normally distributed populations (Ott & Longnecker, 2010).

Procedures based on analyzing a transformation of the original response, *y*, have fared much better. The Levene (1960) approach of performing standard ANOVA with a pooled error term on transformed residuals to test the null hypothesis (1) is actually a family of techniques (Nordstokke & Zumbo, 2007). In between-subjects ANOVA models with *J* groups, the ordinary least squares (OLS) residuals can be calculated as:

$$e_{ij} = (y_{ij} - \bar{y}_j), \tag{2}$$

where $i = 1$ to n_j and $j = 1$ to J are subscripts for the i^{th} subject nested in the j^{th} group, respectively, n_j is the sample size for the j^{th} group, *J* is total number of groups, y_{ij} is the response for i^{th} subject nested in the j^{th} group, and $\bar{y}_j$ is the mean for the j^{th} group.

One approach suggested by Levene ($L1$) is to perform standard ANOVA on the absolute values the residuals from 2 ($|e|$). This may be the most commonly used procedure because it the default test in many software including: `SPSS EXAMINE, T-TEST, ONEWAY,` and `UNIANOVA`; `STATA ROBVAR`; `R LeveneTest` and is available in SAS PROC GLM. Another approach suggested by Levene ($L2$) is to perform standard ANOVA on the squared residuals from 2 (e^2). Miller (1968) showed that ANOVA on absolute values will be asymptotically incorrect if the population is not symmetric and that the problem can be corrected by using medians instead of means to center the variables. Therefore, Brown & Forsythe (1974) suggested a modification to Levene's approach by performing standard ANOVA on the absolute value of differences from the group medians (*BF*):

$$d_{ij} = |y_{ij} - Md_j|, \tag{3}$$

where Md_j is the median for the j^{th} group.

The O'Brien (*OB*) modification to the Levene family of procedures involves performing standard ANOVA on transformed squared residuals (e^2):

$$u_{ij} = \frac{[(W+n_j-2)n_j(y_{ij}-\bar{y}_j)^2]-[W(n_j-1)s_j^2]}{(n_j-1)(n_j-2)}\,. \tag{4}$$

where $s_j^2$is the sample variance for the j^{th} group. O'Brien (1979, 1981) noted that the choice of the value for *W* is rarely critical but suggested *W*=0.5 because the group means of the transformed data are the group variances. This is used as the default in `SAS PROC GLM HOVTEST=OBRIEN`. It should be noted that if *W*=0 is used then equation (4) reduces to

$$u_{ij(W=0)} = \frac{n_j(y_{ij}-\bar{y}_j)^2}{(n_j-1)} \;, \tag{5}$$

which in an ANOVA model with no additional covariates can be shown to be equal to:

$$u_{ij(W=0)} = \frac{(y_i-\hat{y}_i)^2}{(1-h_{ii})} \,, \tag{6}$$

where h_{ii} is the i^{th} diagonal element of the hat matrix ($\mathbf{H} = \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'}$). Keyes and Levy (1997) proposed a modification by taking the square root of the transformed values (6) produced by the O'Brien procedure with *W*=0 ($OB_{W=0}$). It should be noted that with equal sample sizes (n_j) the transformation in equations 5 and 6 are the same for each group:

$$(1\text{-}h_{ii}) = (n_j\text{-}1)/n_j \;,$$

and thus, a linear transformation of *e* or e^2.Therefore, in a between-subjects ANOVA with equal sample sizes and no additional covariates, the O'Brien procedure with *W*=0 is equivalent to the *L*2 test, and the Keyes-Levy (*KL*) procedure is equivalent to the *L*1 test. To our knowledge, the O'Brien family of procedures can only be found in `SAS PROC GLM`.

Brown and Forsythe's (1974) suggestion of using trimmed means to center the data and taking the absolute value. This is a popular modification of these approaches and is available in `SPSS EXAMINE, T-TEST, ONEWAY,` and `UNIANOVA`; `STATA ROBVAR`; and `R LeveneTest`. Keselman et al. (2008) suggested using asymmetric trimming of the means in situations where the residuals are skewed; however, this procedure is not available in any known software.

Another modification often suggested is to use the separate variance approach to ANOVA and apply adjustments for the denominator degrees-of-freedom (e.g., Welch, 1951) to these test statistics (e.g., Beasley, 1995; Keselman et al., 1979). Ramsey (1994) suggested a conditional procedure based on the use of either *BF* or *OB* method, conditional on a test of kurtosis; however, Zimmerman (2004) has warned that using a conditional test to select "the appropriate" subsequent test is problematic. Wang et al. (2017) reported that Lim and Loh's (1997) approach of applying bootstrapping to the *BF* to obtain p-values maintained adequate test size. However, Keselman et al. (2008) suggested that bootstrapping is not necessary because satisfactory Type 1 error rates for *BF* can be obtained without bootstrapping.

Non-Parametric (Rank-Based) Tests for Heteroscedasticity for Between-Subjects Designs.

The Conover (1971) squared ranks test, originally proposed by Taha (1964), is a non-parametric version of the parametric Levene's tests for equality of variance. Procedurally, the absolute deviations from the sample means (i.e., *L1*) are ranked then squared:

$$C_{ij} \;= (\text{RANK}(|(y_{ij} - \bar{y}_j)|)^2. \tag{7}$$

Although Conover and Iman (1987) developed exact tables for using the squared ranks procedure, it is more common to perform a test similar to the Wilcoxon Rank Sum test on these squared ranks:

$$CN = \frac{\sum_{j=1}^{J} n_j(\bar{C}_j-\bar{C}_*)^2}{\sum_{i=1}^{N}(C_{ij}-\bar{C}_*)^2/(N-1)} \,, \tag{8}$$

where C_{ij} is the transformed value for i^{th} subject nested in the j^{th} group, and $\bar{C}_j$ is the mean for the j^{th} group, $\bar{C}_*$ is the overall of mean of the transformed scores, and *N* is the total sample size, N=Σn_j. This test statistic approximates a chi-square distribution with *J*–1 *df*s. To our knowledge, the Conover squared ranks can only be found in SAS PROC NPAR1WAY.

Fligner and Killeen (1976) proposed several procedures as non-parametric tests for homogeneity of group variances based on ranks. The most widely used version of the Fligner Killen (*FK*) approach employs the *BF* approach of calculating absolute values of median centered samples (2), ranking these values,

$$K_{ij} \;= (\text{RANK}(|(y_{ij} - M_j)|). \tag{9}$$

and then weighting these ranks:

$$A_{\text{ij}} = \Phi^{-1}[(1+(K_{ij}/(N+1)))/2] \;; \tag{10}$$

where Φ^{-1} is the inverse cumulative density function of the normal distribution. The *FK* test statistic is calculated as:

$$FK = \frac{\sum_{j=1}^{J} n_j(\bar{A}_j - \bar{A}_*)^2}{\sum_{i=1}^{N}(A_{ij} - \bar{A}_*)^2/(N-1)}\,, \tag{11}$$

where A_{ij} is the transformed value for i^{th} subject nested in the j^{th} group, and $\bar{A}_j$ is the mean for the j^{th} group, and $\bar{A}_*$ is the overall of mean of the transformed scores. This test statistic also approximates a chi-square distribution with J–1 *df*s. To our knowledge, *FK* can only be found in the `R` package `fligner.test`.

Nordstokke and Zumbo (2010) proposed non-parametric rank-based Levene test (*NPL*) that involves simply ranking the pooled data and performing the *L*1 test on the ranks. Shear, Nordstokke, and Zumbo (2018) demonstrated that sampling from populations with unequal means can lead to incorrect Type 1 error rates when error distributions are asymmetric. Given the novelty of this approach, it is not available in any commercially available statistical software.

Linear Regression Model Based Generalizations of Tests for Heteroscedasticity

Levene's approach of analyzing the squared residuals was generalized to linear regression models by Breusch and Pagan (1979). The Breusch-Pagan (*BP*) method uses the likelihood function to obtain a Lagrange multiplier score test. Specifically, the *BP* procedure takes the residuals from the original linear model:

Statistical Model 1: $\boldsymbol{y} = \mathbf{XB} + \boldsymbol{e}$

takes the residual, which in the case of ANOVA models with no covariates is the same a subtracting the group means from each value of *y* (equation 2), squares the residual, and transforms it as follows:

$$g = e^2/(\text{SSE}_{(1)}/N)$$

where $\text{SSE}_{(1)}$ is the Error Sum of Squares from Model 1 and N is the total sample size. This transformation results in the mean of the transformed data equaling one, $\bar{g} = 1$. Typically, a linear model with the same design matrix (**X**) is performed on the transformed data:

Statistical Model 2: $\boldsymbol{g} = \mathbf{XB} + \boldsymbol{u}$

The *BP* test statistic is calculated as:

$$\chi^2_{(\text{BP})} = ½\text{SSM}_{(2)}\,; \tag{12}$$

where $\text{SSM}_{(2)}$ is the Model Sum of Squares from Model 2. This test approximates a chi-square distribution with *k* *df*s, where *k* is the number of predictors in Model 2, which equals the number of groups minus one (*J*-1) for reference cell regression for between-subjects ANOVA models without additional covariates.

Since the *BP* test has been shown to be sensitive to departures from the linear model normality assumption, Koenker (1981) proposed a studentized version of this test (BP_S):

$$\chi^2_{(\text{K})} = NR^2_{(2)} \tag{13}$$

where $R^2_{(2)}$ is the R^2 from Model 2. This test also approximates a chi-square distribution with *k* *df*s. Woolridge (2012) suggested a modified *BP* tests by using the *F*-test from Model 2. We note that the *BP* transformation is a linear transformation of e^2, and thus, the R^2 and *F*-tests for Model 2 would be the same whether the analysis is performed on e^2 or *g*. Therefore, in the context of ANOVA models, the modified *BP F*-test is identical the Levene approach of performing ANOVA on the squared residuals (*L*2) and the Koenker studentized version is equivalent to the White's (1980) test for heteroscedasticity for two-group comparisons and simple regression.

Test for Heteroscedasticity using Generalized Linear Models

Recently, Western and Bloome (2009) advocated the use of Variance Function Regression (*VFR*) to examine heteroscedasticity. *VFR* uses a generalized linear model with the squared residuals (e^2) as the response with a log link function. A gamma distribution is used as the assumed distribution to account to the right-skewed nature of e^2. Although this approach sounds promising (Ng & Cribbie, 2017), it has only been evaluated in limited number of research applications.

Methods

We performed a simulation study to investigate how these methods for testing heteroscedasticity perform in a between-subjects one-way ANOVA model with *J*=2 groups. We anticipate that these results will generalize to other one-way between-subjects ANOVA models. We compared these tests under three distributional and several sample size conditions. To verify computations and results, the first author performed simulation using R and the second author used SAS/IML. Following the work of Nordstokke and Zumbo (2010) and Shear et al. (2018), we varied both Total Sample Size (*N*) and Sample Size Ratio of the two groups (n_1/n_2). The empirical Type 1 Error Rates (i.e., Test Size) were evaluated at two statistical significance levels, $\alpha = 0.05$ and 0.01. The Empirical Power was evaluated under conditions where the tests maintained Test Size. Each sample size by distribution condition simulation employed 10,000 replications.

Distributions

Independent errors (ε) from the unit normal distribution with homoscedastic variances (i.e., each group has a population variance of one; $\sigma^2{}_j = 1$) were generated to evaluate Type 1 Error Rates (i.e., Test Size) for each test under conditions that meet all model assumptions. A symmetric, heavy tailed error distribution was employed to evaluate how kurtosis affects the statistical properties of these tests when symmetry holds. We chose a *t*-distribution with 10 *df*s ($t_{(10)}$), which has a skewness of g^3=0 and excess kurtosis of g^4=1, to examine how slight non-normality affect these tests. A chi-square distribution with *df*=1 ($\chi^2{}_{(1)}$; g^3=2.83; g^4=12) was generated to evaluate how each test performs with an extreme violation of the normality assumption.

The first author used R functions to generate each distribution and standardized the variables to have a zero mean and unit variance using the expected means and standard deviations of each distribution. The second author used the SAS/IML RANNOR function and Headrick's (2004) polynomial method of generating non-normal data with zero mean and unit variance.

Between-Subjects (one-way ANOVA) Designs

Following the work of Nordstokke and Zumbo (2010) and Shear et al. (2018), we varied both Total Sample Size (*N*) and Sample Size Ratio of the two groups (n_1/n_2). Both Balanced (equal sample sizes with Sample Size Ratio of 1:1) and Unbalanced Designs with Sample Size Ratios of 2:1 and 3:1 were investigated. Although we have used several total sample sizes, we have reported the results for four *N*'s that we think are representative.

Tests

The parametric. rank-based (non-parametric), and model-based tests previously reviewed were compared. The parametric tests use Levene's (1960) family of methods that perform standard ANOVA with a pooled variance error term to transformations of centered data. This includes: the absolute value of residuals method (*L*1); the Keyes and Levy (1997) modification of *L*1 (*KL*); Levene's (1960) squared residuals method (*L*2); the O'Brien (1979, 1981) family of tests with *W*=0 and 0.5 (*OB*); the Brown and Forsythe (1974) method of centering the data with medians (*BF*) and centering the data with 10% symmetric trimmed means (TM_{10}) then taking absolute values. Since, several investigators (e.g., Keselman et al., 2008; Mara & Cribbie, 2017) have suggested that using an error estimated with a separated variance approach and applying Welch adjusted *df*s may improve the performance of tests for homogeneity of variances, and therefore, this modification is evaluated for *L*1, *KL*, *L*2, the *OB* tests, *BF*, and TM_{10}. The non-parametric tests included: Conover's square ranks test (*CN*); the Fligner and Killeen (1976) test (*FK*); and the non-parametric Levene test (*NPL*) proposed by Nordstokke and Zumbo (2010). We evaluated the Breusch-Pagan (*BP*) and Koenker's (1981) studentized *BP* (BP_S) generalizations of *L*2. We also evaluated *VFR* because it has only been evaluated in limited number of research applications.

Results

Given the large number of evaluations performed, a Bonferroni correction factor of 50 was used, and thus, α=0.001 (i.e., 0.05/50) was used to obtain a critical value from the normal distribution and build simultaneous confidence intervals (CIs) for the empirical Type 1 error rejection rates. Given 10,000 replications the standard error for Type 1 error rejection rates at α=0.05 is 0.00217945. Using the Bonferroni correction factor of 50, the half-width of the CIs is 0.007, and therefore, empirical Type 1 error rejection rates above 0.057 were considered "significant" inflations of Test Size and are in red in the Tables. This reflects that the procedure does not provide a valid, usable tests. By this same process, rejection rates below 0.043 were considered "significant" suppressions of Test Size and are *italicized*, which reflects the procedure provides a usable but potentially conservative test. Procedures that maintain an adequate test size are **bolded**, indicating a valid, usable procedure. Our tables highlight rejection rates at the α=0.05 significance level, results for α=0.01 were similar and appear as supplementary tables.

Table 1. Empirical Type 1 Error Rates at α=0.05 for Total Sample Size of *N*=24.

Sample Size Ratio	**1:1**			**2:1**			**3:1**		
Sample Sizes	$n_1 = 12$		$n_2 = 12$	$n_1 = 16$		$n_2 = 8$	$n_1 = 18$		$n_2 = 6$
Distribution	**Normal**	$t_{(10)}$	$\chi^2_{(1)}$	**Normal**	$t_{(10)}$	$\chi^2_{(1)}$	**Normal**	$t_{(10)}$	$\chi^2_{(1)}$
Pooled Tests									
*L*1	6.02	6.04	21.58	6.48	6.62	20.88	5.68	5.62	19.38
*KL**				6.71	6.74	21.03	6.33	6.59	20.13
*L*2	**4.84**	**4.62**	6.78	**4.76**	**4.70**	7.41	*4.08*	**4.32**	9.04
$OB_{W=0}$*				**5.31**	**5.23**	7.92	5.61	6.15	10.16
$OB_{W=0.5}$	*3.84*	*3.38*	5.40	*4.12*	*3.93*	6.27	*4.08*	*4.06*	7.02
BF	*3.90*	*3.85*	5.06	*4.20*	*4.24*	5.10	*3.77*	*3.71*	5.40
TM_{10}	**5.35**	**5.24**	9.54	5.93	5.87	10.49	**5.17**	**4.91**	10.23
Welch *df* Tests									
*L*1(W)	5.72	**5.63**	20.12	7.74	7.63	22.35	9.85	10.06	26.70
KL(W) *				7.18	7.15	21.72	8.38	8.57	25.39
*L*2(W)	*4.11*	*3.67*	**5.41**	7.16	6.47	7.19	11.83	11.16	10.61
$OB_{W=0}$(W) *				6.30	5.75	6.85	9.88	9.32	9.67
$OB_{W=0.5}$(W)	*3.11*	*2.62*	**4.32**	**5.14**	**4.74**	5.75	8.38	7.84	7.97
BF(W)	*3.69*	*3.54*	**4.46**	**5.54**	**5.38**	6.37	7.54	7.72	10.19
TM_{10}(W)	**5.13**	**4.95**	8.72	7.07	6.92	12.50	9.02	9.11	16.03
Rank-Based Tests									
NPL	**4.57**	**4.57**	**4.77**	**5.14**	**5.14**	**5.30**	**4.73**	**4.73**	**4.82**
CN	5.76	5.89	38.79	6.18	6.46	36.69	6.09	6.50	35.12
FK	*3.79*	*3.89*	12.56	*4.22*	**4.33**	12.37	*3.78*	*3.82*	12.58
Model Based Tests									
BP	**4.82**	7.73	36.55	*3.94*	6.26	30.72	*2.94*	4.67	16.68
BP_S	**5.24**	**4.88**	7.03	5.08	4.99	7.69	**4.30**	**4.55**	9.28
VFR	8.17	11.04	41.68	9.04	11.77	41.84	9.66	11.92	41.88

Note: * When Sample Sizes are equal, *KL* is equivalent to *L*1 and $OB_{W=0}$ is equivalent to *L*2.

Type 1 Error

Tables 1 through 4 report the Empirical Type 1 Error Rates for four different total samples sizes and three different sample size ratios. With Normally distributed errors, both parametric and non-parametric tests showed occasional slight Test Size inflations with a small sample size of *N*=24 in both balanced (sample size ratio 1:1) and unbalanced designs. These Type 1 Error Rates improved with larger sample sizes. The *BP* and studentized *BP* (BP_S) performed well under normal error distributions. By contrast, *VFR* inflated Test Size with smaller sample sizes and only maintained Type 1 Error Rates with the larger sample size ($N \geq 192$).

The symmetric but heavy-tailed errors sampled from the $t_{(10)}$ distribution had little effect on the Type 1 Error Rate of the parametric and rank-based tests. For the *BP* and *VFR*, however, Test Size was inflated with rejection rates around 10% for α=0.05 regardless of sample size. The studentized version of *BP* (*BPs*) maintained appropriate Type 1 Error Rates.

With skewed errors sampled from the $\chi^2_{(1)}$ distribution, *LV*1, TM_{10}, and rank-based *CN* and *FK* tests drastically inflated test size. Interestingly, the newly proposed *NPL* did not demonstrate inflated rejection rates. Again, *BP* and *VFR* inflated Test Size with rejection rates around 40% for α=0.05 regardless of sample size., while the studentized version of *BP* (*BPs*) maintained appropriate Type 1 Error Rates. It should be noted that applying Welch adjusted *df*s to the tests, did not substantially improve Test Size inflation, and in fact, with skewed residuals made the inflations worse. Also, with small sample size of *N*=24, only *BF* and *NPL* maintained reasonable Type 1 Error Rates in every condition simulated.

Power

Table 5 reports the Empirical Type 1 Error Rates and Power for all tests with Normally distributed errors and a relatively large total sample size of *N*=240. As can be seen, the Test Size was maintained for

Table 2. Empirical Type 1 Error Rates at α=0.05 for Total Sample Size of *N*=36.

Sample Size Ratio	**1:1**			**2:1**			**3:1**		
Sample Sizes	$\boldsymbol{n_1 = 18}$		$\boldsymbol{n_2 = 18}$	$\boldsymbol{n_1 = 24}$		$\boldsymbol{n_2 = 12}$	$\boldsymbol{n_1 = 27}$		$\boldsymbol{n_2 = 9}$
Distribution	**Normal**	$t_{(10)}$	$\chi^2_{(1)}$	**Normal**	$t_{(10)}$	$\chi^2_{(1)}$	**Normal**	$t_{(10)}$	$\chi^2_{(1)}$
Pooled Tests									
*L*1	**5.59**	**5.65**	20.65	**5.48**	**5.47**	19.73	**5.55**	**5.49**	18.31
*KL**				**5.57**	**5.59**	20.15	**5.77**	**5.94**	18.83
*L*2	**4.96**	**4.60**	6.00	**4.63**	**4.30**	5.89	*4.20*	*4.17*	6.01
$OB_{W=0}$*				**4.89**	**4.66**	6.06	**5.05**	**5.12**	7.03
$OB_{W=0.5}$	*4.26*	*3.95*	5.20	*4.27*	*3.91*	5.24	*4.26*	*4.25*	5.69
BF	*4.01*	*3.98*	5.07	*4.13*	*4.06*	4.63	*3.68*	*3.52*	3.80
TM_{10}	**5.19**	**5.28**	10.94	**5.16**	**5.04**	9.12	**5.36**	**5.25**	9.61
Welch *df* Tests									
*L*1(W)	**5.40**	**5.44**	19.82	6.71	6.66	21.32	8.12	8.51	24.18
KL(W) *				6.17	6.19	20.94	7.17	7.39	23.09
*L*2(W)	**4.53**	**4.13**	**5.30**	6.97	6.59	7.15	10.67	10.30	10.05
$OB_{W=0}$(W) *				6.17	5.83	6.95	8.98	8.81	9.31
$OB_{W=0.5}$(W)	*3.89*	*3.56*	**4.50**	**5.56**	**5.14**	6.05	8.16	7.91	8.47
BF(W)	*3.90*	*3.81*	**4.71**	**4.94**	**5.10**	6.68	5.81	6.23	8.88
TM_{10}(W)	**5.02**	**5.06**	10.44	6.16	6.07	11.02	7.78	7.94	15.36
Rank-Based Tests									
NPL	**4.38**	**4.38**	**4.52**	**4.87**	**4.87**	**4.92**	**5.18**	**5.18**	**5.02**
CN	**5.63**	5.71	43.95	**5.45**	5.71	42.11	**5.63**	5.73	39.57
FK	*3.88*	*4.04*	13.39	*3.98*	*4.00*	13.84	*3.62*	*3.69*	9.96
Model Based Tests									
BP	**4.99**	8.75	37.82	*4.14*	7.34	35.07	*3.48*	5.85	29.14
BP_S	**5.24**	**4.81**	6.25	**4.84**	**4.52**	6.09	**4.37**	**4.32**	6.33
VFR	7.21	10.34	40.53	7.19	10.11	40.32	7.88	10.71	39.91

Note: * When Sample Sizes are equal, *KL* is equivalent to *L*1 and $OB_{W=0}$ is equivalent to *L*2.

all tests under these "ideal" conditions, which makes the Power of the tests comparable. With a balanced design (sample ratio 1:1), *VFR* demonstrated a slight power advantage over other tests, such as *BP*, BP_S, *L*2, and the *OB* tests. *L*1 and BF had similar, but lower Power. Welch adjustments did not enhance Power for these tests with equal sample sizes. The non-parametric tests had even less Power with *NPL* showing the lowest rejection rate. With unbalanced designs, *VFR* no longer demonstrated an advantage in power. In fact, *L*2, the *OB* tests, *BP*, and *BPs* had similar power rates and slightly more power than *VFR* and notably more power than *L*1, *KL*, *BF*, and TM_{10}. Importantly, Welch adjustment of the *df*s did enhance Power for these tests depending on the pairing of the variances and sample sizes . With negative variance-sample size pairing (i.e., group with smaller n_j has the larger variance), Welch *df* adjustment decreased power for these tests. By contrast, Welch *df* adjustment increased power with positive variance-sample size pairing (i.e., group with larger n_j has the larger variance).

Tables 6 and 7 report the Empirical Type 1 Error Rates and non-null rejection rates for all tests with skewed residuals sampled from a $\chi^2_{(1)}$ error distribution for a Total Sample Sizes of *N*=60 and 84, respectively. The *L*1, *KL*, TM_{10} *CN*, *FK*, *BP*, and *VFR* inflated test size, and therefore, the rejection rates for these procures under non-null conditions are not considered valid estimates of empirical power. Also, *L*2 and $OB_{W=0}$ showed a slight test size inflation with 3:1 sample size ratio for *N*=60.

Of the eligible tests *NPL* showed superior power; however, there are issues to consider with this approach, which will be discussed later. Among the other eligible tests, *BF* showed more power than tests based on squared residuals (*L*2. *OB*, BP_S) with equal sample sizes. Although Welch ANOVA applied to *L*2, *OB*, and *BF* maintained test size, these procedures did not enhance power. For unbalanced designs where the smaller group had the larger variance (i.e., negative pairing), tests based on squared residuals had similar power to *BF*. For positive pairing of sample size and variance (i.e., larger group has larger variance), *BF*

Table 3. Empirical Type 1 Error Rates at α=0.05 for Total Sample Size of *N*=96.

Sample Size Ratio	**1:1**			**2:1**			**3:1**		
Sample Sizes	n_1 = 48		n_2 = 48	n_1 = 64		n_2 = 32	n_1 = 72		n_2 = 24
Distribution	Normal	$t_{(10)}$	$\chi^2_{(1)}$	Normal	$t_{(10)}$	$\chi^2_{(1)}$	Normal	$t_{(10)}$	$\chi^2_{(1)}$
Pooled Tests									
*L*1	5.46	5.47	19.32	4.98	5.13	18.86	4.92	4.82	18.37
*KL**				5.09	5.20	18.94	4.90	4.99	18.63
*L*2	5.36	4.89	4.62	4.87	4.72	5.18	4.65	4.48	5.01
$OB_{W=0}$*				5.10	4.88	5.61	5.10	4.87	5.33
$OB_{W=0.5}$	5.05	4.64	4.33	4.81	4.59	5.19	4.85	4.60	5.01
BF	4.89	4.84	5.22	4.39	4.49	4.62	4.28	4.21	4.63
TM_{10}	5.40	5.31	10.14	4.91	4.87	8.72	4.84	4.68	8.74
Welch *df* Tests									
*L*1(W)	5.45	5.45	19.17	5.44	5.58	21.37	5.88	6.16	21.23
KL(W) *				5.39	5.42	21.20	5.60	5.85	20.95
*L*2(W)	5.28	4.81	4.39	5.57	5.44	9.20	7.62	8.23	9.54
$OB_{W=0}$(W) *				5.28	5.19	8.75	6.99	7.35	9.06
$OB_{W=0.5}$(W)	4.98	4.48	4.12	5.17	4.95	8.32	6.66	7.12	8.79
BF(W)	4.89	4.80	5.16	4.68	4.81	7.47	5.08	5.49	7.88
TM_{10}(W)	5.37	5.29	10.00	5.27	5.37	11.81	5.67	5.97	12.08
Rank-Based Tests									
NPL	4.98	4.98	4.70	4.95	4.95	5.14	4.64	4.64	4.80
CN	5.30	5.45	58.73	5.19	5.17	52.79	4.85	4.99	53.58
FK	4.78	4.64	21.27	4.47	4.60	17.01	4.25	4.31	18.39
Model Based Tests									
BP	4.99	9.80	42.34	4.55	9.05	39.19	4.37	8.47	39.27
BP_S	5.42	5.01	4.71	4.94	4.80	5.27	4.76	4.49	5.06
VFR	5.93	9.64	41.01	5.56	9.15	40.52	6.08	9.35	40.52

Note: * When Sample Sizes are equal, *KL* is equivalent to *L*1 and $OB_{W=0}$ is equivalent to *L*2.

has the most power. Although Welch ANOVA applied to *L*2, *OB*, and *BF* enhanced power, it comes at the cost of inflated Test Size.

Discussion

The results of this study are consistent with findings from previous similar investigations. Similar to Nordstokke and Zumbo (2007) and Conover et al. (1981), we found that one of the most commonly used tests, *L*1, and the *KL* modification of this approach suggested by Keyes & Levy (1997) drastically inflated Test Size when error distributions were skewed. Similarly, the procedure based on centering the data with 10% symmetric trimming of the means also inflated Test Size when error distributions were skewed, which is consistent with the findings of Brown and Forsythe (1974). One might suggest that symmetric trimming is not appropriate for skewed data (Keselman et al., 2008); however, it should be noted that asymmetric trimming is not available is most software, and therefore, data analysts would have to program their own algorithm to attempt such a procedure. But asymmetric trimming of the means for skewed data as suggested by Keselman et al. (2008) does warrant attention.

Although *L*2, the *OB* tests, and *BF* showed some Test Size inflation with skewed error terms with the smallest sample size (*N*=24) used, these tests performed reasonably well with larger sample sizes, regardless of the shape of the error distribution (Ramsey, 1994) and sample size ratios (Boos & Brownie, 2004). It seems counterintuitive that squared residuals from a skewed error distribution (e.g., *L*2) would fare better than the absolute value of these residuals (e.g., *L*1). Squaring an already skewed distribution drastically increases the skewness and kurtosis. The absolute value can be defined as the square root of a squared value. Thus, taking the square root of a skewed distribution would reduce skewness, and therefore, it would seem that absolute values would perform better in statistical testing. This is not the case; however. Miller (1968) showed that ANOVA on absolute values will be asymptotically incorrect if the population is not symmetric and similar analysis on squared residuals could enlighten the phenomenon.

Table 4. Empirical Type 1 Error Rates at α=0.05 for Total Sample Size of *N*=192.

Sample Size Ratio	**1:1**			**2:1**			**3:1**		
Sample Sizes	$n_1 = 98$		$n_2 = 98$	$n_1 = 128$		$n_2 = 64$	$n_1 = 144$		$n_2 = 48$
Distribution	**Normal**	$t_{(10)}$	$\chi^2_{(1)}$	**Normal**	$t_{(10)}$	$\chi^2_{(1)}$	**Normal**	$t_{(10)}$	$\chi^2_{(1)}$
Pooled Tests									
*L*1	5.22	5.20	18.22	5.44	5.40	18.16	5.02	5.08	17.98
*KL**				5.37	5.39	18.32	5.16	5.25	17.99
*L*2	5.33	5.01	*4.24*	4.94	4.64	4.56	4.63	4.48	5.07
$OB_{W=0}$*				5.02	4.72	4.67	4.95	4.75	5.32
$OB_{W=0.5}$	5.15	4.93	*4.11*	4.87	4.56	4.54	4.83	4.70	5.09
BF	4.78	4.79	4.92	5.10	4.99	5.22	4.62	4.73	4.96
TM_{10}	5.08	4.97	8.77	5.42	5.28	9.30	4.95	4.93	9.00
Welch *df* Tests									
*L*1(W)	5.22	5.20	18.14	5.63	5.62	18.36	5.37	5.49	19.41
KL(W) *				5.57	5.65	18.35	5.10	5.26	19.16
*L*2(W)	5.32	4.98	*4.14*	5.72	5.54	5.71	6.42	6.82	8.62
$OB_{W=0}$(W) *				5.66	5.43	5.58	6.07	6.50	8.34
$OB_{W=0.5}$(W)	5.15	4.86	*4.03*	5.53	5.27	5.46	5.84	6.31	8.04
BF(W)	4.78	4.79	4.92	5.10	5.23	5.70	4.87	5.12	6.46
TM_{10}(W)	5.08	4.96	8.75	5.57	5.66	9.86	5.25	5.41	10.46
Rank-Based Tests									
NPL	4.96	4.96	4.95	5.35	5.35	4.96	5.27	5.27	5.16
CN	5.15	5.22	66.24	5.29	5.35	64.75	4.89	5.00	63.14
FK	4.90	4.83	26.97	5.10	5.04	26.30	4.59	4.61	25.10
Model Based Tests									
BP	4.96	10.48	43.51	5.06	10.15	42.51	4.59	9.64	41.20
BP_S	5.39	5.04	4.29	4.98	4.68	4.60	4.69	4.50	5.12
VFR	5.47	9.66	41.25	5.72	9.65	40.62	5.34	9.52	40.08

Note: * When Sample Sizes are equal, *KL* is equivalent to *L*1 and $OB_{W=0}$ is equivalent to *L*2.

Despite suggestions from other investigators, applying Welch adjusted *df*s to these tests, did not substantially improve Test Size inflation, and in fact, with skewed residuals and smaller sizes made the inflations worse; however, this approach increased power with positive variance-sample size pairing and symmetric error distributions. This is interesting because in ANOVA models for comparing means with negative variance-sample size pairing (i.e., group with smaller n_j has the larger variance) the heteroscedastic standard error tends to be larger than the pooled standard error and the Welch procedure adjusts the denominator *df*s downward drastically, which typically leads to less power. By contrast, with positive variance-sample size pairing (i.e., group with larger n_j has the larger variance) the heteroscedastic standard error tends to be smaller than the pooled standard error and the Welch procedure adjusts the denominator *df* s downward slightly, which can lead to more statistical power.

We also demonstrated that the originally proposed *BP* test and *VFR* are highly sensitive to deviations from normality. In fact, sampling error terms from a symmetric but slightly non-normal distribution, $t_{(10)}$, led to substantial Test Size inflation and sampling error terms from a skewed distribution led to drastic Test Size inflation. However, the rejection rates did not substantially increase with sample size suggesting that these procedures are sensitive to departure from normality but are not testing a different null hypothesis. Although the studentized *BP* showed Test Size inflation with skewed error terms and the smallest sample size (*N*=24), it performed well in most other conditions. Perhaps with smaller sample sizes, some form of randomization test performed on squared residuals might be preferable.

Among the non-parametric rank-based tests, the Conover squared ranks test performed poorly with skewed error distribution and with a small sample size, even under symmetric distributional conditions. Although Conover et al. (1981) suggested *FK* as a procedure robust against departures from normality, it did inflate Test Size with skewed error distributions. With skewed error distributions, both *CN* and *FK* had consistently high rejection rates that increased with total sample size, *N*. When rejection rates increase with

Table 5. Empirical Type 1 Error and Power at α=0.05 for Total Sample Size of *N*=240 with Normal Error Distributions.

Sample Size Ratio	**1:1**			**2:1**			**3:1**		
Sample Sizes	n_1**=120**		n_2**=120**	n_1**=160**		n_2 **= 80**	n_1**=180**		n_2 **= 60**
Variances **Pooled Tests**	$\sigma^2_1=1$ $\sigma^2_2=1$	$\sigma^2_1=1$ $\sigma^2_2=1.5$	$\sigma^2_1=1$ $\sigma^2_2=1.7$	$\sigma^2_1=1$ $\sigma^2_2=1$	$\sigma^2_1=1$ $\sigma^2_2=1.7$	$\sigma^2_1=1.7$ $\sigma^2_2=1$	$\sigma^2_1=1$ $\sigma^2_2=1$	$\sigma^2_1=1$ $\sigma^2_2=1.7$	$\sigma^2_1=1.7$ $\sigma^2_2=1$
*L*1	**4.89**	**53.79**	**76.61**	**5.01**	**72.04**	**71.03**	**4.74**	**65.68**	**62.22**
*KL**				**4.91**	**72.98**	**69.79**	**4.77**	**67.43**	**60.42**
*L*2	**4.88**	**58.49**	**80.94**	**4.52**	**77.90**	**73.93**	**4.62**	**72.28**	**63.42**
$OB_{W=0}$*				**4.59**	**78.85**	**72.83**	**4.62**	**74.06**	**61.42**
$OB_{W=0.5}$	**4.81**	**58.22**	**80.62**	**4.55**	**78.56**	**72.51**	**4.55**	**73.76**	**61.16**
BF	**4.66**	**52.78**	**75.88**	**4.70**	**70.70**	**70.52**	**4.52**	**63.80**	**62.26**
TM_{10}	**4.86**	**53.63**	**76.49**	**5.02**	**71.78**	**71.00**	**4.71**	**65.26**	**62.28**
Welch *df* Tests									
*L*1(W)	**4.89**	**53.78**	**76.56**	**5.08**	**66.66**	**75.56**	**5.10**	**56.45**	**70.68**
KL(W) *				**5.10**	**67.55**	**74.72**	**5.03**	**58.23**	**68.81**
*L*2(W)	**4.85**	**58.45**	**80.92**	**5.26**	**69.18**	**82.00**	6.11	56.45	78.67
$OB_{W=0}$(W) *				**5.13**	**70.31**	**81.14**	5.71	58.36	77.22
$OB_{W=0.5}$(W)	**4.76**	**58.18**	**80.57**	**5.06**	**69.92**	**80.89**	**5.60**	**57.63**	**76.91**
BF(W)	**4.66**	**52.77**	**75.87**	**4.84**	**65.07**	**75.27**	**4.82**	**53.63**	**70.42**
TM_{10}(W)	**4.86**	**53.62**	**76.45**	**5.03**	**66.42**	**75.62**	**5.02**	**55.65**	**70.72**
Rank-Based Tests									
NPL	**4.92**	**40.12**	**60.51**	**4.75**	**53.33**	**57.76**	**4.69**	**45.70**	**51.76**
CN	**5.12**	**48.53**	**71.00**	**4.81**	**64.79**	**66.46**	**4.67**	**57.31**	**59.08**
FK	**4.74**	**51.46**	**74.27**	**4.56**	**68.77**	**69.39**	**4.43**	**61.43**	**61.37**
Model Based Tests									
BP	**4.91**	**59.23**	**81.97**	**4.56**	**78.72**	**75.29**	**4.45**	**73.19**	**64.86**
BP_S	**4.91**	**58.60**	**80.99**	**4.56**	**77.96**	**74.04**	**4.65**	**72.34**	**63.57**
VFR	**5.18**	**60.02**	**82.42**	**4.88**	**76.78**	**78.80**	**4.91**	**69.48**	**71.72**

Note: * When Sample Sizes are equal, *KL* is equivalent to *L*1 and $OB_{W=0}$ is equivalent to *L*2.

sample size, it suggests that sample size is "powering" the tests. This implies that these rank-based tests are testing a different null hypothesis when there are skewed error distributions. It is likely that these procedures are testing a different null hypothesis of "stochastic homogeneity" (Vargha & Delaney, 1998) or equivalence of distribution functions (Lehman, 1975) rather than equivalence of variances (equation 1).

The non-parametric Levene (*NPL*) test (Nordstokke & Zumbo, 2010) performed surprisingly well in all conditions. Shear, Nordstokke, and Zumbo (2018), however, demonstrated that sampling from populations with unequal and unknown means can lead to incorrect (either inflated or decreased, depending on the magnitude of the mean separation) Type I error rates of the *NPL* test when error distributions are asymmetric. Furthermore, centering samples using either sample means, or medians did not correct the Type 1 error rates. We conducted a small simulation study to investigate this phenomenon and were able to reproduce their findings. This again implies issues with the null hypothesis being tested by rank-based tests. Given that the *FK* did not show the extreme inflations show by *CN* and that *NPL* performed well in many conditions, perhaps some method using aspects of each approach could provide a "good" non-parametric test for heteroscedasticity.

Due to the robustness of OLS, if the normality assumption holds for the ANOVA model for mean response, then the Levene family of procedures that perform ANOVA on a transformation of the residuals is generally robust, despite the fact that the *L*1, *L*2, *BF*, and *OB* transformations result in an asymmetric response variable (e.g., $|e|$; e^2; d; u). However, if the original response variable, y, has a skewed error term then these transformed response variables are highly skewed. To demonstrate this, we generated 10,000 replications of a response variable, y, with normally distributed errors for two samples of $n_j = 100$: one with a mean of $\mu_1 = 0$ and variance of $\sigma_1^2 = 1$ and a second group with a mean of $\mu_2 = 0$ and variance of $\sigma_2^2 = 3$.

Table 6. Rejection Rates under Null and Non-Null Variance Structures at α=0.05 for Total Sample Sizes of *N*=60 with Skewed ($\chi^2_{(1)}$) Error Distributions.

Sample Size Ratio	**1:1**			**2:1**			**3:1**		
Sample Sizes	***n*₁= 30**		***n*₂=30**	***n*₁ = 40**		***n*₂ = 20**	***n*₁= 45**		***n*₂ = 15**
Variances **Pooled Tests**	**σ^2_1=1** **σ^2_2=1**	**σ^2_1=1** **σ^2_2=2**	**σ^2_1=1** **σ^2_2=3**	**σ^2_1=1** **σ^2_2=1**	**σ^2_1=1** **σ^2_2=3**	**σ^2_1=3** **σ^2_2=1**	**σ^2_1=1** **σ^2_2=1**	**σ^2_1=1** **σ^2_2=3**	**σ^2_1=3** **σ^2_2=1**
*L*1	19.72	33.78	51.34	19.41	50.87	45.20	18.22	46.87	38.85
*KL**				19.65	52.21	43.98	18.43	48.70	37.29
*L*2	**4.97**	**10.35**	**18.51**	**4.91**	**28.62**	**7.21**	5.74	30.95	3.16
OB$_{W=0}$*				**5.25**	**29.51**	**6.97**	6.14	32.46	3.02
OB$_{W=0.5}$	**4.51**	**9.52**	**17.11**	**4.72**	**27.83**	**6.47**	**5.54**	**30.38**	**2.71**
BF	**5.01**	**12.86**	**25.57**	**5.05**	**28.70**	**16.18**	**4.60**	**26.86**	**10.17**
TM$_{10}$	10.82	21.87	36.96	10.55	39.16	28.40	8.76	34.72	20.93
Welch *df* Tests									
*L*1(W)	19.46	33.42	50.92	21.08	40.49	56.56	22.30	32.92	58.51
KL(W) *				20.78	41.40	55.41	21.62	34.10	56.80
*L*2(W)	**4.61**	**9.59**	**17.28**	6.58	10.06	27.84	9.78	7.63	34.42
OB$_{W=0}$(W) *				6.39	10.29	27.06	9.28	7.98	32.94
OB$_{W=0.5}$(W)	**4.20**	**8.92**	**16.09**	5.99	9.27	25.85	8.72	6.88	31.80
BF(W)	**4.80**	**12.48**	**24.89**	6.25	14.02	33.64	7.98	8.63	37.45
TM$_{10}$(W)	10.55	21.44	36.56	11.71	26.96	43.62	12.83	18.44	45.86
Rank-Based Tests									
NPL	**4.93**	**87.77**	**96.96**	**4.85**	**88.35**	**98.14**	**4.89**	**77.08**	**97.14**
CN	51.12	63.95	77.81	50.03	74.22	74.66	47.26	69.11	69.86
FK	16.53	28.61	43.61	16.51	39.51	41.16	13.85	31.97	36.87
Model Based Tests									
BP	40.71	53.35	67.01	39.12	63.35	64.47	35.63	57.37	59.88
BP$_S$	**5.10**	**10.52**	**18.86**	**5.07**	**28.97**	**7.57**	5.81	31.20	3.27
VFR	40.65	53.10	66.81	40.59	61.68	68.31	39.70	54.95	67.82

Note: * When Sample Sizes are equal, *KL* is equivalent to *L*1 and *OB*$_{W=0}$ is equivalent to *L*2.

The *L*1, *KL*, *L*2, *OB*$_{W=0}$, *OB*$_{W=0.5}$, and *BF* transformations were computed. We then generated samples with skewed errors sampled form a chi-square distribution with *df*=1.

Table 8 shows the Mean, Variance (s^2), Skewness (g^3), excess Kurtosis (g^4) for all 1,000,000 generated cases (10,000 replications of sample size of 100) for both samples. As can be seen in the left side of the table, the original response variable, *y*, has a means of 0 with variances of s_1^2 = 1 and s_2^2 = 3, respectively, and virtually no skewness or excessive kurtosis. As would be expected from distributional theory, transformations based on squared residuals (*L*2, *OB*) from a central unit normal distribution approximates a chi-square distribution with *df*=1. For Group 1 (top left panel), these transformations have mean of 1 (*df*); variance of 2 (2*df*), skewness of 2.83, and excess kurtosis of 12. For the transformations based on absolute values (*L*1, *KL*, *BF*), the variables approximate a chi distribution. In this case, *BF* is virtually equivalent to *L*1 and *KL* because the median equals the mean in a symmetric distribution such as the normal. For Group 2 (bottom left panel), transformations based on squared residuals (*L*2, *OB*) from a central normal distribution with σ_2^2 = 3 approximates the shape of a chi-square distribution with *df*=1; skewness approximately 2.83, and excess kurtosis approximately 12. The means of these transformations are approximately equal to the variance of *y*, (e.g., mean of *OB* ~ σ_2^2 = 3, and the variance of these transformation is approximately 18, $2df(\sigma_2^2)^2$.

For the results on the right side of the table, data were generated from a standardized Chi-Square distribution with *df*=1. As can be seen, the original response variable, *y*, has a means of 0 with approximately the expected skewness of 2.83 and excess kurtosis of 12, and variances of s_1^2 = 1 and s_2^2 = 3, respectively, as assured by the Headrick (2002) method. Transforming a skewed variable by centering then squaring or taking absolute value creates variables with inflated variances. As compared to residuals sampled from a central unit normal, squaring residuals (*L*2, *OB*) from an extremely skewed distribution retains the means

Table 7. Rejection Rates under Null and Non-Null Variance Structures at α=0.05 for Total Sample Sizes of N=84 with Skewed ($\chi^2_{(1)}$) Error Distributions.

Sample Size Ratio	**1:1**			**2:1**			**3:1**		
Sample Sizes	$\boldsymbol{n_1}$**= 42**		$\boldsymbol{n_2}$**=42**	$\boldsymbol{n_1}$ **= 56**		$\boldsymbol{n_2}$ **= 28**	$\boldsymbol{n_1}$**= 63**		$\boldsymbol{n_2}$ **= 21**
Variances **Pooled Tests**	$\sigma^2_1=1$ $\sigma^2_2=1$	$\sigma^2_1=1$ $\sigma^2_2=2$	$\sigma^2_1=1$ $\sigma^2_2=3$	$\sigma^2_1=1$ $\sigma^2_2=1$	$\sigma^2_1=1$ $\sigma^2_2=3$	$\sigma^2_1=3$ $\sigma^2_2=1$	$\sigma^2_1=1$ $\sigma^2_2=1$	$\sigma^2_1=1$ $\sigma^2_2=3$	$\sigma^2_1=3$ $\sigma^2_2=1$
*L*1	19.31	39.69	61.87	18.90	59.68	55.40	18.54	54.86	48.08
*KL**				18.90	60.63	54.49	18.82	56.70	46.60
*L*2	**4.81**	**13.07**	**25.03**	**4.95**	**34.80**	**10.87**	**5.22**	**36.77**	**4.28**
$OB_{W=0}$*				**5.14**	**35.66**	**10.53**	**5.57**	**38.19**	**4.14**
$OB_{W=0.5}$	**4.49**	**12.36**	**23.88**	**4.80**	**34.23**	**9.96**	**5.19**	**36.60**	**3.95**
BF	**5.13**	**17.17**	**35.40**	**4.76**	**37.31**	**25.15**	**4.94**	**35.95**	**16.92**
TM_{10}	9.51	25.36	46.03	9.62	47.36	37.76	8.71	42.80	29.04
Welch *df* Tests									
*L*1(W)	19.14	39.58	61.59	20.60	51.29	64.87	21.83	42.37	63.94
KL(W) *	19.14	39.58	61.59	20.39	52.24	64.03	21.24	43.64	62.51
*L*2(W)	**4.52**	**12.43**	**23.99**	6.55	13.10	36.15	9.75	9.36	41.86
$OB_{W=0}$(W) *	**4.52**	**12.43**	**23.99**	6.45	13.47	35.59	9.41	9.80	40.77
$OB_{W=0.5}$(W)	**4.25**	**11.77**	**22.96**	6.10	12.35	34.77	8.98	8.68	39.87
BF(W)	**5.03**	**16.92**	**34.99**	6.24	21.78	43.09	7.70	14.43	44.48
TM_{10}(W)	9.38	25.04	45.72	10.90	34.73	52.13	12.03	24.35	52.44
Rank-Based Tests									
NPL	**4.75**	**96.34**	**99.63**	**4.93**	**96.99**	**99.83**	**5.15**	**90.58**	**99.57**
CN	55.88	72.82	86.46	54.65	83.24	83.77	52.36	78.47	79.51
FK	19.49	35.34	55.18	18.58	49.79	53.11	16.69	43.58	47.06
Model Based Tests									
BP	41.47	57.91	75.14	40.49	69.95	73.04	38.73	65.19	68.69
BP_S	**4.93**	**13.16**	**25.41**	**5.02**	**35.12**	**11.21**	**5.35**	**37.09**	**4.39**
VFR	40.63	56.92	74.42	40.37	67.64	74.81	40.47	62.17	73.24

Note: * When Sample Sizes are equal, *KL* is identical to *L*1 and $OB_{W=0}$ is identical to *L*2.

of approximately 1 and 3 for Groups 1 and 2, respectively; however, the variances are inflated from approximately 2 to over 13 for Group 1; and for Group 2, the variance increases from 18 to over 120. Interestingly, our previous simulations showed that with adequately large sample sizes, the squared residuals approaches maintained test size; thus, the variance inflation due to skewed residuals does not necessarily affect Type 1 error rate, but it can affect power depending on the variance-sample size configuration.

With skewed residuals, *BF* is no longer equivalent to *L*1 and *KL*. Consistent with many other studies, our previous simulations show that *BF* maintained test size while *L*1 and *KL* inflated Type 1 error rates with skewed residuals. Miller (1968) showed that ANOVA on absolute values of residuals will be asymptotically incorrect if the population is not symmetric, but indicated median centering will provide the correct variance. In the table above, *BF* had a smaller mean than *L*1 and *KL*, suggesting estimation bias. Furthermore, *BF* had a larger variance that *L*1 and *KL*, which implies the test size inflation demonstrated by *L*1 and *KL* may be due to an underestimation of the variance.

To investigate the sampling properties of these transformations, we calculated the average of Mean, Variance, Skewness, and excess Kurtosis for each sample over the 10,000 replications. Table 9 reports the Means and *SD*s for the Mean, Variance (s^2), Skewness (g^3), and excess Kurtosis (g^4) calculated for each sample of n_j = 100 across the 10,000 replications. With normally distributed errors, the means and variances are estimated with reasonable accuracy with relatively small *SD*s surrounding them. The skew and kurtosis for the squared residual transformations (*L*2, *OB*) tend be underestimated in the sample.

When the residuals are sampled from an extremely skewed error distribution, the mean of the squared residual transformations (*L*2, *OB*) are relatively accurate; however, *SD*s around these mean estimates are inflated compared to results under the normality assumption. The variances for these transformations are

Table 8. Mean, Variance (s^2), Skewness (g^3), and Kurtosis (g^4) of each Response Variable over 10,000 replications of n_j = 100 (1,000,000) values for Normal and Chi-Square (*df*=1) Distributions.

	Normal Distribution (0, σ_j^2)				**Standardized Chi-Square ($\chi^2_{(1)}$)**			
Response	**Mean**	s^2	g^3	g^4	**Mean**	s^2	g^3	g^4
y	-0.00018	1.00008	0.00183	-0.00315	0.00003	0.99998	2.81365	11.80644
*L*1	0.79405	0.35944	0.99372	0.85796	0.68261	0.52378	4.20352	26.97209
KL	0.79805	0.36307	0.99372	0.85796	0.68605	0.52907	4.20352	26.97209
BF	0.79185	0.36837	0.99547	0.86054	0.60146	0.78833	3.44044	16.99818
*L*2	0.98995	1.95592	2.80985	11.71095	0.98974	13.27553	13.62516	393.90418
$OB_{W=0}$	0.99995	1.99563	2.80985	11.71095	0.99974	13.54508	13.62516	393.90418
$OB_{W=0.5}$	0.99995	2.01584	2.80985	11.71100	0.99974	13.68220	13.62527	393.90505
	Mean	s^2	g^3	g^4	**Mean**	s^2	g^3	g^4
y	-0.00033	3.00402	-0.00197	-0.00026	0.00037	3.00019	2.83133	12.15842
*L*1	1.37569	1.08114	0.99363	0.86703	1.18217	1.57252	4.25127	28.05752
KL	1.38262	1.09206	0.99363	0.86703	1.18813	1.58840	4.25127	28.05752
BF	1.37187	1.10784	0.99544	0.87003	1.04158	2.36378	3.46728	17.55470
*L*2	2.97367	17.68200	2.83984	12.28015	2.97005	122.75847	14.32639	413.82074
$OB_{W=0}$	3.00371	18.04102	2.83984	12.28015	3.00005	125.25096	14.32639	413.82074
$OB_{W=0.5}$	3.00371	18.22374	2.83984	12.28006	3.00005	126.51944	14.32648	413.82120

drastically inflated with extremely large *SD*s. Again, *BF* had smaller means while having larger variances than *L*1 and *KL*. Furthermore, the *SD*s for the variances were larger than the SDs for *L*1 and *KL*.

Since means and variances are the critical components of the test statistics under investigation, we report their distributional properties to further investigate the sampling properties on these transformations. As can be seen in Table 10, the variance and non-normality of the distribution of the mean increase when the residuals are sampled from a skewed error distribution. For the transformations that employ absolute values (*L*1, *KL*, *BF*), the variance of the variance distribution increases substantially, while non-normality of the distribution of the variance increases slightly when the residuals are sampled from a skewed error distribution. Again, *BF* had a larger variance that *L*1 and *KL*, but had a slightly less skewed distribution. It is interesting that the distributions of *L*1 and *KL* have more skewness and kurtosis but less variance; however, this still implies the test size inflation demonstrated by L1 and *KL* is due to an underestimation of the variance.

Again, the means of the squared residual transformations (*L*2, *OB*) are relatively accurate; however, *SD*s around these mean estimates are inflated compared to results under the normality assumption. The variances for these transformations are drastically inflated with extremely large *SD*s when the residuals are sampled from an extremely skewed error distribution. This variance inflation is likely due to the extreme skewness (*Skew*(s^2) ~ 17.8 for Group 1, σ^2 = 1; *Skew*(s^2) ~ 14 for Group 2, σ^2 = 3) for the distributions of variances for the squared residual transformations (*L*2, *OB*) as compared the skewness of the variance distribution under normality (*Skew*(s^2) ~ 1.2 for Group 1, σ^2 = 1; *Skew*(s^2) ~ 1.4 for Group 2, σ^2 = 3). The fact the test for heterogeneous variance based on square residuals (*L*2, *OB*) are more robust than test based the absolute values of residuals (*L*1, *KL*), despite the noted differences in variances and distributional shape, warrants further investigation in the distributional properties of these tests.

Recommendations

Although Test Size inflations were observed with smaller sample sizes, tests based on squared residuals (*L*2, the *OB* tests, *BPs*) performed well the most conditions simulated and are recommended because generalizing then to more complex statistical models is straightforward. One of the most commonly used tests, *L*1, and the *KL* modification of this approach suggested by Keyes & Levy (1997), cannot be recommended due to the drastically inflated Test Size when error distributions were skewed. Similarly, the trimmed means approach cannot be recommended due to drastically inflated Type 1 error rates. Although asymmetric trimming has been suggested (Keselman et al., 2008), this option is not available is many software, and therefore, data analysts would have to program their own algorithm to perform such a procedure.

Table 9. Means and Standard Deviations (*SD*) for the Mean, Variance (s^2), Skewness (g^3), and Kurtosis (g^4) for each sample of n_j = 100 computed over the 10,000 replications for each response variable with Normal and Chi-Square (*df*=1) Distributions.

Normal	**Mean**		**Variance (s^2)**		**Skewness (g^3)**		**Kurtosis (g^4)**	
Group 1	**Mean**	***SE*(Mean)**	**Mean(s^2)**	***SD*(s^2)**	**Mean(g^3)**	***SD*(g^3)**	**Mean(g^4)**	***SD*(g^4)**
y	-0.00018	0.10065	0.99995	0.14352	0.00099	0.24065	-0.00392	0.47264
*L*1	0.79405	0.06075	0.35934	0.06165	0.95973	0.26737	0.75309	1.11045
KL	0.79805	0.06105	0.36297	0.06227	0.95973	0.26737	0.75309	1.11045
BF	0.79185	0.06065	0.36838	0.06281	0.96131	0.26699	0.75595	1.11077
*L*2	0.98995	0.14209	1.95528	0.72697	2.47145	0.71696	8.04552	6.18553
$OB_{W=0}$	0.99995	0.14352	1.99498	0.74173	2.47145	0.71696	8.04552	6.18553
$OB_{W=0.5}$	0.99995	0.14352	2.01539	0.74931	2.47145	0.71696	8.04552	6.18553
Group 2	**Mean**	***SE*(Mean)**	**Mean(s^2)**	***SD*(s^2)**	**Mean(g^3)**	***SD*(g^3)**	**Mean(g^4)**	***SD*(g^4)**
y	-0.00033	0.17421	3.00371	0.42348	-0.00166	0.24038	-0.00054	0.47641
*L*1	1.37569	0.10393	1.08115	0.18292	0.95951	0.26899	0.75260	1.12559
KL	1.38262	0.10446	1.09207	0.18477	0.95951	0.26899	0.75260	1.12559
BF	1.37187	0.10371	1.10817	0.18681	0.96096	0.26942	0.75565	1.12791
*L*2	2.97367	0.41924	17.68307	6.67686	2.47280	0.72898	8.08353	6.34150
$OB_{W=0}$	3.00371	0.42348	18.04211	6.81242	2.47280	0.72898	8.08353	6.34150
$OB_{W=0.5}$	3.00371	0.42348	18.22668	6.88212	2.47280	0.72898	8.08353	6.34150
$\chi^2_{(1)}$	**Mean**		**Variance (s^2)**		**Skewness (g^3)**		**Kurtosis (g^4)**	
Group 1	**Mean**	***SE*(Mean)**	**Mean(s^2)**	***SD*(s^2)**	**Mean(g^3)**	***SD*(g^3)**	**Mean(g^4)**	***SD*(g^4)**
y	0.00003	0.10116	0.99974	0.37615	2.46668	0.71333	7.99230	6.11124
*L*1	0.68261	0.10604	0.51772	0.25253	3.48424	1.00186	16.21985	10.25452
KL	0.68605	0.10658	0.52295	0.25509	3.48424	1.00186	16.21985	10.25452
BF	0.60146	0.08960	0.78818	0.34153	2.98486	0.78696	11.17758	7.56060
*L*2	0.98974	0.37239	13.26955	26.20130	5.79360	1.58126	39.63311	21.37640
$OB_{W=0}$	0.99974	0.37615	13.53898	26.73330	5.79360	1.58126	39.63311	21.37640
$OB_{W=0.5}$	0.99974	0.37615	13.67749	27.00678	5.79360	1.58126	39.63311	21.37640
Group 2	**Mean**	***SE*(Mean)**	**Mean(s^2)**	***SD*(s^2)**	**Mean(g^3)**	***SD*(g^3)**	**Mean(g^4)**	***SD*(g^4)**
y	0.00037	0.17361	3.00005	1.12113	2.47370	0.72983	8.07091	6.37247
*L*1	1.18217	0.18075	1.55540	0.76144	3.49326	1.02328	16.33284	10.59997
KL	1.18813	0.18166	1.57112	0.76913	3.49326	1.02328	16.33284	10.59997
BF	1.04158	0.15359	2.36383	1.01894	2.99358	0.80857	11.28490	7.88690
*L*2	2.97005	1.10992	122.75408	246.65710	5.79571	1.60448	39.70314	21.72725
$OB_{W=0}$	3.00005	1.12113	125.24649	251.66524	5.79571	1.60448	39.70314	21.72725
$OB_{W=0.5}$	3.00005	1.12113	126.52777	254.23980	5.79571	1.60448	39.70314	21.72725

Welch *df* adjustment applied to *L*2 and the *OB* tests showed inconsistent Test Size. These modifications inflated Test size with sample size *N* < 100 with skewed errors in our simulations; however, these procedures increase power with symmetric error distributions and positive pairing of variances and sample sizes. Therefore, we do not recommend that Welch *df* adjustment be applied to *L*2 or the *OB* tests with smaller sample sizes. More research into the degree of skewness in context with sample size for these procedures to be robust needs to be conducted. In a preliminary investigation not reported, we found that with a total sample size of *N*=500 with a 3:1 sample size ratio (n_1=375; n_2=125), the Welch df correction applied to square residuals (*L*2; *OB*) did not maintain test size when residuals were sample from the $\chi^2_{(1)}$ distribution.

Despite a slight advantage in power under ideal circumstances, the original *BP* and *VFR* cannot be recommended due to their sensitivity to even slight departures from normality. The *FK* and *NPL* rank-based tests performed well in a number of conditions simulated, however, more work in the area non-parametric tests of homogenous variances is needed.

As previously mentioned, another reason that methods based on squared residuals, namely the studentized Breusch-Pagan (*BPs*), *L*2, and $OB_{W=0}$ tests, are recommended is that it is straightforward to

Table 10. Mean, Variance, Skewness, and Kurtosis for the Mean and Variance (s^2) for each sample of n_j = 100 computed over the 10,000 replications for each response variable with Normal and Chi-Square (*df*=1) Distributions.

Normal	**Mean (*M*)**				**Variance (s^2)**			
Group 1	***Mean*(*M*)**	***Var*(*M*)**	***Skew*(*M*)**	***Kurt*(*M*)**	**Mean(s^2)**	***Var*(s^2)**	***Skew*(s^2)**	***Kurt*(s^2)**
y	-0.00018	0.01013	0.02877	0.08932	0.99995	0.02060	0.27936	0.23981
*L*1	0.79405	0.00369	0.08060	0.12591	0.35934	0.00380	0.37072	0.19126
KL	0.79805	0.00373	0.08060	0.12591	0.36297	0.00388	0.37072	0.19126
BF	0.79185	0.00368	0.08254	0.13175	0.36838	0.00395	0.37382	0.18439
*L*2	0.98995	0.02019	0.27936	0.23981	1.95528	0.52848	1.18323	3.05746
$OB_{W=0}$	0.99995	0.02060	0.27936	0.23981	1.99498	0.55016	1.18323	3.05746
$OB_{W=0.5}$	0.99995	0.02060	0.27936	0.23981	2.01539	0.56147	1.18323	3.05746
Group 2	***Mean*(*M*)**	***Var*(*M*)**	***Skew*(*M*)**	***Kurt*(*M*)**	**Mean(s^2)**	***Var*(s^2)**	***Skew*(s^2)**	***Kurt*(s^2)**
y	-0.00033	0.03035	-0.02885	0.07944	3.00371	0.17933	0.29235	0.15844
*L*1	1.37569	0.01080	0.13270	0.05539	1.08115	0.03346	0.39531	0.41264
KL	1.38262	0.01091	0.13270	0.05539	1.09207	0.03414	0.39531	0.41264
BF	1.37187	0.01076	0.13194	0.04873	1.10817	0.03490	0.38930	0.35906
*L*2	2.97367	0.17576	0.29235	0.15844	17.68307	44.58041	1.43693	4.78333
$OB_{W=0}$	3.00371	0.17933	0.29235	0.15844	18.04211	46.40912	1.43693	4.78333
$OB_{W=0.5}$	3.00371	0.17933	0.29235	0.15844	18.22668	47.36352	1.43693	4.78333
$\chi^2_{(1)}$	**Mean (*M*)**				**Variance (s^2)**			
Group 1	***Mean*(*M*)**	***Var*(*M*)**	***Skew*(*M*)**	***Kurt*(*M*)**	**Mean(s^2)**	***Var*(s^2)**	***Skew*(s^2)**	***Kurt*(s^2)**
y	0.00003	0.01023	0.31804	0.30827	0.99974	0.14149	1.32709	4.16387
*L*1	0.68261	0.01124	0.33265	0.29534	0.51772	0.06377	1.89296	8.78852
KL	0.68605	0.01136	0.33265	0.29534	0.52295	0.06507	1.89296	8.78852
BF	0.60146	0.00803	0.31284	0.31941	0.78818	0.11664	1.45320	4.82212
*L*2	0.98974	0.13867	1.32709	4.16387	13.26955	686.50827	17.84739	619.38288
$OB_{W=0}$	0.99974	0.14149	1.32709	4.16387	13.53898	714.66908	17.84739	619.38288
$OB_{W=0.5}$	0.99974	0.14149	1.32709	4.16387	13.67749	729.36617	17.84739	619.38288
Group 2	***Mean*(*M*)**	***Var*(*M*)**	***Skew*(*M*)**	***Kurt*(*M*)**	**Mean(s^2)**	***Var*(s^2)**	***Skew*(s^2)**	***Kurt*(s^2)**
y	0.00037	0.03014	0.26036	0.11903	3.00005	1.25694	1.31566	3.82258
*L*1	1.18217	0.03267	0.30905	0.15622	1.55540	0.57978	1.93774	8.35067
KL	1.18813	0.03300	0.30905	0.15622	1.57112	0.59156	1.93774	8.35067
BF	1.04158	0.02359	0.27453	0.12332	2.36383	1.03824	1.49332	4.99183
*L*2	2.97005	1.23193	1.31566	3.82258	122.75408	60839.72	13.97286	386.57797
$OB_{W=0}$	3.00005	1.25694	1.31566	3.82258	125.24649	63335.39	13.97286	386.57797
$OB_{W=0.5}$	3.00005	1.25694	1.31566	3.82258	126.52777	64637.88	13.97286	386.57797

generalize these methods to more complex linear and mixed linear models. The *BF* test is recommended for ANOVA designs because tests for heteroscedasticity can be generalized to factorial designs based on procedures using medians to center the data (Boos & Brownie, 2004), or applying some rank transformation; however, generalizing these methods to more complex modeling procedures (e.g., multiple linear regression) would not be straightforward. Furthermore, rank-based methods would have to potentially align the ranks to obtain a valid procedure for factorial designs (e.g., Higgins & Tashtoush, 1994; Beasley, 2002).

In summary, testing for heteroscedasticity can be complicated by many factors, even in a simple two-group ANOVA. In many circumstances, a valid test may be elusive. We agree with other researchers (Zimmerman, 2004) and do not recommend performing preliminary tests of the homoscedasticity assumption to decide which statistical analysis to perform. If researchers are interested in modeling means and are concerned about violations to the homoscedasticity assumption, then they should use robust procedures that do not require the homogeneity of variances assumption (Wilcox et al., 1986), such as the ANOVA with separate variance error terms and Welch *df* adjustments for between-subjects designs. Although not investigated in this study this would generalize to using heteroscedasticity consistent covariance matrices (HCCMs) to estimate standard error is linear regression (Long & Ervin, 2000; White, 1980), and heteroscedastic models with Kenward and Roger (1997) dfs for mixed linear models (Littell et al., 2006).

If researchers are interested in testing for heteroscedasticity as an outcome of interest, then we recommend planning accordingly. This paper demonstrates the problems encountered when error distributions are skewed. If researchers suspect skewness related to the outcome of interest (e.g., reaction time), then they should consider some form of a between-subjects (or other ANOVA-type) design where the *BF* test can be used. If researcher do not suspect serious departures from normality, then tests based on squared residuals may provide a slight advantage in power and applying Welch ANOVA with adjusted *df*s to the squared residual approaches may provide additional but slight increases in power. Furthermore, analyzing squared residuals has a straightforward generalization to more complex statistical analyses; however, the performance of such approaches needs to be investigated.

References

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

Beasley, T. M. (1995). Comparison of general linear model approaches to testing variance heterogeneity in true and quasi-experiments. *Multiple Linear Regression Viewpoints*, *22*, 36-54.

Beasley, T. M. (2002). Multivariate aligned rank test for interactions in multiple group repeated measures designs. *Multivariate Behavioral Research*, *37*, 197-226.

Breusch, T. S., & Pagan, A. R. (1979). A Simple Test for Heteroskedasticity and Random Coefficient Variation. *Econometrica*, *47*(*5*), 1287-1294.

Brown, M. B., & Forsythe, A. B. (1974). Robust Tests for 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.

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

Conover, W. J. (1971). *Practical Nonparametric Statistics*. Wiley.

Conover, W. J., & Iman, R. L. (1987). Some Exact Tables for the Squared Ranks Test. *Communications in Statistics: Simulation and Computation*, *7*(*5*), 491-513.

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. *Technometrics*, *23*, 351-361

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

Higgins, J. J., & Tashtoush, S. (1994). An aligned rank transform test for interaction. *Nonlinear World 1*(*2*), 201-211.

Kenward, M. G., & Roger, J. H. (1997). Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood. *Biometrics, 53*, 983–997.

Keselman, H.J., Games, P.A., & Clinch, J. J. (1979). Tests for homogeneity of variance. *Communications in Statistics: Simulations and Computation*, *B8*, 113-129.

Keselman, H. J., Wilcox, R. R., Algina, J., Othman, A. R., & Fradette, K. (2008). A comparative study of robust test for spread: Asymmetric trimming strategies. *British Journal of Mathematical and Statistical Psychology*, *61*, 235-253.

Kowalski, S. M., Taylor, J. A., Askinas, K. M., Wang, Q., Zhang, Q., Maddix, W. P., & Tipton, E. (2020). Examining Factors Contributing to Variation in Effect Size Estimates of Teacher Outcomes from Studies of Science Teacher Professional Development. *Journal of Research on Educational Effectiveness*, *13*(*3*), 430-458.

Lehmann, E. L. (1975). Nonparametrics: Statistical Methods Based on Ranks, San Francisco. Holden-Day.

Levene, H. (1960). Robust Tests for the Equality of Variance. In I. Olkin, S. G. Ghurye, W. Hoeffding, W. G. Madow, & H. B. Mann (Eds.) *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, (pp. 278–292). Palo Alto, CA: Stanford University Press.

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

Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., & Schabenberger, O. (2006). SAS® for Mixed Models (2nd ed.). Cary NC: SAS® Institute.

Long, J. S., & Ervin, L. H. (2000) Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician 54*, 217–224.

Mara, C., & Cribbie, R. A. (2017). Equivalence of population variances: Synchronizing the objective and analysis. *Journal of Experimental Education*, *86*, 442-457.

Mattison, J. A., Colman, R. J., Beasley, T. M., Allison, D. B., Kemnitz, J. W., Roth, G. S., Ingram, D. K., Weindruch, R., de Cabo, R., & Anderson, R. M. (2017). Caloric restriction improves health and survival of rhesus monkeys. *Nature Communications, 8,* 14063. doi:10.1038/ncomms14063.

Miller, R. G. (1968). Jackknifing Variances. *Annals Mathematical Statistics*, *39*, 567–582.

Muller, K.E. (2009). Analysis of variance concepts and computations. *Wiley Interdisciplinary Reviews Computational Statistics*, *1*, 271-282.

Ng, V. K., & Cribbie, R.A. (2019). The gamma generalized linear model, log transformation, and the robust Yuen-Welch test for analyzing group means with skewed and heteroscedastic data. *Communications in Statistics: Simulation and Computation*, *48*, 2269-2286.

Nordstokke, D. W., & Zumbo, B. D. (2007). A cautionary tale about Levene's tests for equality of variances. *Journal of Educational Research and Policy Studies*, *7*, 1-14.

Nordstokke, D.W. & Zumbo, B.D. (2010). A new nonparametric Levene test for equal variances. *Psicologica*, *31*, 401-430

Nordstokke, D. W., Zumbo, B. D., Cairns, S. L., & Saklofske, D. H. (2011). The operating characteristics of the nonparametric Levene test for equal variances with assessment and evaluation data. *Practical Assessment, Research & Evaluation*, *16*(*5*), 1–8.

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.

O'Brien, R. G. (1981). A Simple Test for Variance Effects in Experimental Designs. *Psychological Bulletin*, *89*, 570–574.

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

Parra-Frutos, I. (2009). The behaviour of the modified Levene's test when data are not normally distributed. *Computational Statistics*, *24*, 671-693.

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

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

Shear, Benjamin R., Nordstokke, David W., & Zumbo, Bruno D. (2018) A Note on Using the Nonparametric Levene Test When Population Means Are Unequal. *Practical Assessment, Research & Evaluation*, *23*(*13*). Available online: http://pareonline.net/getvn.asp?v=23&n=13

Snedecor G. W., & Cochran W. G. (1989). *Statistical Methods* (8th ed.). Ames: Iowa State University Press.

Taha, M. A. H. (1964). Rank test for scale parameter for asymmetrical one-sided distributions. In *Publications de L'Institute de Statistiques*, Vol. 13, (pp. 169-180). Paris: de L'Université de

Vargha, A., & Delaney, H. D. (1998). The Kruskal-Wallis test and stochastic homogeneity. *Journal of Educational and Behavioral Statistics*, *23*(*2*), 170–192.

Wang, Y., Rodríguez de Gil, P., Chen, Y. H., Kromrey, J. D., Kim, E. S., Pham, T., Nguyen, D., & Romano, J. L. (2017). Comparing the Performance of Approaches for Testing the Homogeneity of Variance Assumption in One-Factor ANOVA Models. *Educational & Psychological Measurement*, *77*, 305-329.

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

Western, B. & Bloome, D. (2009). Variance function regressions for studying inequality. *Sociological Methodology*, *39*(*1*), 293–326.

White, H. (1980) A heteroskedastic consistent covariance matrix estimator and a direct test of heteroskedasticity. *Econometrica*, *48*, 817–838.

Wilcox, R.R., Charlin, V.L., & Thompson, K.L. (1986). New Monte Carlo results on the robustness of the ANOVA F, W, and F* statistics. *Communications in Statistics: Simulation and Computation*, *B15*, 933-943.

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

Zimmerman, D. W. (1996). A Note on Homogeneity of Variance of Scores and Ranks. *Journal of Experimental Education*, *64*(*4*), 351-362.

Zimmerman, D. W., & Zumbo, B. D. (1993a). Rank transformations and the power of the Student t-test and Welch's t-test for non-normal populations with unequal variances. *Canadian Journal of Experimental Psychology*, *47*, 523-539.

Zimmerman, D. W., & Zumbo, B. D. (1993b). The relative power of parametric and nonparametric statistical methods. In G. Keren and C. Lewis (Eds.) *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 481-517). Hillsdale, NJ: Erlbaum.

Send correspondence to:

T. Mark Beasley
University of Alabama at Birmingham
Email: mbeasley@uab.edu