Comparison of General Linear Model Approaches to Testing Variance Heterogeneity in True and Quasi-Experiments

T. Mark Beasley
St. John's University, New York

Simulation results indicated that when groups were sampled from the same platykurtic population the O'Brien (1981) transformation was preferred except when a positive sample size/variance correlation existed, then the Welch test performed on the O'Brien scores was more powerful. Also consistent with previous research, when grouped data were sampled from the same leptokurtic population the Brown & Forsythe (1974) transformation was preferred for equal sample sizes. The O'Brien test was more powerful with an indirect sample size/variance relationship regardless on distribution shape(e.g., Algina et al., 1989; Olejnik & Algina, 1987, 1988). The study also demonstrated that the Welch test performed on Brown-Forsythe scores was more powerful when a positive sample size/variance correlation existed in leptokurtic data. Furthermore, choosing a test of variance based on an initial test of kurtosis may improve power (Ramsey, 1994). When data were sampled from populations with drastically different shapes (kurtosis), the Type I error rate of most tests was unstable excluding the Hartley *Fmax* test which performed surprisingly well.

The analysis of variance (ANOVA) is one of the most widely used statistical procedures in educational research. Namely, it is the technique of choice for True Experiments in which members sampled from the same population are randomly assigned to treatment conditions. In field research and Quasi-Experiments, comparisons among groups are also of interest; however, there exists the possibility that these groups are sampled from different populations. In either case, behavioral researchers often compare groups with different distributional properties, which may be a result of (a) sampling from different populations or (b) an experimental treatment affecting something other than central location. Thus, as far as analytic procedures are concerned, the distinction between True and Quasi-Experiments becomes ambiguous. For the purposes of this study, a True Experiment is defined as sampling data from a single population and randomly assigning cases to groups. A Quasi-Experiment is defined by separately sampling data from populations which differ in distributional shape (i.e., skew and kurtosis).

One of the most critical conditions for any linear modeling procedure involves the assumption of homoscedasticity across levels of the independent variable. In the ANOVA, it follows that heterogeneous variances may obscure the magnitude of test statistics for comparisons among means. Thus, testing variance equality appropriately is important in checking a vital assumption of the ANOVA. Furthermore, despite the existence of differences in central location, heterogeneous variances may constitute substantive and theoretically valuable results. That is, it may be interesting to know that the responses of two separately sampled populations differ in scale or that an experimental treatment significantly affects response variability.

Traditional tests of variance homogeneity (e.g., Hartley's *Fmax*) can be very simple, calculating the ratio of two sample variances. The *Fmax* test, however, has long been known to be extremely sensitive to deviations in kurtosis (Box, 1953; Scheffe', 1959). Slight departures from normality which involve kurtosis have been shown to make substantial difference in the Type I error rate of the *Fmax* test (Pearson & Please, 1975). For instance, Hartley's *Fmax* test has been shown to be conservative for platykurtic distributions and liberal when distributions have positive kurtosis (Durrand, 1969). Although several tests of variance have been proposed, the *Fmax* remains popular in a variety of applied studies because of its simplicity.

In a simulation study, Conover, Johnson, and Johnson (1981) compared several procedures for testing homogeneous variances and found that most are liberal (i.e., the Type I error rate was considerably larger that the nominal alpha). Thus few tests exist that actually control the Type I error rate. Over the past two decades, robust tests of variances based on applying the ANOVA to transformed scores (e.g., Brown & Forsythe, 1974; O'Brien, 1981) have been proposed. Under conditions of a "True Experiment" and equal sample sizes, these tests have been shown to be powerful in a variety of population distributions (Algina, Olejnik, & Ocanto, 1989; Olejnik & Algina, 1987; Ramsey, 1994; Ramsey & Brailsford, 1990). The Brown-Forsythe (BF) test has been criticized because it has low power for small, odd

sample sizes and only moderate power for platykurtic and normal populations (O'Brien, 1981; Olejnik & Algina, 1987). Under these same conditions, the most common form of the O'Brien (OB) procedure has been shown to be more powerful than BF. Also for unequal sample sizes, OB has been suggested for platykurtic distributions and BF with symmetric and/or leptokurtic distributions (Algina et al., 1989).

Despite these recommendations based on the kurtosis of distributions, a criterion for identifying population shape was not suggested. Ramsey and Brailsford (1990) noted that tests of kurtosis could be used to decide between BF and *Fmax*. Following the suggestions of previous studies, Ramsey (1994) has recently suggested two conditional procedures based on testing kurtosis for each group separately. Ramsey's results confirmed the robustness of OB and BF but indicated that optimal power can be established with the conditional procedure of testing kurtosis to decide between the these tests. However, the power of these conditional procedures has been shown to be dependent on the power of the test of kurtosis. Also, Ramsey's results are limited in the sense that only the conditions of a True Experiment were simulated. That is, the two groups were sampled from the same population. In field research and Quasi-Experiments, comparisons of groups sampled from different populations are often of interest and the suggested conditional procedures have yet to be fully investigated under such conditions.

Olejnik and Algina (1988) found that both OB and BF held the Type I error rate for a limited number of distributions which differed in location and form. The OB tended to be most powerful with equal sample sizes and with an inverse relationship between sample sizes and population variances (i.e., larger sample has the smaller variance). When sample sizes and population variances had a direct relationship (i.e., larger sample has the larger variance), using OB transformed scores as dependent variables and performing the Welch (1951) statistic was the most powerful procedure.

A variety of nonparametric tests of variance are also available; however, they have presented problems with robustness and low power. Two of the better known procedures were proposed by Klotz (1962) and Siegel and Tukey (1960). When data were sampled from a normal population, both tests demonstrated the appropriate Type I error rate (Penfield & Koffler, 1985; Olejnik & Algina, 1985). Also, the Klotz test had power equal to or greater than the power of OB or BF when samples differed in variance only. However, both the Siegel-Tukey and Klotz tests were strongly affected by differences in central location (Moses, 1963). When the sampled distributions share the same asymmetric shape but differ in location, the tests are liberal. Yet, both tests become less powerful as location parameters increase when groups share the same symmetric shape (Olejnik & Algina, 1985). To date, attempts to modify these tests through mean- and median-alignment have not drastically improved their statistical properties (e.g., Conover et al., 1981; Olejnik & Algina, 1988).

Thus the purpose of this study was to investigate the robustness and power of OB, BF, and the use of the Welch statistic on these transformed scores (W_{OB} and W_{BF}, respectively) under conditions of a *True Experiment* (i.e., groups are randomly constructed from the same population) and a *Quasi-Experiment* (i.e., groups are sampled from two different populations). Furthermore, the effectiveness of conditional procedures based on tests of kurtosis (e.g., Ramsey, 1994) was examined. Although tests of variance are themselves of interest, in most educational research differences in central location are to be expected; therefore, nonparametric procedures such as the Siegel-Tukey and Klotz tests were excluded from this study. Under several circumstances, the results were expected to replicate those of Ramsey (1994) and Olejnik and Algina (1987, 1988). Furthermore, the findings of this study should address the issue of the appropriate procedure for testing variances in Quasi-Experiments in which the populations may differ in variance and form and the samples differ in size.

Statistics for Testing Variances

Although many statistical tests for comparing population variances have been developed, only a few of these procedures have demonstrated robustness when populations are nonnormal (i.e., Conover et al., 1981). Of these tests, the general linear model procedures, which involve performing the ANOVA (or some variant) on transformed scores, have shown both robustness and superior power.

Hartley's Fmax test. This test was investigated because of its wide use and known properties when kurtosis deviates form normality. The *Fmax* test is the ratio of the largest to the smallest of *J* variance,

$$Fmax = \frac{s^2_{largest}}{s^2_{smallest}}. \qquad (1)$$

The degrees of freedom are ($n_{largest}$ - 1) for the numerator and ($n_{smallest}$ - 1) for the denominator. Although it is often recommended that the *Fmax* test only be used with approximately equal sample *n*'s, its statistical properties were examined under all condition of this study. Critical values were obtained from the sampling distribution derived by Hartley (1950).

Brown-Forsythe Transformation. To test differences in variances, Levene (1960) proposed using the ANOVA but replacing each score, y_{ij}, of subject *i* within group *j* with the absolute deviation from its respective group mean. Although this procedure is fairly robust, it was found not to be adequately powerful (Conover et al., 1981). Brown and Forsythe (1974)

proposed applying the ANOVA to absolute deviations from respective group medians, m_j, such that:

$$b_{ij} = |y_{ij} - m_j| \quad . \qquad (2)$$

O'Brien Transformation. O'Brien (1979) proposed that the original score, y_{ij}, of subject i in group j be replaced with

$$r_{ij}(w) = \frac{(w + n_j - 2)\, n_j (y_{ij} - \bar{y}_j)^2 - w s_j^2 (n_j - 1)}{(n_j - 1)(n_j - 2)} \qquad (3)$$

where w is a parameter ranging between zero and one and $\bar{y}_j$ equals the mean, s_j^2 equals the variance, and n_j equals the sample size of group j. For most cases, O'Brien (1981) has recommended a value of $w = 0.5$ from which the group means for r in (3) are the variances of each group y: $\bar{r}_j = s_j^2$. The ANOVA is performed on the transformed r values.

Welch Statistic. It is not known whether the OB or BF tests are asymptotically distribution free. Furthermore, because the variance of r is dependent on sample size, O'Brien (1981) suggested using a Welch (1951) approximate degrees of freedom analysis on r values in place of the ANOVA when sample sizes are not equal (WOB). This procedure may also be performed on BF transformed scores (WBF). The Welch statistic is calculated by

$$W = \frac{\sum_{j=1}^{J} c_j (\bar{y}_j - \tilde{y})^2 / (J - 1)}{1 + \frac{2(J-2)}{J^2 - 1} \sum_{j=1}^{J} (1 - \frac{c_j}{c})^2 / (n_j - 1)} \qquad (4)$$

where J equals the number of groups, $c_j = n_j / s_j^2$, $c_* = \sum c_j$, and $\tilde{y} = \sum c_j r_j / c$. The Welch statistic is approximately distributed as F with degrees of freedom equal to $(J - 1)$ and

$$\left[\frac{3}{J^2 - 1} \sum_{j=1}^{J} (1 - \frac{c_j}{c})^2 / (n_j - 1) \right]^{-1} \qquad (5)$$

For $J = 2$ groups, the degrees of freedom in (5) follow the Satterthwaite (1946) formula.

Conditional Tests. Based on the simulation studies of Olejnik & Algina (1987, 1988), BF is preferred for leptokurtic populations, while OB is recommended for normal and platykurtic distributions. To achieve optimal power, Ramsey (1994) proposed two procedures for testing variances that are conditioned on applying a test for kurtosis.

Pearson's traditional sample measure of population kurtosis, γ_2, in group j is $b_2 = m_4 / m_2^2$, where $m_r = \sum (y_{ij} - \bar{y}_j)^r / n_j$. Thus m_2 is the second moment about the mean, the biased sample variance. Although standardized population moments for skewness and kurtosis provide popular significance tests, Ramsey and Ramsey (1993) have supplied a detailed and accurate table of critical values for b_2, which are used to test kurtosis against the null hypothesis (Ho: $\beta_2 = 3$).

For the tests proposed by Ramsey, tests of kurtosis are applied in each of the two samples at the $\alpha = .05$ significance level. A score of -1, 0, or +1 is recorded depending on whether the test of b_2 indicates that the distribution was significantly platykurtic, nonsignificant, or significantly leptokurtic, respectively. Combining scores from the two samples results in a total score, S, ranging from -2 to +2. In a J-group study, S would range from $-J$ to $+J$. The test of kurtosis is taken as identifying the population for the entire experiment as platykurtic if $S \leq -1$, mesokurtic if $S = 0$, and leptokurtic if $S \geq +1$. In one conditional procedure, OB_{BF}, kurtosis is tested and OB is applied if the samples are platykurtic or mesokurtic ($S \leq 0$) and BF if the distributions are significantly leptokurtic ($S \geq +1$). This approach is based on the recommendations of Olejnik and Algina (1987) but does not control the Type I error rate under certain distributional conditions; therefore, Ramsey (1994) suggested another conditional procedure that demonstrated superior power and adequate robustness. This approach, BF_{OB}, involves testing the fourth moment and applying OB with significantly platykurtic distributions ($S \leq -1$) and BF otherwise ($S \geq 0$).

Methods

Consistent with previous studies (e.g., Miller, 1968; Olejnik & Algina, 1987, 1988; Ramsey & Brailsford, 1990), the present investigation was restricted to the two-group case. These studies yielded results congruent with multi-group studies. Furthermore, the restriction to two groups allows more careful consideration of other factors. Since previous studies have indicated that shifts in central location have little to no effect on general linear model tests of

variance, (Beasley & O'Connor, 1995; Olejnik & Algina, 1988), three population variables were manipulated: shape in the form of kurtosis (γ_2), variance in one group, (σ^2); sample size (n_j).

Conditions

Population Kurtosis. Previous studies have indicated that skewness affects the robustness and power of nonparametric tests (Olejnik & Algina, 1988) but only affects the statistical properties of parametric tests in combination with nonnormal kurtosis (Conover et al., 1981; Olejnik & Algina, 1988; Pearson & Please, 1975). The normal and six nonnormal distributions that had no skewness but varied in kurtosis were simulated. They are presented in ascending order from platykurtic to leptokurtic. The first population was extremely platykurtic (XPLT) and continuous with skewness (γ_1) equal to zero and kurtosis (γ_2) equal to -1.80. The second population was also platykurtic (PLAT) and continuous with skewness (γ_1) equal to zero and kurtosis (γ_2) equal to -1.00. It was chosen because it has been used in a variety of other simulation studies (e.g., Olejnik & Algina, 1987, 1988). The third population was slightly platykurtic (SPLT) with γ_1 = 0.0 and γ_2 = -0.50. It was selected as a continuous distribution which closely matches the moments of one of Micceri's (1989) data sets. The fourth population was the normal distribution (NORM) generated with the SAS RANNOR function. The fifth population (LEP1) was selected as a slightly leptokurtic, γ_2 = +1.00, continuous distribution with no skew comparable to the second population (PLAT). The sixth (LEP3) and seventh (XLEP) were selected as highly leptokurtic (γ_2 = +3.00 and +3.75, respectively) with no expected skewness.

Group Size and Variance Ratio Parameters. Equal sample sizes of n_j = 10, 13, and 20 and unequal sample sizes of (10, 20) and (13, 20) were employed. To investigate power, variance ratios of VR = 2.0 and 5.0 were imposed by taking the population from which Group Two was sampled and multiplying it by constant equal to the square root of VR.

Because Olejnik and Algina (1988) found that tests of variance were differentially powerful depending on the relationship between group size and population variance, all conditions were crossed when power was investigated. For example, when the variance ratio was VR = 2.0 and Group One, with n_1 = 13, was sampled from the normal distribution, while Group Two (n_j = 20) was sampled from a platykurtic distribution, an inverse relationship between group size and population variance (negative condition) was imposed. In order to create a positive condition, the sample sizes were reversed so that the larger group had the larger variance. Also because power and robustness may depend on population shape, the seven populations were systematically manipulated as long as conditions did not duplicate (e.g., when investigating Type I error rate for equal sample sizes all population combinations are not necessary). Table 1 shows the sample size conditions for the analyses in this study. Note that two sample size configurations were added to impose positive and negative sample size/variance correlations for investigating power in True Experiments. For Quasi-Experiments, all possible sample size conditions were used since the Type I error rate has been shown to depend on sample size/kurtosis configurations. In examining power, variance constants were imposed on both Group One and Group Two because power has also been shown to depend on the configuration of sample size, kurtosis, and variance (Beasley & O'Connor, 1995; Olejnik & Algina, 1988).

Procedure

The second through seventh populations were generated separately for each group using the RANNOR function is SAS/IML, which provides a clock generated pseudorandom standard normal deviate, z_{ij} (SAS Institute, 1990). Fleishman's (1978) method was used to transform these distributions into non-normal data with specified mean, variance, skewness, and kurtosis values via a polynomial equation of the form,

$$y_{ij} = a + bz_{ij} + cz_{ij}^2 + dz_{ij}^3 \quad . \qquad (6)$$

Since the minimum kurtosis derived by Fleishman is γ_2 = -1.00, the first population (XPLT) was simulated by combining three uniform distributions that varied in central location. A small distribution of 20 cases that centered around 0 and two larger distributions of 990 cases each which centered around -0.75 and 0.75 were concatenated to create this heavy-tailed distribution. Linear transformations were used in order to have the expected variances used in this study. During the simulation procedures, observations were randomly sampled from these distributions during each replication. For each condition elaborated, 5,000 replications were completed. The proportions of rejections at the α = .05 level of significance were used as measures of empirical power and Type I error rate.

Since 5,000 replications were conducted in each condition with α = .05, the standard error is .0031. Thus, any Type I error rate of .0562 or greater exceeded two standard errors and was considered a significant inflation of the Type I error rate. Other less stringent criteria include upper limits of .06 (Cochran, 1954) and .075 (Bradley, 1978). In order to avoid the problems with making multiple comparisons within this study, the standard error of simulation was used as a general heuristic rather than as a statistical test when comparing empirical power estimates. Furthermore, if the

empirical Type I error rate of a test exceeded the nominal alpha by two standard errors, its power was interpreted cautiously. If its Type I error rate exceeded Cochran's limit of .06, its power estimate was not reported.

Results

Simulation Accuracy

When multiplied as in (6), the resulting mean, variance, skewness, and kurtosis of y_j approximate the characteristics of the distribution of interest. It should be noted, however, that the simulated data are not governed by a known mathematical function. Rather, the simulated data represent a distribution with the same skewness and kurtosis as the desired distribution. Table 2 demonstrates the adequacy of the Fleishman simulation method in this study. Values for the mean (μ), variance (σ^2), skew (γ_1) and kurtosis (γ_2) for each group of n_j were taken across 15,000 replications for n_j = 10 and 13 and across 30,000 replications for n_j =20. For all seven populations, μ, σ^2, and, γ_1 were adequately simulated. Furthermore, kurtosis (γ_2) was reasonably simulated for platy- and mesokurtic distributions, especially with n_j =20. For leptokurtic distributions, however, the kurtosis of the group was drastically underestimated which is most likely due to the small sample sizes used.

True Experiments

Type I Error. Table 3 shows the empirical Type I error rate for the seven sampled populations under the conditions of a True Experiment (i.e., both groups drawn from the same population). As would be expected the Hartley's *Fmax* test showed a conservative rejection rate with platykurtic populations (e.g., XPLT, PLAT, & SPLT) but more importantly was liberal when the data were sampled from leptokurtic populations. Furthermore, when sample sizes are unequal, the suspension of the *Fmax* test is often suggested. However, the Type I error rate remained under the nominal alpha of .05 even with unequal samples under the meso- and platykurtic conditions. All other tests, except for WOB which exhibited minor inflation with disparate sample sizes, held the Type I error rate under the nominal alpha.

Power. Tables 4 and 5 show representative results from the comparative power analysis of True Experiments with different populations and variance ratios of *VR* = 2.0 and 5.0, respectively. For all tests, except *Fmax*, it can be seen that heavy-tailed distributions presented a more powerful situation when testing variance heterogeneity. For example, the empirical power estimates were higher when data were sampled from the PLAT population as compared to normally distributed data. Also, higher power estimates were yielded when data were sampled from the normal distribution as compared to the leptokurtic populations(e.g., LEP1 and XLEP, see Tables 4 & 5).

Under the conditions of a normally distributed population, the Hartley's *Fmax* was robust and demonstrated superior power, except when there was a positive relationship between sample size and group variance. In this case, the WOB was most powerful. When the sample size/variance correlation was negative *Fmax* and OB were of similar power.

When data were sampled from the PLAT distribution, the OB transformation and Ramsey's OB_{BF} were the clear choices in low power situations (see Table 4). However, with a variance ratio of *VR* = 5.0, the *Fmax* test was more powerful except when the sample size/variance correlation was positive. Thus, in cases where the smaller group had the smaller variance, WOB was most powerful namely because neither *Fmax* nor the Ramsey's procedures make provisions for such situations. In high power situations (*VR* = 5.0) whether the data were meso- or platykurtic, the *Fmax* was more powerful. However, both *Fmax* and OB are very likely to reject the null hypothesis in such cases. Thus, if the data sampled are platykurtic, one should consider the O'Brien transformation in low power situations. In all conditions with meso- and platykurtic data and a positive relationship between sample size and variance, performing the Welch statistic on O'Brien scores was the most powerful procedure.

When data were sampled from leptokurtic populations, the *Fmax* test was disqualified because it inflated the Type I error rate (see Table 3). Of the remaining tests, BF and BF_{OB} had similar empirical power estimates with small (n_j = 10) equally sized samples. Similarly, as was observed with platykurtic distribution, BF_{OB} was more powerful than BF, which indicates that the Ramsey conditional procedures can provide more power. When sample sizes were unequal and positively related to the group variances, the WBF was more powerful. This finding seems consistent with previous research but has yet to be reported in the literature. When a negative relationship between sample sizes and group variances existed, OB was more powerful regardless of the leptokurtosis of the sampled population. Increasing the group Variance Ratio to *VR* = 5.0 magnified these findings. However, under these more powerful conditions, the power estimates of BF were more competitive and actually exceeded those of OB when there was an inverse sample size/variance relationship. For example in Table 5, when n_1 = 20, σ_1^2 = 1.0, n_2 = 13, and σ_2^2 = 5.0, the power of BF, .5590 was much higher than that of OB, .5110. This indicates that OB is only more powerful under negative sample size/variance conditions in low power situations (i.e., small sample sizes, small differences in variance). That is, if samples are rather large and drawn from leptokurtic populations, the BF and WBF may be better

choices for testing variances. Thus, consistent with previous research, when there is a negative correlation between sample sizes and variances, the advantage in power of OB over BF seems to dissipate with increasing (a) sample sizes for both groups, (b) variance for the smaller group, and/or (c) kurtosis of the sampled population (Olejnik & Algina, 1988; Ramsey, 1994).

Quasi-Experiments

Type I Error. When one group was sampled from a population with extremely negative kurtosis (XPLT), while the second group was sampled from population of varying shapes, the Type I error rates for all tests, except for *Fmax*, were unstable and generally above the nominal alpha of .05 (see Table 6). However, as the extremity of platykurtosis declined, the Type I error rates became more stable for most tests (see Table 7).

When the two groups were sampled from populations with similar positive kurtosis, the results were predictable from the Type I Error results for True Experiments. Table 8 shows that when both groups had positive kurtosis most tests, except for *Fmax*, held the Type I error rate at the nominal alpha of .05. However, WOB showed inflations when the larger group was sampled from a less leptokurtic distribution. These results extended to situations where one group is sampled from a slightly leptokurtic distribution (LEP1) and the other is sampled from a slightly platykurtic distribution (SPLT).

Although the mixture of LEP1 and the normal distribution did not affect the Type I error rate of most tests (see Table 8), when the variance of a normally distributed sample was tested against the variance of data sampled from more leptokurtic populations (e.g., LEP3, XLEP), the Type I Error rate of all tests were affected when sample sizes were unequal (see Table 7). When the normally distributed data were compared to samples from platykurtic populations (PLAT), the Type I error rate was controlled for all tests with equal sample sizes. When sample sizes were not equal, only *Fmax* and BF were consistently robust to these violations to the normality assumption. When the larger group was more platykurtic, OB, WOB, and OB_{BF} tended to inflate the Type I error rate (see Table 7). Thus it would appear that if data are sampled from different populations with similar kurtosis, keeping group sizes approximately equal would be a reasonable step in controlling the Type I error rate.

In some situations where the kurtosis of the sampled distributions differed in sign, the Type I error rate of *Fmax* remained under the nominal alpha of .05. However, when the disparity in kurtosis increased this was not the case. For example, in comparing the variance of data sampled form the extremely platykurtic population (XPLT, γ_2 = -1.80) to the variance of samples from highly leptokurtic distributions (LEP3, γ_2 = 3.00), no test was robust (see Table 6). Thus, it appears that if the kurtosis of distributions differ in sign to the same absolute degree, then the *Fmax* test of variance is robust. This supposition was confirmed in an *ad-hoc* simulation in which the variance of data sampled from the XPLT distribution was tested against the variance of two leptokurtic distributions with population kurtosis values of γ_2 = 1.75 and 2.00. When comparing these variances under the null hypothesis, the Type I error rate of *Fmax* remained under the nominal alpha of .05 while all other tests were not robust.

Power. It should be noted that since Type I error rates for these tests of variance were dependent on the sample size and population kurtosis configuration, power was also dependent on combinations of sample size, population kurtosis, and group variance. In general, when the group with the larger variance was sampled from the heavier-tailed distribution there was more power for the tests of variance. When the more leptokurtic distribution was more variant, a reduction in power was observed. Therefore, results comparing the power of these tests are reported for both situations.

In quasi-experimental situations in which one group was sampled from the extremely platykurtic population, only *Fmax* controlled the Type I error. Therefore, only *Fmax* can be validly used for testing variances when only one group is sampled from an extremely platykurtic population. As this negative kurtosis increased in value and became less extreme, more comparisons were possible.

When one group was sampled from the platykurtic population (PLAT, γ_2 = -1.00) while the other group was normally distributed, all tests of variance held the Type I error rate for equal sample sizes and are comparable (see Table 7). Table 9 shows that under these conditions, OB and OB_{BF} were the most powerful. With unequal sample sizes, OB, WOB, and OB_{BF}, tended to inflate the Type I error rate, and therefore, *Fmax* and BF seem to be the most dependable tests. Furthermore, when there was a positive sample size/variance correlation, WBF was robust and more powerful as long as the disparity in sample sizes was not extreme. With an inverse sample size/variance relationship, *Fmax* is robust and adequately powerful. However, one may consider that OB, WOB, and OB_{BF} only inflated the Type I error rate when the more platykurtic group was larger in size. Thus, under conditions where the sample sizes are equal or the smaller group is more platykurtic, OB and OB_{BF} were more powerful except when the larger (more leptokurtic) sample had the larger variance, in which case, WOB was more powerful.

As with the extremely platykurtic population, the Type I error rate was controlled by *Fmax* when the platykurtic distribution (PLAT, γ_2 = -1.00) was compared to a group sampled from a population with an equal degree of leptokurtosis (LEP1, γ_2 = 1.00); however, no other test was robust (see Table 8). Thus

for a test of variance to be valid when one group is platykurtic, the other group must be either (a) similarly platykurtic, (b) symmetric, or (c) leptokurtic to the same degree. If the sampled distributions are similarly platykurtic, OB or WOB are preferred. If a second group is symmetric in shape then overall, *Fmax* is adequate; however, if the platykurtic distribution has more variance, OB, WOB, BF, WBF may be considered. If the kurtosis of groups differ in sign to the same degree, only *Fmax* is adequate.

Table 9 also shows that when normally distributed scores were compared to data sampled from a leptokurtic distribution with $\gamma_2 = 1.00$, all tests of variance that did not violate the Type I error rate were similarly powerful. Since OB and BF exhibited similar power, one of the conditional procedures may be used to decide which test to perform. That is, BF_{OB} or OB_{BF} can provide more power (Ramsey, 1994). With a positive sample size/variance correlation, the Welch procedures (WOB and WBF) showed more power relative to the other tests. With a negative sample size/variance correlation, OB remained the test of choice. Similar findings extend to situations where one group was leptokurtic ($\gamma_2 = 1.00$) and the other was slightly platykurtic ($\gamma_2 = -0.50$; see Table 10). However, in this situation the Welch procedures were more likely to inflate the Type I error rate, and BF should be considered when the sample size-variance correlation is positive.

As was the case when samples were selected from the same leptokurtic distribution, sampling from different leptokurtic populations demonstrated the superiority of the BF procedure and its variants. For example, Table 10 shows comparative power estimates for the tests of variance when one group was sampled from an extremely leptokurtic population (LEP3, $\gamma_2 = 3.00$) while the other group was less leptokurtic ($\gamma_2 = 1.00$). With equal sample sizes, BF and BF_{OB} were more powerful. As was the case in True Experiments, the high power of BF_{OB} relative to BF indicates the effectiveness of testing kurtosis before applying a test of variance (Ramsey, 1994). When the sample size/variance correlation is positive, WBF was clearly the most powerful procedure, while OB was more powerful with a negative relationship. As with the results for True Experiments, the advantage of OB with an inverse sample size/variance relationship dissipated in high power situations (*VR* =5.0, results not shown).

Discussion

Summary

The results demonstrated that when data were sampled from the same population and randomly assigned (i.e., True Experiments) to equally sized groups, Hartley's *Fmax* test was only robust when the population kurtosis was near or below zero. This confirms the findings of many other studies and establishes the need for analytic alternatives for testing variances when data are nonnormal. When data had a negative kurtosis, the O'Brien (1981) transformation was generally the best choice, while the Brown & Forsythe (1974) transformation was robust and showed superior power for testing variances in leptokurtic data. Also consistent with previous studies, the O'Brien test was generally more powerful when sample sizes and variances were negatively correlated, regardless of the shape of the distribution (Algina et al., 1989; Olejnik & Algina, 1988; Ramsey, 1994). The Welch procedure performed on O'Brien scores was more powerful when the sample size/variance correlation was positive in platykurtic samples (Algina et al., 1989).

Furthermore, under the conditions of a positive sample size/variance correlation in leptokurtic samples, the Welch test applied to Brown-Forsythe scores was robust and demonstrated superior power. Although this finding seems reasonable given previous research, it had yet to be empirically confirmed until this study. The results also demonstrated that choosing a test of variance based on an initial test of kurtosis can increase power (Ramsey, 1994); however, the power of these conditional tests has been shown to be dependent on the power of the test of kurtosis (Beasley & O'Connor, 1995). Thus, if tests of kurtosis are to be used to determine the most powerful and appropriate test of variance to perform, one must be concerned with the power of both tests.

This study also presented many new findings about the statistical properties of testing variances when groups were not sampled from the same population (i.e., Quasi-Experiments). When both groups were sampled from similarly platykurtic or similarly leptokurtic distributions, the results were predictable from the results of True Experiments. However, when one group was extremely platykurtic, only the *Fmax* tests controlled the Type I error rate. Furthermore, if the kurtosis of the groups differed in sign to the same absolute degree, the *Fmax* test was robust.

When the variance of normally distributed data were tested against the variance of data sampled from a platykurtic population, the Type I error rate of many tests were less stable which in turn affected the validity of power estimates and recommendations for use. Most notably, when the larger group was normally distributed with a larger variance and the smaller, platykurtic group was less variable, WOB was robust and more powerful. *Fmax* was preferable when sample sizes were equal or negatively correlated to variances. However, when the larger group was platykurtic, the Type I error rates of OB and WOB were inflated. Thus, when the more platykurtic group had a larger variance, the OB exhibited more power when the sample sizes were equal or inversely related to variance (platykurtic group was smaller in size). For a positive sample size/variance correlation (i.e., larger platykurtic group had larger variance) only BF was robust and adequately powerful.

When the variance of normally distributed data was tested against the variance of data sampled from leptokurtic populations, the Type I error rate of *Fmax* is extremely inflated and the BF is preferred when sample sizes are equal; however, OB showed similar power. When the sample size/variance correlation was positive the Welch test applied to BF scores is generally more powerful, while OB was more powerful when the smaller group had a larger variance despite the leptokurtic shape of one group.

Recommendations

Educational researchers are typically interested in estimating change and differences. However, simply examining shifts in central location does not fully address these issues, all distributional differences should be investigated. Thus testing all moments in the distribution is recommended when comparing groups whether they are intact or randomly constructed. Not only does this approach test the major assumptions for the ANOVA, but it also investigates the issue of whether a treatment condition affected the shape or response variability of a distribution of scores in True Experiments. In this case, tests such as the Kolmogorov-Smirnov test may be used to answer the question "Did the treatment affect the distribution of scores?" If intact groups are compared in central location or if differences in scale of the dependent variable are of interest, a test of variance is needed. The results demonstrate that the shape of the distributions should be examined before choosing a test of variance.

Table 11 shows a summary of these recommendations based on the kurtosis of the distributions and whether the sample sizes are equal, positively correlated, or negatively correlated with the variances. Entries on the diagonal exhibit recommendations for data sampled from the same (i.e., True Experiments) or similar populations. Off-diagonal entries reveal the recommendations for Quasi-Experiments and field research. Since the conditional tests examined are used to select one of these tests of variance (i.e., OB and BF), they are not represented. Furthermore, conditionally choosing the most powerful test based on sample characteristics may capitalize on chance differences in the data and inflate the Type I error rate.

In evaluating the recommendations in Table 11, one should consider that educational data tends to be platykurtic in nature (Micceri, 1989). It should also be noted that the recommendations for situations where the groups are either both leptokurtic or both platykurtic extend to most values of kurtosis. However, one should be aware that for situations in which one group is leptokurtic and the other is platykurtic the recommendations in Table 11 apply only if the kurtosis is of similar absolute value. Thus it is suggested that all relevant tests of variance be performed and agreement among the results assessed. If all tests reject the null hypothesis under conditions in which the Type I error rate is controlled, then the statistical significance is likely to represent a valid result. If there is disagreement among tests, then the consistency of disagreements with empirical findings should be assessed. For example, if equally sized, platykurtic samples are tested for variance heterogeneity and only the O'Brien test rejects the null hypothesis, there is indication of statistical significance because the O'Brien test is robust and most powerful in this situation (see Table 11).

Although this investigation was limited to the two-sample tests, it is believed that these results extend to most multi-group situations. For True Experiments, other studies have shown this to be the case (e.g., Miller, 1968). For Quasi-Experiment recommendations, one should consider the several factors. If the kurtosis values for all groups indicate similar positive or similar negative kurtosis, then the recommendations for leptokurtic and platykurtic groups in Table 11 should be valid. Also if about half of the groups are mesokurtic while the other half are either lepto- or platykurtic, then Table 11 can be used. If the groups are mostly leptokurtic, using the leptokurtic recommendations is advised; however, if the groups are mostly platykurtic, recommendations are more difficult to make. If the groups have drastically different shapes, the results indicated that *Fmax* was the preferred test in the two group situation, but one must consider that the *Fmax* only uses the data of two groups. Thus, if multiple groups are present and the groups with the largest and smallest variances (the values used for *Fmax*) have kurtosis estimates of opposite signs, *Fmax* may be allowable as long as the kurtosis values have approximately the same absolute value.

References

Algina, J., Olejnik, S. F., & Ocanto, R. (1989). Error rates and power estimates forselected two-sample tests of scale. *Journal of Educational Statistics, 14,* 373-384.

Beasley, T. M., & O'Connor, S. A. (1995, April). Testing heterogeneous variances in true and quasi-experiments. Paper presented at the meeting of the American Educational Research Association. San Francisco, CA.

Box, G. E. P. (1953). Non-normality and tests on variances. *Biometrika, 40,* 318-335.

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

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

Cochran, W. G. (1954). Some methods for strengthening the common χ^2 tests. *Biometrics, 10*, 417-451.

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.

Durrand, A. L. (1969). *Comparative power of various tests of homogeneity of variance.* Unpublished Master's Thesis, University of Colorado, Boulder, Colorado.

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

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

Klotz, J. (1962). Nonparametric tests for scale. *Annals of Mathematical Statistics, 33*, 495-512.

Levene, H. (1960). Robust tests for equality of variances. In I. Olkin (Ed.), *Contributions to probability and statistics* (pp. 278-292). Palo Alto, CA: Stanford University Press.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*, 156-166.

Miller, R. G., Jr. (1968). Jackknifing variances. *Annals of Mathematical Statistics, 39*, 567-582.

Moses. L. E. (1963). Rank tests of dispersion. *Annals of Mathematical Statistics, 34*, 973-983.

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

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

Olejnik, S. J., & Algina, J. (1985, April). *Power analysis of selected parametric and nonparametric test for heterogeneous variances in non-normal distributions.* Paper presented at the meeting of the American Educational Research Association. Chicago, IL.

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

Olejnik, S. J., & Algina, J. (1988). Test of variance equality when distributions differ in form and location. *Educational and Psychological Measurement, 48*, 317-329.

Pearson, E. S., & Please, N. W. (1975). Relation between the shape of population distribution and the robustness of four simple test statistics. *Biometrika, 62*, 223-242.

Penfield, D. A., & Koffler, S. (1985, April). *A power study of selected nonparametric k-sample tests.* Paper presented at the meeting of the American Educational Research Association. Chicago, IL.

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

Ramsey, P. H., & Brailsford, E. A. (1990). Robustness and power of tests of variability on two independent groups. *British Journal of Mathematical and Statistical Psychology, 43*, 113-130.

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

SAS Institute. (1990). *SAS/IML user's guide* (Release 6.04). Cary, NC: Author.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics, 2*, 110-114.

Scheffe', H. (1959). *The analysis of variance.* New York: Wiley.

Siegel, S., & Tukey, J. W. (1960). A nonparametric sum of ranks procedure for relative spread in unpaired samples. *American Statistical Association Journal, 55*, 429-445.

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

Table 1. Summary of conditions analyzed for Sample Size and Population configurations

	True Experiment		Quasi-Experiment		
	Type I	Power	Type I	Power	Power
(n_1, n_2)	Error	$(\sigma_1^2 < \sigma_2^2)$	Error	$(\sigma_1^2 < \sigma_2^2)$	$(\sigma_1^2 > \sigma_2^2)$
(10, 10)	*	*	*	*	*
(13, 13)	*	*	*	*	*
(20, 20)	*	*	*	*	*
(10, 20)	*	+	*	+	-
(13, 20)	*	+	*	+	-
(20, 10)	U	-	*	-	+
(20, 13)	U	-	*	-	+

Note. * indicates the analysis was completed. U indicates the analysis was unnecessary and not completed. - indicates a negative relationship between sample size and variance. + indicates a positive relationship between sample size and variance.

Table 2. Average population parameters across Type I error simulations.

	Population Parameter			
Population	μ	σ^2	γ_1	γ_2
1. **XPLT** $E(\gamma_2) = -1.80$				
$n = 10$	+0.003937	1.009481	-0.001320	-0.964967
$n = 13$	+0.003508	1.011025	-0.006166	-1.205637
$n = 20$	-0.003517	1.006298	-0.010828	-1.447347
2. **PLAT** $E(\gamma_2) = -1.00$				
$n = 10$	+0.005843	1.007704	-0.012320	-0.421073
$n = 13$	+0.005508	1.009243	-0.008166	-0.541785
$n = 20$	+0.013517	1.011201	-0.030828	-0.686890
3. **SPLT** $E(\gamma_2) = -0.50$				
$n = 10$	+0.000090	1.017263	+0.039203	-0.172836
$n = 13$	+0.001756	1.015814	+0.046087	-0.233373
$n = 20$	-0.004001	1.012889	+0.058859	-0.302452
4. **NORM** $E(\gamma_2) = 0.00$				
$n = 10$	-0.002803	1.006276	-0.027654	-0.003053
$n = 13$	-0.001429	1.001976	-0.034611	-0.009850
$n = 20$	+0.000057	1.003101	-0.028943	-0.010843
5. **LEP1** $E(\gamma_2) = +1.00$				
$n = 10$	-0.000713	1.024512	+0.061602	+0.261224
$n = 13$	+0.058517	1.022624	+0.065461	+0.339038
$n = 20$	+0.023376	1.027412	+0.052777	+0.493708
6. **LEP3** $E(\gamma_2) = +3.00$				
$n = 10$	+0.011568	1.028524	+0.032652	+0.553391
$n = 13$	+0.008883	1.012206	+0.035397	+0.765899
$n = 20$	+0.005923	1.020455	+0.213038	+1.218371
7. **XLEP** $E(\gamma_2) = +3.75$				
$n = 10$	+0.010142	1.013801	+0.040116	+0.662791
$n = 13$	+0.005774	1.028358	+0.017173	+0.904860
$n = 20$	+0.004539	0.998350	+0.017142	+1.357478

Table 3. Empirical Type I Error Rate for seven procedures in True Experiments with no differences in central location.

Pop.	n_1,n_2	*Fmax*	OB	BF	WOB	WBF	OB_{BF}	BF_{OB}
1. **XPLT** (γ_2 = -1.80)								
	10, 10	.0006	.0248	.0338	.0242	.0292	.0248	.0284
	13, 13	.0018	.0272	.0058	.0262	.0054	.0272	.0272
	20, 20	.0006	.0262	.0186	.0256	.0172	.0262	.0272
	10, 20	.0030	.0372	.0366	.0288	.0316	.0372	.0380
	13, 20	.0026	.0296	.0232	.0288	.0114	.0294	.0308
2. **PLAT** (γ_2 = -1.00)								
	10, 10	.0152	.0438	.0344	.0380	.0324	.0444	.0392
	13, 13	.0142	.0464	.0256	.0430	.0248	.0468	.0400
	20, 20	.0082	.0474	.0386	.0456	.0378	.0476	.0444
	10, 20	.0124	.0494	.0384	.0580*	.0436	.0498	.0474
	13, 20	.0110	.0464	.0326	.0510	.0316	.0464	.0424
3. **SPLT** (γ_2 = -0.50)								
	10, 10	.0354	.0396	.0330	.0320	.0312	.0400	.0362
	13, 13	.0340	.0416	.0306	.0366	.0294	.0422	.0346
	20, 20	.0306	.0486	.0436	.0458	.0432	.0498	.0456
	10, 20	.0368	.0498	.0458	.0604*	.0526	.0518	.0494
	13, 20	.0332	.0458	.0382	.0472	.0414	.0468	.0406
4. **NORM** (γ_2 = 0.00)								
	10, 10	.0484	.0336	.0402	.0258	.0370	.0352	.0414
	13, 13	.0528	.0384	.0312	.0320	.0274	.0402	.0346
	20, 20	.0460	.0416	.0366	.0392	.0358	.0428	.0368
	10, 20	.0462	.0364	.0392	.0530	.0492	.0380	.0400
	13, 20	.0478	.0422	.0344	.0442	.0356	.0440	.0380
5. **LEP1** (γ_2 = 1.00)								
	10, 10	.0648*	.0352	.0404	.0260	.0368	.0366	.0412
	13, 13	.0648*	.0362	.0324	.0316	.0298	.0376	.0358
	20, 20	.0678*	.0380	.0398	.0352	.0382	.0412	.0388
	10, 20	.0624*	.0368	.0366	.0460	.0456	.0388	.0370
	13, 20	.0664*	.0408	.0328	.0396	.0348	.0416	.0362
6. **LEP3** (γ_2 = 3.00)								
	10, 10	.1466*	.0278	.0328	.0188	.0286	.0294	.0338
	13, 13	.1538*	.0332	.0346	.0278	.0316	.0364	.0360
	20, 20	.1766*	.0292	.0364	.0256	.0348	.0362	.0368
	10, 20	.1476*	.0366	.0358	.0438	.0494	.0420	.0368
	13, 20	.1567*	.0316	.0384	.0302	.0396	.0368	.0392
7. **XLEP** (γ_2 = 3.75)								
	10, 10	.1400*	.0240	.0340	.0168	.0278	.0266	.0348
	13, 13	.1496*	.0294	.0338	.0234	.0298	.0326	.0346
	20, 20	.1752*	.0346	0412	.0300	.0380	.0438	.0418
	10, 20	.1502*	.0382	.0372	.0448	.0506	.0416	.0390
	13, 20	.1662*	.0340	.0370	.0344	.0412	.0388	.0384

Note. * indicates the Type I error rate exceeded .0562 and is 2 standard errors above the nominal alpha of α = .05.

Table 4. Empirical Power for seven procedures in True Experiments with no differences in central location and σ_1^2 = 1.0 and σ_2^2 = 2.0.

Pop.	n_1,n_2	*Fmax*	OB	BF	WOB	WBF	OB_{BF}	BF_{OB}
1. **XPLT** (γ_2 = -1.80)								
	10, 10	.0228	.4566	.1842	.4424	.1742	.4558	.4502
	13, 13	.0336	.6310	.0444	.6168	.0420	.6296	.6196
	20, 20	.0856	.8662	.2306	.8608	.2278	.8662	.8630
Pos.	10, 20	.0282	.6502	.2004	.7602	.2024	.6494	.6470
Pos.	13, 20	.0392	.7640	.2044	.8330	.1952	.7642	.7620
Neg.	20, 10	.0546	.7038	.1846	.5224	.2192	.7036	.6964
Neg.	20, 13	.0686	.7544	.0600	.6422	.0302	.7542	.7458
2. **PLAT** (γ_2 = -1.00)								
	10, 10	.0984	.1412	.1116	.1188	.1050	.1424	.1262
	13, 13	.1364	.2206	.1348	.2020	.1300	.2204	.1648
	20, 20	.2680	.4086	.2950	.3950	.2910	.4084	.3392
Pos.	10, 20	.1230	.1600	.1716	.3420*	.2480	.1620	.1804
Pos.	13, 20	.1530	.2424	.2144	.3612	.2668	.2442	.2384
Neg.	20, 10	.1898	.3176	.1740	.1184*	.0918	.3176	.2142
Neg.	20, 13	.2150	.3390	.1912	.1952	.1264	.3392	.2390
4. **NORM** (γ_2 = 0.00)								
	10, 10	.1508	.0994	.1070	.0782	.0978	.1030	.1130
	13, 13	.2022	.1454	.1298	.1262	.1226	.1484	.1368
	20, 20	.2850	.2348	.2182	.2190	.2122	.2412	.2250
Pos.	10, 20	.1620	.0750	.1230	.2502	.2172	.0806	.1244
Pos.	13, 20	.2116	.1330	.1732	.2416	.2282	.1406	.1758
Neg.	20, 10	.2530	.2452	.1616	.0650	.0766	.2454	.1734
Neg.	20, 13	.2572	.2382	.1654	.1072	.0952	.2396	.1752
5. **LEP1** (γ_2 = 1.00)								
	10, 10	---	.0686	.0882	.0522	.0782	.0728	.0904
	13, 13	---	.1040	.1114	.0870	.1044	.1118	.1148
	20, 20	---	.1610	.1864	.1470	.1812	.1850	.1888
Pos.	10, 20	---	.0462	.1016	.1934	.2158	.0562	.1002
Pos.	13, 20	---	.0766	.1372	.1710	.1976	.0952	.1376
Neg.	20, 10	---	.1912	.1528	.0400	.0630	.1982	.1580
Neg.	20, 13	---	.1798	.1544	.0668	.0834	.1904	.1576
7. **XLEP** (γ_2 = 3.75)								
	10, 10	---	.0602	.0830	.0444	.0724	.0672	.0848
	13, 13	---	.0830	.1010	.0654	.0900	.0910	.1020
	20, 20	---	.1438	.1776	.1322	.1712	.1764	.1792
Pos.	10, 20	---	.0302	.0916	.1680	.2020	.0438	.0910
Pos.	13, 20	---	.0746	.1376	.1568	.2034	.0984	.1384
Neg.	20, 10	---	.1620	.1336	.0280	.0492	.1688	.1366
Neg.	20, 13	---	.1520	.1386	.0536	.0786	.1668	.1420

Note. * indicates that the Type I error rate exceeded the nominal alpha by 2 standard errors and should be interpreted cautiously. Blank entries indicate that the Type I error rate exceeded Cochran's limit of .06. Pos. indicates a positive correlation between sample size and variance; Neg. indicates an inverse sample size-variance relationship.

Table 5. Empirical Power for seven procedures in True Experiments with no differences in central location and σ_1^2 = 1.0 and σ_2^2 = 5.0.

Pop.	n_1,n_2	*Fmax*	OB	BF	WOB	WBF	OB_{BF}	BF_{OB}
1. **XPLT** (γ_2 = -1.80)								
	10, 10	.6612	.5450	.4874	.4762	.4532	.5476	.5134
	13, 13	.8402	.7594	.6582	.7186	.6332	.7600	.6974
	20, 20	.9724	.9590	.9104	.9526	.9062	.9592	.9406
Pos.	10, 20	.7794	.6856	.7010	.9348	.8618	.6878	.7170
Pos.	13, 20	.9040	.8550	.8138	.9452	.8858	.8554	.8352
Neg.	20, 10	.8782	.8684	.7238	.4664	.4442	.8688	.7788
Neg.	20, 13	.9278	.9158	.8086	.7322	.6380	.9160	.8544
2. **PLAT** (γ_2 = -1.00)								
	10, 10	.6612	.5450	.4874	.4762	.4532	.5476	.5134
	13, 13	.8402	.7594	.6582	.7186	.6332	.7600	.6974
	20, 20	.9724	.9590	.9104	.9526	.9062	.9592	.9406
Pos.	10, 20	.7794	.6856	.7010	.9348*	.8618	.6878	.7170
Pos.	13, 20	.9040	.8550	.8138	.9452	.8858	.8554	.8352
Neg.	20, 10	.8782	.8684	.7238	.4664*	.4442	.8688	.7788
Neg.	20, 13	.9278	.9158	.8086	.7322	.6380	.9160	.8544
4. **NORM** (γ_2 = 0.00)								
	10, 10	.6396	.3616	.4136	.2876	.3772	.3698	.4200
	13, 13	.7690	.5528	.5646	.4924	.5392	.5630	.5726
	20, 20	.9350	.8428	.8330	.8198	.8266	.8594	.8396
Pos.	10, 20	.7264	.3800	.5680	.7796	.7604	.3964	.5654
Pos.	13, 20	.8388	.5860	.6978	.7894	.8008	.6052	.6992
Neg.	20, 10	.8148	.7458	.6470	.2860	.3664	.7538	.6584
Neg.	20, 13	.8648	.7880	.7084	.4954	.5342	.7986	.7224
5. **LEP1** (γ_2 = 1.00)								
	10, 10	---	.2600	.3432	.1984	.3086	.2762	.3446
	13, 13	---	.4088	.4924	.3574	.4640	.4446	.4944
	20, 20	---	.6656	.7536	.6384	.7438	.7454	.7550
Pos.	10, 20	---	.2186	.4674	.6032	.6792	.2690	.4628
Pos.	13, 20	---	.3838	.5948	.6250	.7142	.4524	.5934
Neg.	20, 10	---	.6160	.5600	.1838	.2890	.6372	.5650
Neg.	20, 13	---	.6400	.6362	.3452	.4530	.6852	.6406
6. **LEP3** (γ_2 = 3.00)								
	10, 10	---	.2062	.2990	1590	.2636	.2268	.2986
	13, 13	---	.2948	.4090	.2490	.3866	.3502	.4094
	20, 20	---	.4962	.6662	.4616	.6532	.6556	.6676
Pos.	10, 20	---	.1400	.3666	.4794	.6086	.2082	.3598
Pos.	13, 20	---	.2610	.4902	.4750	.6274	.3678	.4878
Neg.	20, 10	---	.5262	.5166	.1256	.2366	.5696	.5184
Neg.	20, 13	---	.5110	.5590	.2358	.3732	.5888	.5606

Note. * indicates that the Type I error rate exceeded the nominal alpha by 2 standard errors and should be interpreted cautiously. Blank entries indicate that the Type I error rate exceeded Cochran's limit of .06. Pos. indicates a positive correlation between sample size and variance; Neg. indicates an inverse sample size-variance relationship.

Table 6. Empirical Type I Error Rates for seven procedures in Quasi-Experiments with no differences in central location and Group One is platykurtic $\gamma_2 = -1.80$.

Group 2 Pop.	n_1,n_2	*Fmax*	OB	BF	WOB	WBF	OB_{BF}	BF_{OB}
2. **PLAT** ($\gamma_2 = -1.00$)								
	10, 10	.0092	.0620*	.0484	.0584*	.0460	.0640*	.0632*
	13, 13	.0068	.0642*	.0114	.0616*	.0110	.0654*	.0634*
	20, 20	.0022	.0542	.0476	.0524	.0460	.0550	.0550
	10, 20	.0048	.0282	.0408	.0402	.0658*	.0348	.0364
	13, 20	.0044	.0362	.0160	.0454	.0126	.0388	.0408
	20, 10	.0062	.1178*	.0532	.0808*	.0460	.1180*	.1178*
	20, 13	.0054	.0840*	.0494	.0600*	.0418	.0842*	.0844*
3. **SPLT** ($\gamma_2 = -0.50$)								
	10, 10	.0148	.0734*	.0600*	.0694*	.0568*	.0752*	.0738*
	13, 13	.0116	.0750*	.0160	.0742*	.0150	.0766*	.0742*
	20, 20	.0084	.0674*	.0582*	.0656*	.0566*	.0678*	.0680*
	10, 20	.0082	.0358	.0458	.0462	.0682*	.0420	.0452
	13, 20	.0068	.0468	.0162	.0560	.0122	.0492	.0492
	20, 10	.0096	.1238*	.0660*	.0936*	.0580*	.1240*	.1230*
	20, 13	.0108	.0992*	.0560	.0782*	.0500	.0994*	.0990*
4. **NORM** ($\gamma_2 = 0.00$)								
	10, 10	.0216	.0860*	.0668*	.0832*	.0628*	.0880*	.0858*
	13, 13	.0170	.0856*	.0178	.0830*	.0156	.0874*	.0826*
	20, 20	.0136	.0748*	.0676*	.0738*	.0644*	.0756*	.0786
	10, 20	.0090	.0438	.0464	.0560	.0752*	.0480	.0516
	13, 20	.0122	.0606*	.0202	.0686*	.0132	.0636*	.0610*
	20, 10	.0234	.1530*	.0814*	.1168*	.0780*	.1528*	.1508*
	20, 13	.0156	.1238*	.0752*	.0988*	.0652*	.1244*	.1266*
5. **LEP1** ($\gamma_2 = 1.00$)								
	10, 10	.0324	.1090*	.0712*	.1052*	.0674*	.1106*	.1054*
	13, 13	.0262	.1038*	.0248	.1022*	.0230	.1052*	.0992*
	20, 20	.0252	.1036*	.0872*	.1022*	.0836*	.1040*	.1074*
	10, 20	.0160	.0614*	.0548	.0706*	.0838*	.0642*	.0652*
	13, 20	.0224	.0818*	.0252	.0886*	.0144	.0846*	.0790*
	20, 10	.0460	.1930*	.1034*	.1732*	.1068*	.1934*	.1934*
	20, 13	.0354	.1624*	.1078*	.1440*	.1018*	.1630*	.1626*
6. **LEP3** ($\gamma_2 = 3.00$)								
	10, 10	.0538	.1392*	.0934*	.1354*	.0884*	.1400*	.1366*
	13, 13	.0564*	.1470*	.0398	.1454*	.0366	.1470*	.1380*
	20, 20	.0608*	.1316*	.1172*	.1300*	.1140*	.1316*	.1438*
	10, 20	.0328	.0844*	.0680*	.0888*	.1012*	.0856*	.0856*
	13, 20	.0446	.0990*	.0404	.1036*	.0234	.1006*	.0942*
	20, 10	.0820*	.2430*	.1338*	.2182*	.1518*	.2426*	.2398*
	20, 13	.0804*	.2188*	.1418*	.1890*	.1496*	.2190*	.2178*

Note. * indicates the Type I error rate exceeded .0562 and is 2 standard errors above the nominal alpha of $\alpha = .05$.

Table 7. Empirical Type I Error Rates for seven procedures in Quasi-Experiments with no differences in central location and Group One is normally distributed $\gamma_2 = 0$.

Group 2 Pop.	n_1, n_2	*Fmax*	OB	BF	WOB	WBF	OB_{BF}	BF_{OB}
2. **PLAT** (γ_2 = -1.00)								
	10, 10	.0362	.0458	.0432	.0420	.0404	.0470	.0484
	13, 13	.0280	.0472	.0340	.0434	.0328	.0478	.0412
	20, 20	.0280	.0532	.0500	.0516	.0492	.0534	.0506
	10, 20	.0304	.0588*	.0536	.0860*	.0668*	.0606*	.0588*
	13, 20	.0304	.0566*	.0484	.0654*	.0536	.0578*	.0534
	20, 10	.0304	.0516	.0390	.0468	.0404	.0532	.0456
	20, 13	.0228	.0422	.0328	.0360	.0282	.0430	.0400
3. **SPLT** (γ_2 = -0.50)								
	10, 10	.0446	.0400	.0396	.0312	.0370	.0416	.0412
	13, 13	.0408	.0444	.0312	.0368	.0296	.0454	.0372
	20, 20	.0428	.0482	.0414	.0444	.0412	.0498	.0430
	10, 20	.0442	.0404	.0404	.0670*	.0592	.0416	.0406
	13, 20	.0422	.0474	.0424	.0522	.0460	.0490	.0442
	20, 10	.0424	.0394	.0354	.0454	.0424	.0410	.0380
	20, 13	.0370	.0442	.0310	.0402	.0338	.0454	.0356
6. **LEP3** (γ_2 = 3.00)								
	10, 10	.1048*	.0478	.0500	.0374	.0440	.0496	.0510
	13, 13	.1060*	.0460	.0442	.0370	.0404	.0494	.0464
	20, 20	.1188*	.0536	.0562*	.0488	.0542	.0598*	.0572*
	10, 20	.0960*	.0586*	.0470	.0292	.0366	.0602*	.0492
	13, 20	.1018*	.0564*	.0454	.0340	.0364	.0584*	.0482
	20, 10	.1194*	.0434	.0614*	.0990*	.0926*	.0474	.0624*
	20, 13	.1184*	.0524	.0638*	.0730*	.0746*	.0568*	.0640*
7. **XLEP** (γ_2 = 3.75)								
	10, 10	.1076*	.0422	.0494	.0326	.0458	.0446	.0500
	13, 13	.1174*	.0480	.0474	.0406	.0452	.0524	.0510
	20, 20	.1228*	.0564*	.0658*	.0526	.0648*	.0660*	.0650*
	10, 20	.1056*	.0584*	.0468	.0280	.0336	.0614*	.0484
	13, 20	.1166*	.0554	.0510	.0344	.0384	.0606*	.0524
	20, 10	.1272*	.0496	.0680*	.1040*	.0996*	.0550	.0686*
	20, 13	.1376*	.0548	.0676*	.0802*	.0814*	.0594*	.0672*

Note. * indicates the Type I error rate exceeded .0562 and is 2 standard errors above the nominal alpha of $\alpha = .05$ test.

Table 8. Empirical Type I Error Rates for seven procedures in Quasi-Experiments with no differences in central location and Group One is leptokurtic $\gamma_2 = 1.00$.

Group 2 Pop.	n_1,n_2	*Fmax*	OB	BF	WOB	WBF	OB_{BF}	BF_{OB}
2. **PLAT** (γ_2 = -1.00)								
	10, 10	.0406	.0524	.0434	.0436	.0410	.0530	.0492
	13, 13	.0444	.0622*	.0460	.0566*	.0436	.0632*	.0546
	20, 20	.0448	.0684*	.0616*	.0656*	.0606*	.0696*	.0656*
	10, 20	.0538	.0730*	.0720*	.1134*	.0980*	.0756*	.0740*
	13, 20	.0478	.0752*	.0744*	.0954*	.0826*	.0768*	.0756*
	20, 10	.0268	.0520	.0362	.0400	.0336	.0520	.0438
	20, 13	.0350	.0548	.0352	.0402	.0322	.0550	.0446
3. **SPLT** (γ_2 = -0.50)								
	10, 10	.0602*	.0436	.0412	.0350	.0368	.0448	.0442
	13, 13	.0548	.0472	.0350	.0394	.0316	.0474	.0404
	20, 20	.0528	.0510	.0464	.0460	.0458	.0534	.0480
	10, 20	.0622*	.0484	.0556	.0922*	.0798*	.0498	.0576*
	13, 20	.0586*	.0468	.0494	.0630*	.0562*	.0492	.0502
	20, 10	.0458	.0546	.0380	.0386	.0410	.0558	.0404
	20, 13	.0436	.0452	.0358	.0354	.0288	.0474	.0394
4. **NORM** (γ_2 = 0.00)								
	10, 10	.0702*	.0322	.0378	.0242	.0348	.0336	.0396
	13, 13	.0702*	.0370	.0302	.0302	.0272	.0382	.0342
	20, 20	.0772*	.0470	.0460	.0424	.0444	.0516	.0492
	10, 20	.0734*	.0444	.0444	.0714*	.0616*	.0466	.0452
	13, 20	.0700*	.0400	.0432	.0514	.0496	.0428	.0444
	20, 10	.0628*	.0436	.0378	.0402	.0420	.0464	.0402
	20, 13	.0670*	.0400	.0342	.0346	.0306	.0422	.0360
6. **LEP3** (γ_2 = 3.00)								
	10, 10	.0858*	.0326	.0370	.0264	.0328	.0336	.0380
	13, 13	.0812*	.0366	.0354	.0316	.0320	.0384	.0382
	20, 20	.0982*	.0472	.0472	.0426	.0460	.0512	.0480
	10, 20	.0858*	.0462	.0392	.0350	.0408	.0496	.0402
	13, 20	.0936*	.0434	.0390	.0304	.0342	.0480	.0404
	20, 10	.1134*	.0384	.0476	.0682*	.0722*	.0424	.0478
	20, 13	.1248*	.0380	.0448	.0522	.0540	.0428	.0456
7. **XLEP** (γ_2 = 3.75)								
	10, 10	.1012*	.0352	.0382	.0254	.0338	.0380	.0396
	13, 13	.1078*	.0374	.0360	.0296	.0326	.0400	.0360
	20, 20	.1200*	.0402	.0492	.0360	.0472	.0500	.0494
	10, 20	.0996*	.0492	.0382	.0282	.0342	.0530	.0040
	13, 20	.1110*	.0464	.0438	.0350	.0338	.0526	.0444
	20, 10	.1264*	.0408	.0514	.0704*	.0734*	.0452	.0534
	20, 13	.1210*	.0382	.0496	.0568*	.0584*	.0432	.0502

Note. * indicates the Type I error rate exceed .0562 and is 2 standard errors above the nominal alpha of $\alpha = .05$ test.

Table 9. Empirical Power Estimates for seven procedures in Quasi-Experiments with no differences in central location and Group One is normally distributed $\gamma_2 = 0.00$.

									Group Two
PLAT	γ_2 = -1.00			σ_1^2 = 1.0			σ_2^2 = 2.0		
Pop.	n_1, n_2	*Fmax*	OB	BF	WOB		WBF	OB_{BF}	BF_{OB}
	10, 10	.1540	.1724	.1550	.1496	.1462	.1748	.1682	
	13, 13	.1868	.2386	.1862	.2234	.1784	.2418	.2046	
	20, 20	.2956	.3974	.3526	.3888	.3488	.4032	.3672	
Pos.	10, 20	.1684	.2120*	.2358	----	----	----	.2380*	
Pos.	13, 20	.2064	.2746*	.2758	----	.3274	.2812*	.2842	
Neg.	20, 10	.2104	.2906	.2048	.1242	.1248	.2940	.2198	
Neg.	20, 13	.2276	.3222	.2246	.2038	.1542	.3256	.2456	
Group Two **PLAT**	γ_2 = -1.00			σ_1^2 = 2.0			σ_2^2 = 1.0		
	10, 10	.1292	.0710	.0722	.0540	.0652	.0722	.0760	
	13, 13	.1810	.1150	.0856	.0964	.0806	.1154	.0960	
	20, 20	.2998	.2264	.1748	.2086	.1706	.2268	.1978	
Neg.	10, 20	.2406	.2520*	.1386	----	----	----	.1762*	
Neg.	13, 20	.2686	.2396*	.1340	----	.0718	.2402*	.1700	
Pos.	20, 10	.1268	.0474	.0930	.2142	.1762	.0512	.0908	
Pos.	20, 10	.1890	.0976	.1232	.2182	.1770	.0992	.1262	
Group Two **LEP1**	γ_2 = 1.00			σ_1^2 = 1.0			σ_2^2 = 2.0		
	10, 10	----	.0640	.0798	.0458	.0722	.0670	.0814	
	13, 13	----	.0940	.0890	.0772	.0798	.0968	.0924	
	20, 20	----	.1690	.1690	.1556	.1640	.1844	.1736	
Pos.	10, 20	----	.0404	.0878	.1816	.1820	.0486	.0882	
Pos.	13, 20	----	.0692	.1130	.1618	.1670	.0794	.1130	
Neg.	20, 10	----	.1852	.1238	----	----	.1884	.1314	
Neg.	20, 13	----	.1676	.1224	.0608	.0682	.1732	.1306	
Group Two **LEP1**	γ_2 = 1.00			σ_1^2 = 2.0			σ_2^2 = 1.0		
	10, 10	----	.1038	.1148	.0826	.1054	.1088	.1178	
	13, 13	----	.1460	.1420	.1218	.1346	.1514	.1472	
	20, 20	----	.2526	.2646	.2386	.2598	.2690	.2658	
Neg.	10, 20	----	.2320	.1758	.0608	.0818	.2352	.1822	
Neg.	13, 20	----	.2348	.1920	.1058	.1118	.2404	.1984	
Pos.	20, 10	----	.0880	.1654	----	----	.0990	.1622	
Pos.	20, 13	----	.1412	.1998	.2518	.2632	.1574	.2010	

Note. * indicates that the Type I error rate exceeded the nominal alpha by 2 standard errors and should be interpreted cautiously. Blank entries indicate that the Type I error rate exceeded Cochran's limit of .06. Pos. indicates a positive correlation between sample size and variance; Neg. indicates an inverse sample size-variance relationship.

Table 10. Empirical Power Estimates for seven procedures in Quasi Experiments with no differences in central location and Group One has positive kurtosis $\gamma_2 = 1.00$.

Group Two **SPLT**	$\gamma_2 = -0.50$			$\sigma_1^2 = 1.0$			$\sigma_2^2 = 2.0$	
Pop.	n_1, n_2	*Fmax*	OB	BF	WOB	WBF	OB_{BF}	BF_{OB}
	10, 10	---	.1390	.1362	.1102	.1272	.1442	.1418
	13, 13	.2428	.1954	.1782	.1722	.1682	.2012	.1854
	20, 20	.3410	.3082	.3042	.2960	.3006	.3218	.3084
Pos.	10, 20	---	.1278	.1866	----	----	.1382	.1864*
Pos.	13, 20	.2652*	.1878	.2338	----	.2974*	.1994	.2330
Neg.	20, 10	.2680	.2582	.2040	.0844	.1114	.2622	.2086
Neg.	20, 13	.2858	.2706	.2174	.1454	.1462	.2778	.2244
Group Two **SPLT**	$\gamma_2 = -0.50$			$\sigma_1^2 = 2.0$			$\sigma_2^2 = 1.0$	
	10, 10	---	.0588	.0674	.0452	.0602	.0610	.0700
	13, 13	.1740	.0796	.0748	.0638	.0674	.0814	.0782
	20, 20	.2730	.1408	.1286	.1298	.1244	.1486	.1362
Neg.	10, 20	.2242	.1832	.1098	.0350	.0424	.1854	.1218
Neg.	13, 20	.2552	.1682	.1150	.0614	.0574	.1718	.1258
Pos.	20, 10	---	.0318	.0698	----	----	.0364	.0690*
Pos.	20, 13	.2058*	.0694	.1032	----	.1510*	.0768	.1044
Group Two **LEP3**	$\gamma_2 = 3.00$			$\sigma_1^2 = 1.0$			$\sigma_2^2 = 2.0$	
	10, 10	---	.0460	.0630	.0344	.0530	.0496	.0650
	13, 13	---	.0690	.0782	.0558	.0704	.0768	.0812
	20, 20	---	.1048	.1238	.0978	.1196	.1278	.1236
Pos.	10, 20	---	.0272	.0722	.1450	.1654	.0362	.0716
Pos.	13, 20	---	.0444	.0832	.1174	.1386	.0576	.0836
Neg.	20, 10	---	.1388	.1052	----	----	.1460	.1086
Neg.	20, 13	---	.1272	.1062	.0436	.0530	.1386	.1096
Group Two **LEP3**	$\gamma_2 = 3.00$			$\sigma_1^2 = 2.0$			$\sigma_2^2 = 1.0$	
	10, 10	---	.0932	.1174	.0712	.1002	.0982	.1200
	13, 13	---	.1112	.1354	.0920	.1234	.1232	.1374
	20, 20	---	.1898	.2402	.1770	.2346	.2288	.2412
Pos.	20, 10	---	.0732	.1490	.2386	.2654	.0876	.1476
Pos.	20, 13	---	.1020	.1812	.2104	.2450	.1272	.1796
Neg.	10, 20	---	.1962	.1716	----	----	.2102	.1746
Neg.	13, 20	---	.2010	.1876	.0796	.1112	.2150	.1902

Note. * indicates that the Type I error rate exceeded the nominal alpha by 2 standard errors and should be interpreted cautiously. Blank entries indicate that the Type I error rate exceeded Cochran's limit of .06. Pos. indicates a positive correlation between sample size and variance; Neg. indicates an inverse sample size-variance relationship.

Table 11. Recommendations based on Variance Ratios and whether the sample sizes are equal, positively correlated, or negatively correlated with the variances for both True and Quasi-Experiments.

Smaller Variance	Larger Variance			
	PLAT	***SPLT***	***NORM***	***LEPT***
PLAT				
Equal	OB	OB, *Fmax*	*Fmax*	*Fmax*
Positive	WOB	WOB	WOB	*Fmax*
Negative	OB	OB	*Fmax*	*Fmax*
SPLT				
Equal	OB, *Fmax*	OB, *Fmax*	*Fmax*, OB	OB, BF
Positive	WOB	WOB	WOB	BF
Negative	OB	OB, *Fmax*	*Fmax*, OB	OB
NORM				
Equal	OB	*Fmax*, OB	*Fmax*	BF, OB
Positive	BF	WOB	WOB	WBF, WOB
Negative	OB	*Fmax*, OB	*Fmax*	OB
LEPT				
Equal	*Fmax*	OB, BF	BF, OB	BF
Positive	*Fmax*	BF	BF, WBF	WBF
Negative	*Fmax*	OB	OB	OB

Note. Entries on the diagonal represent recommendations for True Experiments, while off-diagonal entries are for Quasi-Experiments. PLAT = platykurtic; SPLT = slightly platykurtic; NORM = Normal; LEPT = leptokurtic; OB = O'Brien BF = Brown-Forsythe; W refers to performing Welch procedure