Parametric Test for Two Population Means: An Empirical Comparison of Type I Error Control and Statistical Power

Patricia Rodríguez de Gil[1], Yi-Hsin Chen[1], Eun Sook Kim[1]
Diep Nguyen[1], Anh Kellermann[1], Aarti Bellara[1] , Jeffrey Kromrey[1]
[1]University of South Florida, 4202 E. Fowler Avenue EDU 105, Tampa, FL 33620

Abstract

A simulation study was conducted to explore the Type I error rates and statistical power of the independent means t-test, the Satterthwaite's approximate t-test, and the conditional t-test based on a preliminary test of variances. Factors manipulated were total sample size (10, 20, 50, 100, 200, 300, 400), sample size ratio between the groups (1:1, 2:3, 1:4), population variance ratio (1:1, 1:2, 1:4, 1:8, 1:12, 1:16, 1:20), population effect size (0, .2, .5, .8), and alpha for both the test of treatment effect and the test of variances. Normal population distributions and distributions with varying kurtosis (γ_1) and skewness (γ_2) values (γ_1=1.00, γ_2=3.00; γ_1=1.50, γ_2=5.00; γ_1=2.00, γ_2=6.00; γ_1=0.00, γ_2=25.00) were included. As expected, the independent means *t*-test showed great dispersion of Type I error control. The Satterthwaite *t*-test provided adequate Type I error control in nearly all conditions and the conditional t-test evidenced notable improvement in Type I error control relative to the independent means *t*-test as the level of alpha for the test of variances increased. This study provides guidance on the proper use of parametric tests with nonnormal, heteroscedastic populations.

Key Words: Type I error control, statistical power, parametric tests, Satterthwaite's approximate test, Conditional *t*-test

1. The Independent T-Test and Alternatives

1.1 Introduction

Elementary statistics courses typically introduce significance testing and inferential techniques using the independent means *t*-test, which provides a smooth transition into concepts such as statistical assumptions, robustness, Type I error control, and power. The independent means *t*-test relies on the assumptions of population normality and equal variances (homoscedasticity). Alternative approaches (e.g., Satterthwaite's approximate test) relax these assumptions, approximating the *t* distribution and the corresponding degrees of freedom. Although the independent means *t*-test is "the most powerful unbiased test" (Bridge and Sawilowski, 1999; p. 229) for detecting true mean differences under the assumption of normality, statisticians to date are still evaluating the various conditions and factors for which this test is robust under the violation of the equality of variances assumption and departures from normality. Many statistical textbooks (e.g., Cody & Smith, 1997; Schlotzhauer & Littell, 1997) continue recommending what Hayes and Cai (2007) call the "conditional decision rule" (p. 217), that researchers screen their samples for variance homogeneity by conducting preliminary tests (e.g., the Folded *F*-test). That is, the *t*-test assumes that the distributions of the two groups being compared are normal with equal variances. The preliminary test of the null hypothesis that $\sigma_1^2 = \sigma_2^2$ versus the alternative $\sigma_1^2 \neq \sigma_2^2$ is conducted using the test statistic: $F = \frac{S_1^2}{S_2^2}$.

Common practice has been that if the Folded F-test is not statistically significant (e.g., $p > .05$), then the test of $\mu_1 = \mu_2$ versus $\mu_1 \neq \mu_2$ is calculated using the independent means t-test:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{X_1X_2} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

On the other hand, if the preliminary test is statistically significant ($p < .05$) and in addition there are unequal sample sizes, the independent means t-test should be avoided and the Satterthwaite's approximate t-test should be used instead (Moser, Stevens, & Watts, 1989):

$$t' = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}} \quad \text{with} \quad df = \frac{\left(\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}\right)^2}{\frac{\left(\frac{\hat{\sigma}_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{\hat{\sigma}_2^2}{n_2}\right)^2}{n_2 - 1}}$$

Due to the unlikeliness of encountering real data that are normally distributed (see Micceri, 1989) and the serious disadvantages of performing preliminary tests of equality of variances (e.g., Moser et al., 1989; Zimmerman, 2004), researchers have questioned the robustness of the conditional t-test with respect to Type I error and statistical power when the assumptions of normality and homoscedasticity are not met.

Thus, the purpose of this study was to explore the performance of the independent t-test and alternatives (i.e., Satterthwaite's t test and the conditional t-test) by conducting a series of simulations under various manipulated conditions. Some guidelines for researchers on the selection of an appropriate test were also recommended.

1.2 Previous Research on Conditional Testing

While some statistics textbooks do not even mention the assumption of homogeneity of variance (e.g., Gravetter & Wallnau, 2011) as one required for the t-test, homoscedasticity is basic and necessary for hypothesis testing because the violations of this assumption "alter Type I error rates, especially when sample sizes are unequal" (Zimmerman, 2004; p. 173). However, some research on preliminary tests suggests that the choice between the independent means t-test and the Satterthwaite's approximate test, conditioning on a preliminary test of the assumption of homogeneity of variance is not effective.

Moser et al. (1989) examined the effect of the significance level of the preliminary test of variance on the size and power of the t-test and Satterthwaite's tests of means and noted that when $\alpha = 0$ or $\alpha = 1$ was established for the significance level of the test of variances, it allowed applying directly the t-test or Satterthwaite's, respectively. In addition, they suggested that for equal sample sizes ($n_1 = n_2$), the t-test and the Satterthwaite's had the same power and provided very stable sizes close to the nominal alpha prescribed for the test of means. For unequal sample sizes ($n_1 \neq n_2$), the Satterthwaite's test still provided reasonable and stable sizes close to the nominal significance level. Based on their study, Moser et al. (1989) recommended applying directly the Satterthwaite's test for testing the equality of means from two independent and normally distributed populations where the ratio of the variance is unknown. Both Zimmerman (2004) and Rasch, Kubinger, and Moder (2011) found similar optimal results for the Welch-Satterthwaite separate-variance t-test if applied unconditionally whenever sample sizes were unequal and noted that the power of this test deteriorated if it was conditioned by a preliminary test. Grissom (2000) argued that it is realistic to expect heteroscedasticity in data as well as outliers, and examined the effect of these factors on variance. He also addressed issues of robustness (i.e., control of Type I error rate) in the presence of heteroscedasticity and departures from normality, for which he suggested trimming as a way to stabilize variances.

2. Method

A completely crossed factorial mixed design included six between-subjects factors: (a) total sample size (10, 20, 50, 100, 200, 300, and 400), (b) sample size ratio between groups (1:1, 1:3, and 1:4), (c) variance ratio between populations (1:1, 1:2, 1:4, 1:8, 1:12, 1:16, and 1:20), (d) effect size for mean difference between populations ($\Delta = 0, .2, .5, .8$), (e) alpha set for testing treatment effects ($\alpha = .01, .05, .10, .15, .20,$ and $.25$), (f) alpha set for testing homogeneity assumption for the conditional *t*-test ($\alpha = .01, .05, .10, .15, .20, .25, .30, .40, .45,$ and $.50$), and one within-subjects, (g) population distributions with varying kurtosis and skewness values ($\gamma_1 = 1.00$, $\gamma_2 = 3.00$; $\gamma_1 = 1{:}50$, $\gamma_2 = 5{:}00$; $\gamma_1 = 2.00$, $\gamma_2 = 6.00$; $\gamma_1 = 0.00$, $\gamma_2 = 25.00$, including the normal distribution data from previous research, $\gamma_1 = 0.00$, $\gamma_2 = 0.00$). For each condition, 100,000 samples were simulated. The use of 100, 000 replications provided a maximum standard error of an observed proportion (e.g., Type I error rate estimates) of .0015 and a 95% confidence interval no wider than ± .003 (Robey & Barcikowski, 1992). This crossed factorial design (7 x 3 x 7 x 4 x 6 x 10 x 5) provided a total of 176,400 conditions.

3. Results

3.1 Type I Error Control for the Tests of Means

An overall view of the Type I error control of the tests is provided in Figure 1. These boxplots describe the distributions of the Type I error rate estimates under a nominal alpha level of .05 across all conditions in which the population means were identical. The first two plots are for the independent means *t*-test and Satterthwaite's approximate *t*-test, respectively. The remaining plots delineate the Type I error rate estimates for the conditional *t*-test across the different conditioning rules (i.e., the alpha levels for the Folded *F*-test) that were investigated. For instance, the plot for C (01) provides the distribution of the Type I error rates for the conditional *t*-test when an alpha level of .01 was used with the Folded *F*-test as the rule to choose between the independent means *t*-test and Satterthwaite's approximate *t*-test.

Note that in Figure 1 the independent means *t*-test has the great dispersion of the Type I error rates. In some conditions, this testing approach provides appropriate control of the Type I error probability while in others the Type I error rate is very different from the nominal alpha level. In contrast, Satterthwaite's approximate *t*-test provides adequate Type I error control in nearly all of the conditions simulated. The series of plots for the conditional *t*-test illustrate that the conditional test provides a notable improvement in Type I error control relative to the independent means *t*-test and the improvement increases as the alpha level for the Folded *F*-test is increased. This improvement occurs because the statistical power of the Folded *F*-test increases as the alpha level is increased. That is, the ability of this test to detect variance heterogeneity (and to subsequently steer us away from the independent means *t*-test and steer us to Satterthwaite's approximate *t*-test) increases with the alpha level for this test, which supports the argument of insufficient power when using a more conservative nominal alpha level for a preliminary analysis.

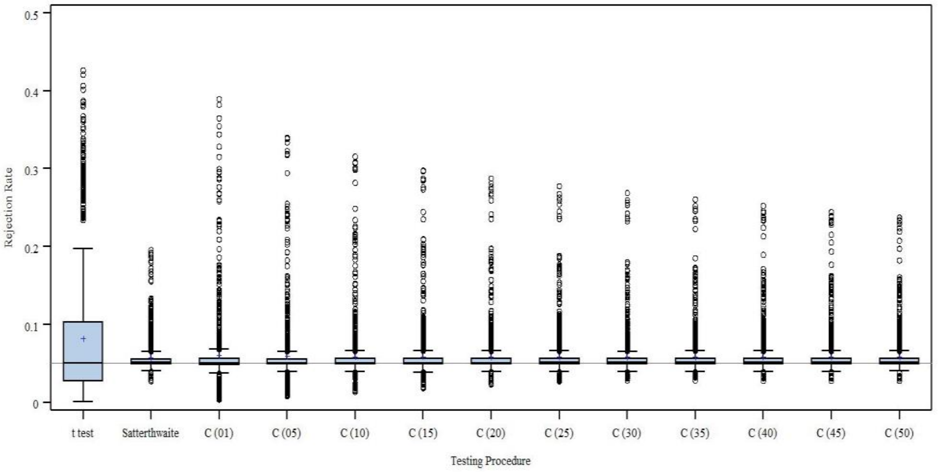

Figure 1: Distributions of estimated Type I error rates for testing approaches (Nominal alpha = .05)

3.2 Power for the Folded F-Test

Figure 2 provides the distributions of statistical power estimates for the Folded *F*-test across all conditions simulated in which population variances were not equal. As expected, when the alpha level used for the Folded *F*-test is small (e.g., .01 or .05) the average power is low. However, the power of the Folded *F*-test increases when the applied alpha level increase.

Nominal alpha levels of .05 and .25 for the Folded *F*-test were selected for further analysis of power. The average power of the Folded *F*-test based on simulation design factors is presented in Table 1. As seen in the table, the power remains stable regardless of distribution shapes, yet using the alpha level of .25 consistently yielded more power. The average powers for .05 and .25 alpha levels were around .81 and .90, respectively, across normal and non-normal distributions. Further, as the value of variance ratio increases, the power of the Folded *F*-test increase as well. Using the alpha level of .25 provided substantially more power when the variance ratios were small (i.e., variance ratio = 1:2 and 1:4). As the variance ratios increased, the power differences between the two nominal alpha levels decreased.

It is well-known that the power increases when the sample size increases. Using an alpha level of .05 for the Folded *F*-test yielded average power of .80 with sample size of 50 and of .90 with 100. In contrast, the average power reached .80 with as few as 20 observations and .90 with 50 observations using an alpha level of .25. The use of extremely unbalanced samples (sample size ratios of 1:4 or 4:1) reduced the power of the Folded *F*-test, but power advantages of the more liberal alpha level remained evident.

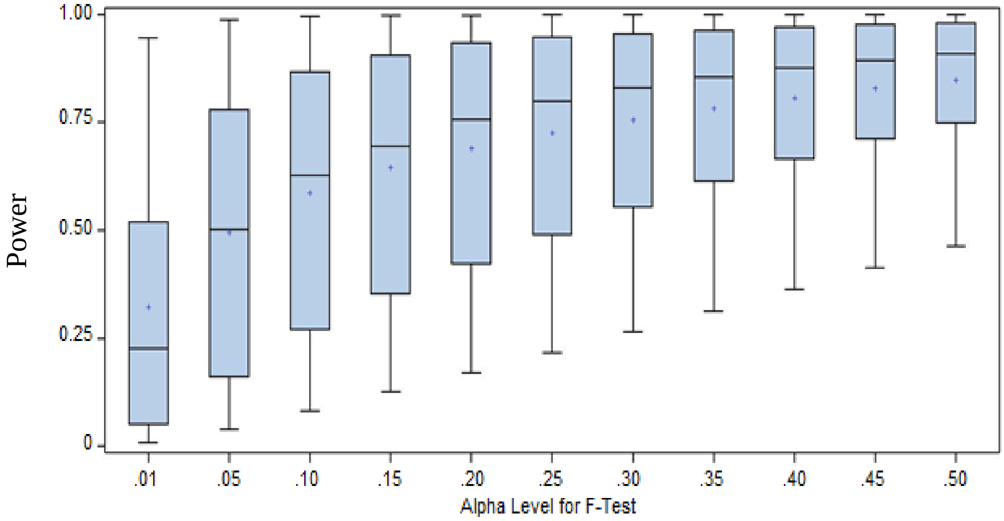

Figure 2: Distributions of statistical power for the Folded *F*-Test by nominal alpha level

Table 1: *The Power of the Folded F-Test Using α = .05 and α = .25*

Condition	*α = .05*	*α = .25*	*Condition*	*α = .05*	*α = .25*
Total N			*Variance ratio*		
10	0.36	0.62	1:2	0.55	0.73
20	0.64	0.82	1:4	0.76	0.87
50	0.85	0.92	1:8	0.85	0.93
100	0.92	0.96	1:12	0.89	0.93
200	0.96	0.98	1:16	0.91	0.96
300	0.98	0.99	1:20	0.92	0.97
400	0.99	0.99			
N ratio			*Shape*		
1:4	0.74	0.86	0,0	0.82	0.91
2:3	0.83	0.92	1,3	0.81	0.90
1	0.85	0.93	1.5,5	0.81	0.90
3:2	0.85	0.92	2,6	0.81	0.89
4:1	0.80	0.82	0,25	0.81	0.91

Note: For shape, the two values indicate skewness and kurtosis, respectively

3.3 Impact of Variance Heterogeneity on the Test of Means

The large dispersion of Type I error rates for the independent means *t*-test (Figure 3) results in large part from the variance heterogeneity that was included in the simulation conditions. Figure 3 presents the distributions of Type I error rates for the independent means *t*-test with the results disaggregated by population variance ratio. Note that as the population variance ratio increases, both the average Type I error rate and the dispersion of Type I error rates increase.

Figures 4 and 5 present the analogous distributions of Type I error rates for Satterthwaite's approximate *t*-test and the conditional *t*-test, respectively, with an alpha level of .25 as a decision rule for the Folded-*F* test. Both the Satterthwaite's approximate *t*-test and the conditional *t*-test provided good control of Type I error rate even if the population variances in the two groups are heterogeneous.

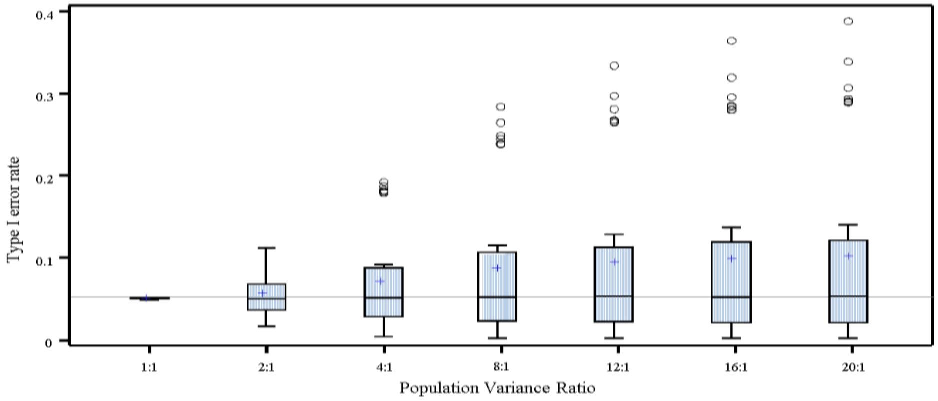

Figure 3: Distributions of estimated Type I Error rates by variance ratio at α = .05 for the Independent Means *T*-Test

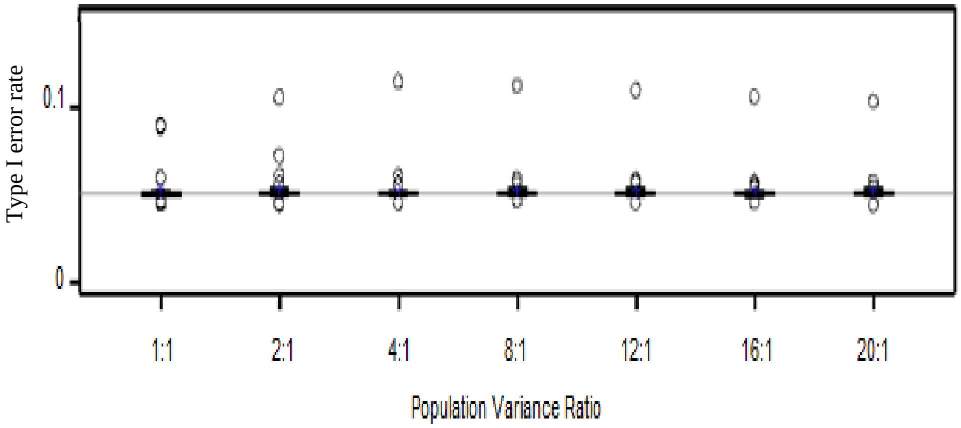

Figure 4: Distributions of estimated Type I error rates by variance ratio at α = .05 for the Satterthwaite's Approximate *T*-Test

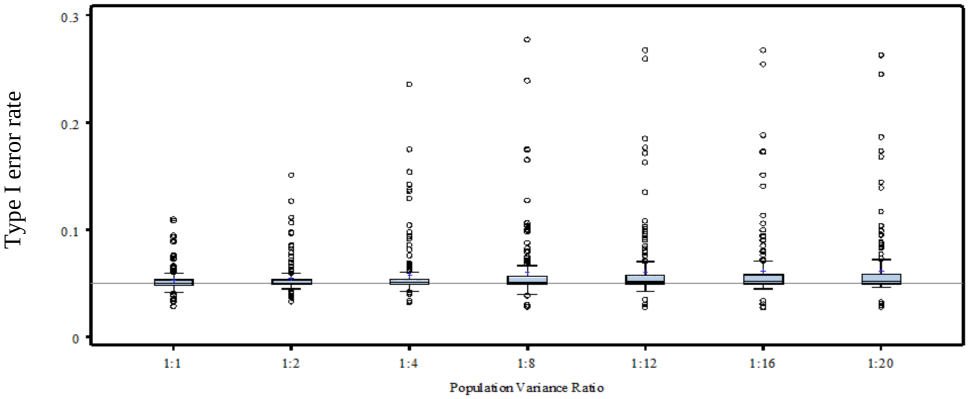

Figure 5: Distributions of estimated Type I error rates by variance ratio at α = .05 for the Conditional *T*-Test Using α = .25 for the Folded *F*-Test.

Of course, the independent means *t*-test is known to be relatively robust to violations of the assumption of variance homogeneity if the sample sizes in the two groups are equal. This phenomenon is illustrated in Figure 6. Note that the Type I error rate for the independent means *t*-test is maintained near the nominal .05 level if sample sizes are equal. With disparate sample sizes in the two groups, the independent means *t*-test either becomes conservative (Type I error rates lower than the nominal alpha level) or liberal

(Type I error rates higher than the nominal level) depending upon the relationship between sample size and population variance. In contrast, both Satterthwaite's approximate *t*-test (Figure 7) and the conditional *t*-test (Figure 8) evidence much improved Type I error control under variance heterogeneity when samples sizes are unequal.

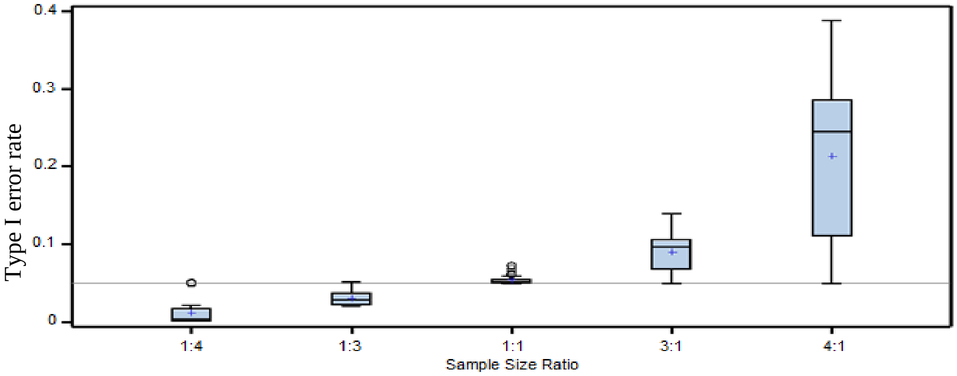

Figure 6: Distributions of estimated Type I error rates by sample size ratio for the Independent Means *T*-Test

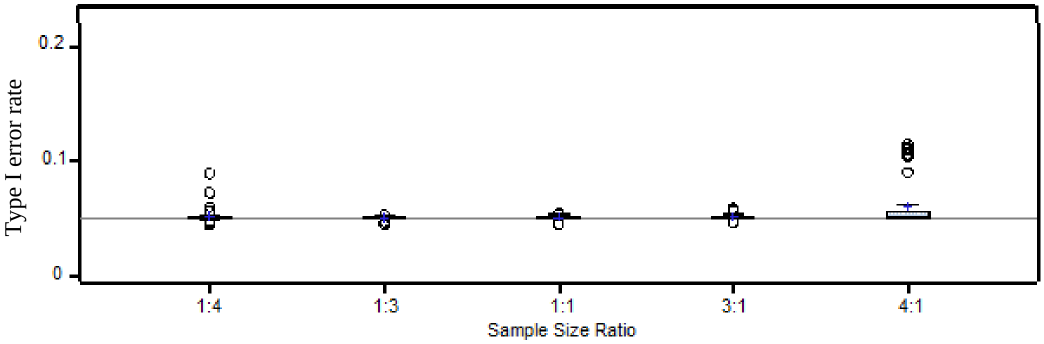

Figure 7: Distributions of estimated Type I error rates by sample size ratio for Satterthwaite's Approximate *T*-Test

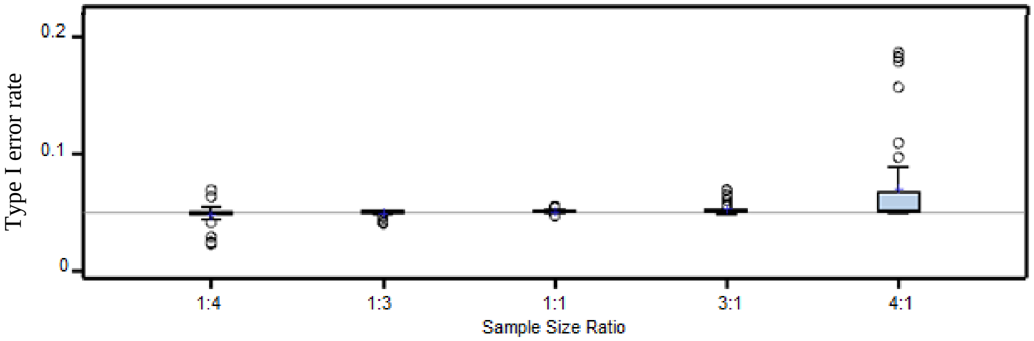

Figure 8: Distributions of estimated Type I error rates by sample size ratio for the conditional *t*-test.

Using a larger sample size does not improve the performance of the independent means *t*-test (Figure 9), but larger samples provide substantial improvements to the Type I error control of both Satterthwaite's approximate *t*-test (Figure 10) and the conditional *t*-test (Figure 11).

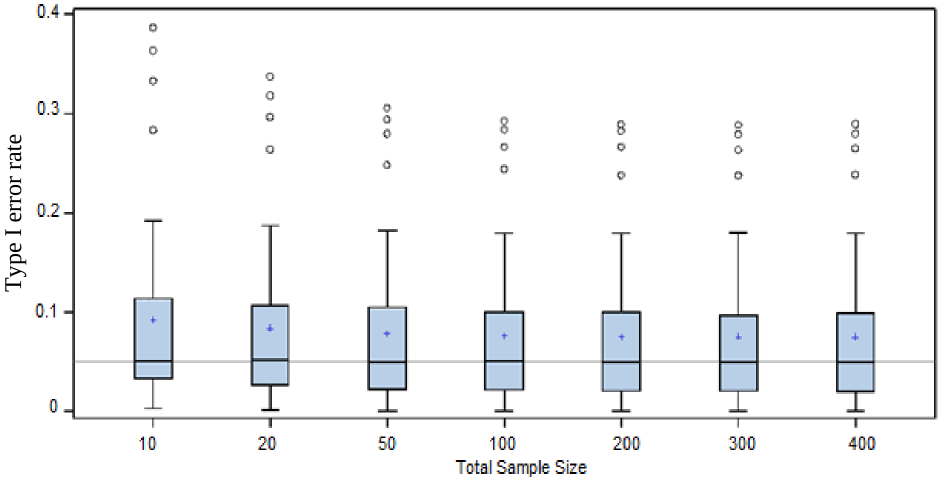

Figure 9: Distributions of estimated Type I error rates for the independent *t*-test by total sample size

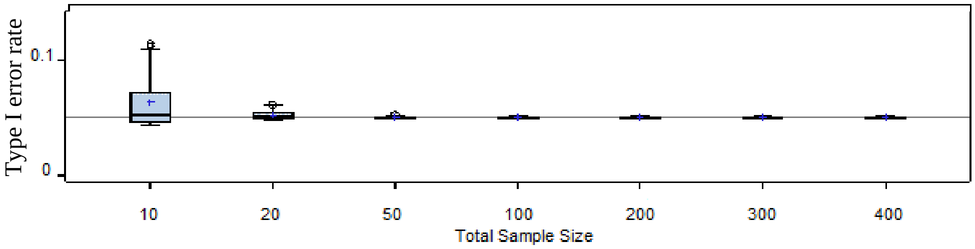

Figure 10: Distributions of estimated Type I error rates for Satterthwaite's approximate *t*-test by total sample size

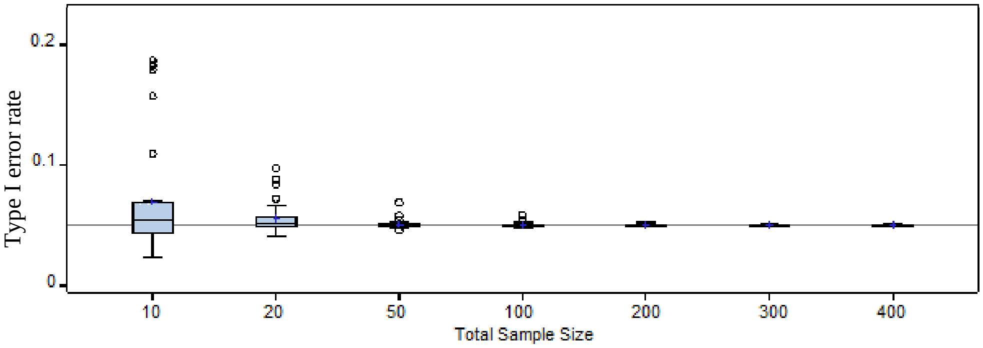

Figure 11. Distributions of estimated Type I error rates for the conditional *t*-test by total sample size

The variability in the estimated Type I error rates for the three tests of means was analyzed by computing the eta-squared value associated with each simulation design factor and the first-order interactions. For the independent means *t*-test, the factors associated with variability in estimated Type I error rates were sample size ratio (η^2 = 0.69) and the interaction between sample size ratio and variance ratio (η^2 = 0.22). For Satterthwaite's approximate *t*-test and the conditional *t*-test respectively, major factors were sample size ratio (η^2 = 0.15; η^2 =0.18), total sample size (η^2 = 0.18; η^2 =0.14), and the interaction between sample size ratio and total sample size (η^2 = 0.26; η^2 =0.36). An analysis of the sole impact of distribution shape on Type I error rates of the three tests showed that Type I error rate of Satterthwaite's approximate *t*-test was most affected (η^2 = 0.07). While Type I error rate of the independent means t-test was least impacted by the distribution shape (η^2 = 0.001), Type I error rate of the conditional *t*-test was also impacted (η^2 = 0.04) but to a much lesser degree in comparison with that of Satterthwaite's approximate *t*-test.

The mean Type I error rates by total sample size and distribution shape for the three test types under the nominal alpha level of .05 are presented in Figures 12, 13, and 14. Figure 12 shows that for the independent means *t*-test, the mean Type I error rates are much above the nominal alpha level in all conditions of distribution shapes and total sample sizes. Although the mean estimated Type I error rates decrease with larger samples, they remain substantially greater than .05. In contrast, both Satterthwaite's approximate *t*-test and the conditional *t*-test provided much better Type I error control except for extremely small sample sizes (i.e., total sample size = 10) or the extremely skewed distribution (i.e., skewness = 2) (see Figure 13 and 14, respectively).

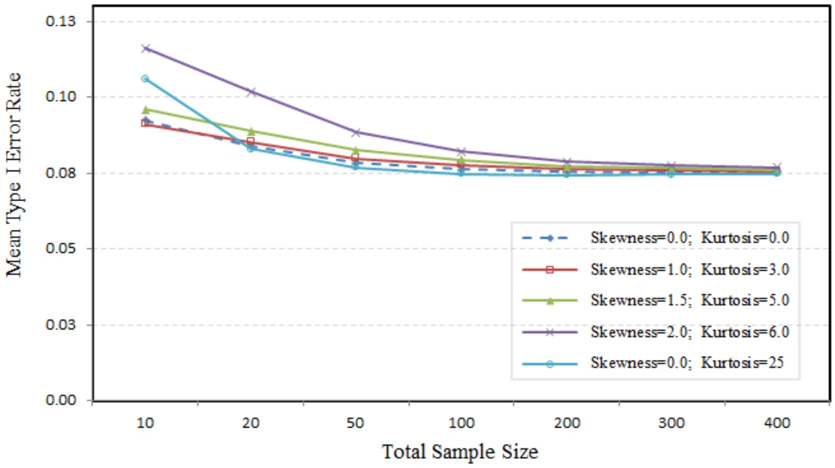

Figure 12: Mean Type I error rate for the Independent *t*-test by total sample size and distribution shape

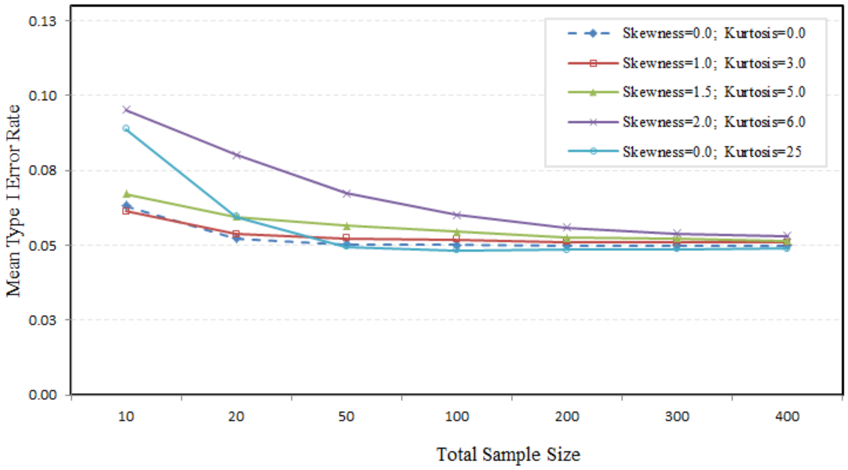

Figure 13: Mean Type I error rate for Satterthwaite's approximate *t*-test by total sample size and distribution shape

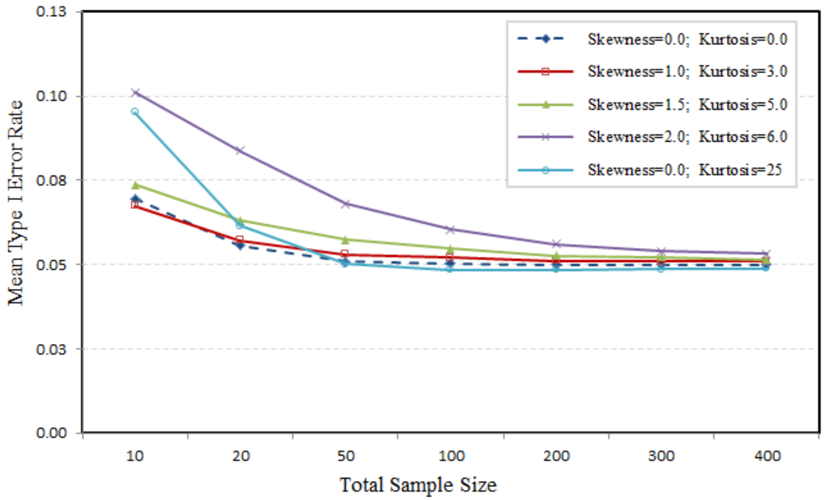

Figure 14: Mean Type I error rate for the conditional *t*-test by total sample size and distribution shape

3.4. Bradley's Liberal Criterion for Robustness of Type I Error Control

The Type I error rate estimates were further evaluated on the basis of the liberal criterion for robustness suggested by Bradley (1978). Given a nominal alpha level, Bradley's liberal criterion provides a plausible range of Type I error rates in which a test can be considered robust. The liberal criterion for the robustness is set at 0.5α around the nominal alpha. For example, when α = .05, a test is considered robust when the Type I error rate falls between .025 (= 0.5*.05) and .075 (= 1.5*.05).

For the independent means *t*-test, the overall proportions of cases meeting the Bradley's liberal criterion were .35, .44, and .52 for α = .01, .05, and .10, respectively. In contrast, the respective proportions for the Satterthwaite's approximate test were .77, .91, and .96. Depending on the alpha of the Folded *F*-test, the proportions of the conditional *t*-test are generally bound between those of independent means *t*-test as a lower limit and those of Satterthwaite's approximate tests as an upper limit. When the significance level of the Folded-*F* test was .25, the conditional *t*-test showed the largest proportion meeting the criterion close to that of Satterthwaite. The overall proportions meeting the criterion at α = .05 for independent means *t*-test, conditional *t*-test and Satterthwaite's approximate test are presented in Table 2.

Table 2: *The Proportions of Cases Meeting the Bradley's Liberal Criterion at α = .05*

Test	*t*-test	Conditional *t*-test											Satterthwaite
α[a]		.01	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50	
Proportion	.44	.84	.87	.88	.89	.90	.91	.88	.88	.88	.88	.88	.91

[a] The significance level of Folded *F*-test.

The factors related to the Bradley proportions vary across tests (Table 3). For the independent means *t*-test, the sample size ratio and the variance ratio between the two populations emerged as primary factors making an impact on the Type I error control. Although the overall proportions of conditions for the independent means *t*-test meeting the Bradley's criterion were very low (below 50%), the Type I error rates were perfectly controlled when the homogeneity of variance assumption was met (i.e., variance ratio between groups = 1:1). As the disproportion of two group variances became larger to 1:20, the Type I error control of independent means *t*-test diminished considerably. When the two groups have equal sample size, the independent means *t*-test adequately controlled the Type I error rates within the Bradley's criterion in 91% of the conditions. The imbalance of sample size between groups also worsened the Type I error control noticeably. On the other hand, the adequacy of Type I error control of the independent means *t*-test appears independent of total sample size and the shape of distribution. The proportions meeting the Bradley's criterion were consistently low irrespective of total sample size and distribution shape.

On the contrary, the impact of variance ratio and sample size ratio on the Type I error control of the Satterthwaite's approximate test and the conditional *t*-test appear minimal. Both tests showed adequate levels of Type I error control in the majority of conditions regardless of variance ratio and sample size ratio. Instead, total sample size and the skewness of the distribution were associated with the Bradley proportions of both the Satterthwaite's approximate test and the conditional *t*-test. When the total sample size was 10, the proportions meeting the criterion dropped to about 65%. In this total sample size condition (N = 10), the conditional *t*-test showed slightly better control of Type I error than the Satterthwaite's approximate test. Interestingly, the Type I error control of both tests was affected by skewness but not by kurtosis (see Table 3).

Table 3: *The Proportions of Cases Meeting the Bradley's Liberal Criterion by Tests and Conditions at α = .05*

Condition	*t*-test	Conditional	Satterthwaite	Condition	*t*-test	Conditional	Satterthwaite
Total N				Variance ratio			
10	0.45	0.68	0.65	1:1	1.00	0.94	0.92
20	0.49	0.76	0.82	1:2	0.62	0.93	0.95
50	0.45	0.93	0.95	1:4	0.40	0.91	0.93
100	0.43	0.97	0.97	1:8	0.29	0.90	0.91
200	0.42	1.00	1.00	1:12	0.27	0.89	0.89
300	0.41	1.00	1.00	1:16	0.25	0.89	0.89
400	0.41	1.00	1.00	1:20	0.23	0.89	0.90
N ratio				Shape			
1:4	0.18	0.98	0.97	0,0	0.43	0.96	0.97
2:3	0.67	0.98	0.98	1,3	0.44	0.96	0.97
1	0.91	0.97	0.97	1.5,5	0.44	0.93	0.94
3:2	0.28	0.91	0.91	2,6	0.46	0.77	0.78
4:1	0.14	0.70	0.74	0,25	0.41	0.91	0.91

Note. Conditional = conditional *t*-test at α = .25 of Folded *F*-test. For shape, the two values indicate skewness and kurtosis, respectively.

3.5 Statistical Power Analysis

Although Satterthwaite's approximate *t*-test provides superior Type I error control, it is not always the best test to select because of the potential for power differences. When the assumptions are met, the independent means *t*-test is the most powerful test for mean differences. For this simulation study, power comparisons were made only for conditions in which both the Satterthwaite's approximate *t*-test and the conditional *t*-test procedures evidenced adequate Type I error control by Bradley's (1978) benchmark. Figure 15 shows the distributions of power estimates for the conditional *t*-test and Satterthwaite's approximate *t*-test (at a nominal alpha of .05). As evident in this figure, the power differences between the tests were small.

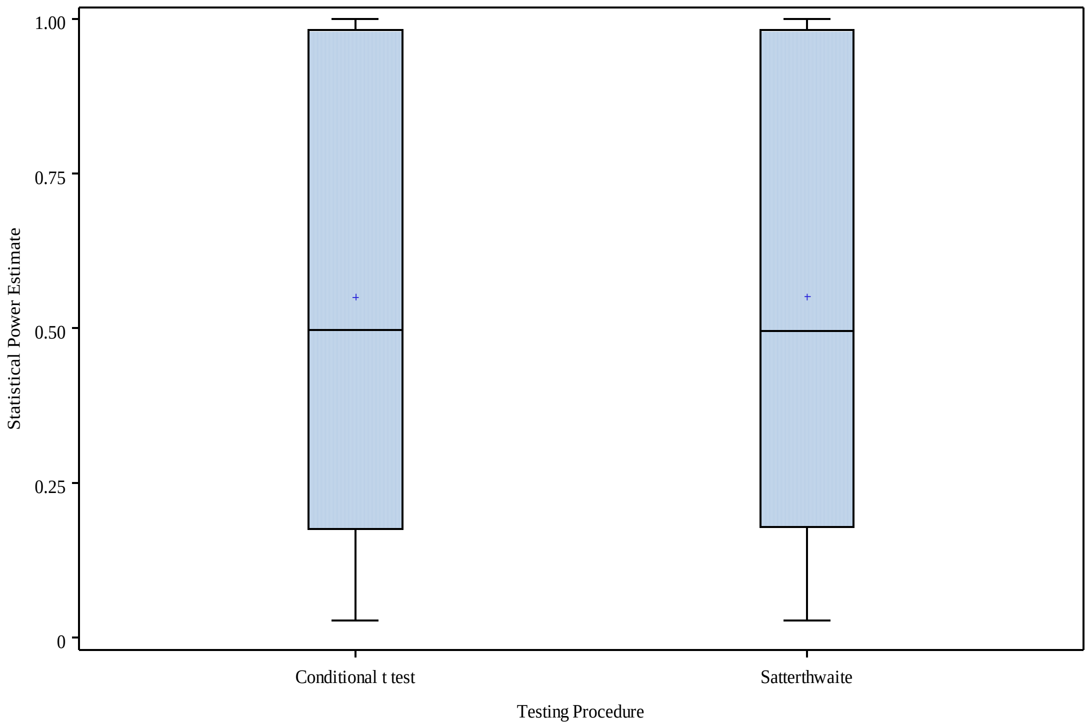

Figure 15: Distributions of estimated statistical power (Nominal alpha = .05)

Figure 16 presents a scatter plot of the power estimates for the conditional *t*-test and Satterthwaite's approximate *t*-test (using an alpha level of .25 for the Folded-*F* test). Data points above the line represent conditions in which the Satterthwaite test was more powerful than the conditional *t*-test, while those below the line are conditions in which the conditional *t*-test is more powerful. Overall, the conditional *t*-test, using alpha = .25 for the Folded *F*-test of variances, was more powerful in 29% of the conditions while Satterthwaite's approximate *t*-test was more powerful in only 23% of the conditions (identical power estimates were obtained in the other conditions).

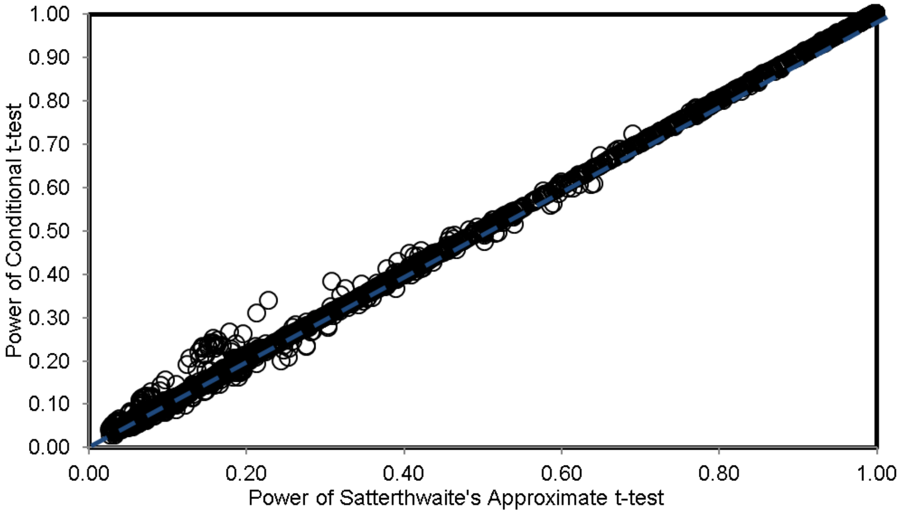

Figure 16: Scatterplot of power estimates for the conditional *t*-test and Satterthwaite's approximate *t*-test

To identify research design factors associated with power differences between these two tests, the percentages of conditions in which each test evidenced power advantages were disaggregated by the simulation design factors (Table 4). For conditions with homogeneous variances, the conditional *t*-test evidenced more power than the Satterthwaite's test in 61.64% of the conditions, while the Satterthwaite's test was more powerful in 20.55% of the conditions (in the remaining conditions, the two tests evidenced equal power). As the variance ratios increased, the power advantages of the conditional *t*-test diminished, such that the Satterthwaite's test was more often the more powerful test when the population variance ratio was 1:8 or larger.

Table 4: *Percentage of Simulation Conditions by Research Design Factors in which the Conditional t-test and Satterthwaite's Test Were More Powerful*

Condition	Conditional	Satterthwaite	Condition	Conditional	Satterthwaite
N			Variance ratio		
10	48.97	51.03	1:1	61.64	20.55
20	54.89	44.11	1:2	43.15	34.97
50	44.79	28.83	1:4	31.03	29.35
100	29.22	20.98	1:8	21.02	22.08
200	19.43	12.38	1:12	16.56	18.06
300	11.62	9.52	1:16	14.62	17.63
400	8.19	7.81	1:20	13.25	15.81
N ratio			Shape		
1:4	10.59	53.95	0,0	29.20	15.67
2:3	9.21	44.77	1,3	29.60	21.98
1	35.58	0.00	1.5,5	28.17	24.63
3:2	47.68	4.04	2,6	21.69	24.51
4:1	48.52	4.54	0,25	34.98	27.50

Note. Conditional = conditional *t*-test at $\alpha = .25$ of Folded *F*-test. For shape, the two values indicate skewness and kurtosis, respectively.

With balanced samples the conditional *t*-test was more powerful in 35.38% of the conditions and Satterthwaite's test was never more powerful. With unbalanced samples in which the larger sample is drawn from the population with the larger variance (in heterogeneous populations), the Satterthwaite's test presents notable power advantages (44.77% and 51.03% of the cases with sample size ratios of 2:3 and 1:4, respectively). In contrast, when the larger sample is drawn from the population with the smaller variance, the conditional *t*-test evidences more power than the Satterthwaite's test (47.68% and 48.52% of the conditions with sample size ratios of 3:2 and 4:1, respectively). The results by total sample size show that the conditional *t*-test is more powerful in more conditions, except for the smallest sample sizes examined ($N = 10$). Finally, the conditional *t*-test is more powerful in more conditions for all distribution shapes except for the most skewed distribution examined (skewness = 2, kurtosis = 6).

4. Conclusions and Recommendations

The testing of differences between two population means is a fundamental statistical application, but controversy about the appropriate test to use has been evident for many years. When conducting the independent means *t*-test, major statistical software programs (e.g, SAS and SPSS) automatically produce the results of the independent means *t*-test and the alternative Satterthwaite's approximate test. Depending on the statistical significance of homogeneous variance testing (Folded *F*-test in SAS and Levene's *F* test in SPSS), researchers are recommended to follow one of the options, which has been a common practice in studies comparing two population means. However, recent studies on the conditional *t*-tests in comparison to Satterthwaite's test have strongly supported the

Satterthwaite's test over the conditional *t*-test and suggested even abandoning the conventional practice of selecting one of the options based on the results of the homogeneity of variance test. Considering the ongoing controversy surrounding these tests and the frequency with which two means are compared in applied research, this simulation study was intended to investigate the performance of the independent means *t*-test, Satterthwaite's approximate *t*-test, and the conditional *t*-test under the manipulated conditions of population distribution shape, total sample size, sample size ratio between groups, variance ratio between populations, the difference in means between populations, alpha level for testing the treatment effect, and alpha level for testing the homogeneity assumption for the conditional *t*-test. Type I error control and power analysis were used to examine the performance of these testing procedures.

As expected, the independent means *t*-test performed very well on the Type I error control when the homogeneity assumption was met regardless of the tenability of the normality assumption. This reminds us of the long-known property that the independent means *t*-test requires the homogeneity assumption to be met and this test is robust to violations of the normality assumption when two population variances are equal. Furthermore, the independent means *t*-test showed adequate Type I error control when sample sizes in the two groups were equal under the normal distribution. This re-emphasizes another well-known property that the independent means *t*-test is robust to violations of the homogeneity assumption when the sample sizes are equal under the normal distribution. Under these conditions, the independent means *t*-test is the best method to test the difference between two independent means. This testing procedure also provides more statistical power. On the other hand, the *t* test evidenced poor Type I error control under heterogeneous variances with non-normal distributions. Thus, two alternatives, Satterthwaite's approximate *t*-test and the conditional *t*-test, were considered in this study.

This study also found that the Type I error rate of the conditional *t*-test was affected by the alpha level for the Folded *F*-test that was used to test the homogeneity assumption of population variances. The more conservative alpha levels for the Folded *F*-test resulted in larger Type I error rates for the conditional test because of lower statistical power, such that the Folded *F*-test might not be able to detect the true difference between population variances. This leads us to re-consider the conventional procedures for examining the difference between two population means. Thus, the conditional *t*-test with using a relatively large alpha level for the Folded *F*-test may be an appropriate alternative.

Overall, Satterthwaite's approximate test performed best in control of Type I error rate but the conditional *t*-test also yielded comparable results using a large alpha level of .25 for the Folded *F*-test. Both alternatives made a tremendous improvement in Type I error control, compared to the independent means *t*-test, when group variances were unequal. Extreme skewness (e.g., skewness = 2) contaminated the Type I error control for both alternative testing procedures. Kurtosis seemed not to have this kind of impact. Increasing total sample size was found in this study to improve Type I error control for both testing procedures, but not for the independent *t*-test. When total sample size was 200 or more, Bradley's rates were 100% for both alternative testing procedures. Although Satterthwaite's approximate *t*-test provides slightly better Type I error control, the use of the conditional *t*-test may have a slight power advantage.

So, what can we recommend for researchers on the selection of an appropriate test? First, with equal sample size the independent means *t*-test is the appropriate testing procedure to examine the difference of two independent group means because it provides adequate Type I error control and more statistical power. With unequal sample size the Folded *F*-

test can provide reasonable guidance in the choice between the independent *t*-test and Satterthwaite's approximate test. A large alpha level of .25 is recommended to evaluate the results of the Folded *F*-test. If the *F* value is not statistically significant at this large alpha level, then the independent means *t*-test should be used. In contrast, if the *F* value is statistically significant at this large alpha level, then Satterthwaite's approximate test should be chosen. Finally, the confidence in this conditional testing procedure increases as the sample sizes become larger. To adequately control for Type I error rate in the conditional testing procedure, a total sample size of at least 200 is recommended with extremely skewed populations (e.g., skewness = 2). For less skewed populations, a total sample size of at least 100 is recommended. With a total sample size smaller than these recommended in the corresponding conditions, the Type I error control resulting from any of these testing procedures may be questionable.

References

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

Bridge, P. D., & Sawilowsky, S. S. (1999). Increasing physicians' awareness of the impact of statistics on research outcomes: Comparative power of the t-test and Wilcox Rank-Sum test in small samples applied research. *Journal of Clinical Epidemiology, 52*(3), pp. 229-235.

Gravetter, F. J., & Wallnau, L. B. (2011). *Essentials of Statistics for the Behavioral Sciences* (6th Ed.). Belmont, CA: Wadsworth, Cengage Learning.

Grissom, R. J. (2000). Heterogeneity of variance in clinical data. *Journal of Consulting and Clinical Psychology*, *68*, 155-165.

Hayes, A. F. & Cai, L. (2007). Further evaluating the conditional decision rule for comparing independent means. *British Journal of Mathematical and Statistical Psychology*, *60*, 217-244.

Heiman, G. W. (2011). Basic Statistics for the Behavioral Sciences (6th Ed.). Belmont, CA: Wadsworth Cengage Learning.

Moser, B. K., Stevens, G. R., & Watts, C. L. (1989). The two-sample t-test versus Satterthwaite's approximate F test. *Communications in Statistics: Theory and Methods*, *18*, 3963-3975.

Rasch, D., Kubinger, K. D., & Moder, K. (2011). The two-sample t-test: pre-testing its assumptions does not pay off. Statistical Papers, 52, 219-231.

Robey, R. R., & Barcikowski, R. S. (1992). Type I error and the number of iterations in Monte Carlo studies of robustness. *British Journal of Mathematical and Statistical Psychology, 45*, 283-288.

Sawilowsky, S. S. & Blair, R. C. (1992). A more realistic look at the robustness and type II error properties of the t test to departures from population normality. *Psychological Bulletin, 111*, 352–360.

Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing* (2nd edition). San Diego, California: Academic Press.

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