A Simulation Study of the Independent Means *t*-test, Satterthwaite's Approximate *t*-test and the Trimmed *t*-test Under Normal and Non-normal Distributions

Anh P. Kellermann, Diep Nguyen, Patricia Rodríguez de Gil,
Eun Sook Kim, Jeffrey D. Kromrey
University of South Florida, 4202 E. Fowler Avenue, Tampa, FL 33620

Abstract
While the independent means *t*-test is popular for testing the equality of two population means, it is sensitive to violations of the population normality and homogeneity of variance assumptions. Satterthwaite's approximate t-test and Yuen's trimmed *t*-test are recommended as robust alternatives which relax those assumptions. This simulation study compared the performance of the *t*-test, Satterthwaite's *t*-test, and the trimmed *t*-test under normal and non-normal distributions. The two latter tests were conducted both unconditionally and conditionally on a preliminary test of variances. Simulation conditions manipulated were total sample size, group sample size ratio, population variance ratio, population effect size, alpha sets for both the treatment effect and group variances tests, and population distribution shape. As expected, the independent means *t*-test showed great dispersion of Type I error rates. The other tests (both conditional and unconditional) evidenced notable improvement in Type I error control relative to the independent means *t*-test. Power comparisons (for conditions in which Type I error control was adequate) were used to identify the most powerful test among the set.

Key Words: heteroscedasticity, non-normality, Independent means *t*-test, Satterthwaite approximate *t*-test, trimmed means, simulation

1. The Independent Means *T*-test and Alternatives

1.1 Introduction

The independent means *t*-test is popular for testing the equality of two population means. However, it is sensitive to violations of the population normality and homogeneity of variance assumptions. If the data for one or both of the samples violates these assumptions, then the *t*-test may provide misleading results. Satterthwaite's approximate *t*-test and Yuen's trimmed *t*-test are recommended as robust alternatives which relax those assumptions. The purpose of the present paper is to compare the Type I error control and power of the trimmed *t*-test against the independent sample *t*-test and Satterthwaite's approximate test under normal and non-normal distributions.

1.2 Problems with the independent means *t*-test

Non-normality and Heterogeneity of Variance. Testing for the equality of means across independent groups is "a common inferential problem" (Keselman, Wilcox, Lix, Algina, & Fradette, 2007; p. 267). The independent means *t*-test relies on a strong assumption of equal variances (homoscedasticity) as the test statistic is a ratio of the difference in sample means to an estimate of the standard error of the difference, using a pooled variance estimate. Alternative approaches (e.g., Satterthwaite's approximate test)

relax this assumption, approximating the *t* distribution and the corresponding degrees of freedom. Although the *t*-test may be one of the most basic and widely used statistical procedures to compare two group means (Hayes & Cai, 2007), 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. Research on preliminary tests suggests that the choice between the *t*-test and the Satterthwaite's test, conditioning on a preliminary test of the assumption of homogeneity of variance, is not effective (Grissom, 2000; Hayes & Cai, 2007; Moser, Stevens, & Watts, 1989; .Rasch, Kubinger, & Moder, 2011; Zimmerman, 2004, 2010).

Keselman, Wilcox, Othman, and Fradette (2002) have suggested the use of trimmed means to achieve robustness in the presence of non-normality and variance heterogeneity. Lix and Keselman (1998) studied the performance of the Welch test (a close relative of Satterthwaite's test) in addition to the performance of the Alexander and Govern, James, and Brown-Forsythe tests for testing mean equality in the presence of unequal variances. These tests can generally control Type I error rate when group variances are heterogeneous and data are normally distributed. However, these tests become liberal when the assumptions of normality and homogeneity of variances are violated, and they become even more liberal with unbalanced groups. For all the investigated distributions in the Lix and Keselman study, a symmetric trimming was applied by removing 20% of the observations from each tail of the groups' set of scores. Their results showed that the studied methods generally exhibited a very good Type I error control rate when computed with trimmed means and Winsorized variances. Using a one-way completely randomized experiment, Keselman, Wilcox, Algina, Fradette, and Othman (2004) compared seven methods known to be robust to the effects of non-normality and variance heterogeneity. Six methods (WJ or Welch-James-type heteroscedastic tests) known to provide good Type I error control and power (Algina & Keselman, 1998), using either symmetric or asymmetric trimming, Winsorized means and variances, were applied. The power of these tests was compared to the power of the one-step-M-Estimator trimmed means (MOMT; Wilcox & Keselman; 2003), test for the detection of treatment effects. Preliminary power results showed minor differences between the WJ tests due to data transformation or sample size. However, there were power differences favoring the WJ tests over the MOMT (.13).

Robust methods such as modified F_t and modified S_1 have been also recommended to overcome the sensitivity of the *t*-test to variance heterogeneity. Yusof, Abdullah, Yahaya, and Othman (2012) proposed the use of the trimmed mean, as a central tendency measure in the F_t test, and the median as the central tendency measure in S_1 when comparing the equality of two groups. These methods were compared in terms of Type I error under conditions of normality and non-normality represented by skewed-*g* and *h*-distributions.

Nguyen, Rodríguez de Gil, Kim, Bellara, Kellermann, Chen, and Kromrey (2012) conducted a simulation study to investigate the performance of the independent means *t*-test, Satterthwaite's approximate *t*-test, and the conditional *t*-test in terms of Type I error control and statistical power. Factors manipulated in the study included total sample size, sample size ratio between groups, variance ratio between groups, population effect size, alpha for testing treatment effect, and alpha for testing the homogeneity of variance. For each condition, 100,000 replications were simulated, which provided a maximum standard error of an observed proportion (e.g., Type I error rate estimates) of .0015, and a 95% CI no wider than ± .003 (Robey & Barcikowski, 1992). Overall, the Satterthwaite's approximate *t*-test performed best in control of Type I error rates under all conditions.

Results indicated that to maintain adequate Type I error control, the independent means *t*-test required that the homogeneity of variance assumption was met in addition to equal sample size between groups, regardless of the tenability of the normality assumption. The alpha level used for the Folded *F* test had an impact on Type I error control for the conditional *t*-test. The more conservative the alpha level, the larger Type I error rate. Because of lower statistical power of the Folded *F* test, the study recommended the conditional *t*-test using relatively large alpha levels for the test of variances. The results also showed that an increase in total sample size did not improve the control of Type I error rate for the independent means *t*-test, but larger samples provided better Type I error control for the conditional *t*-test.

Kellermann, Bellara, Rodríguez de Gil, Nguyen, Kim, Chen, and Kromey (2013) extended the Nguyen et al. (2012) study to investigate the performance of the *t*-test, Satterthwaite's approximate *t*-test, and conditional *t*-test under heteroscedastic populations. In addition to the normal population, four non-normal populations were studied, with varying values of skewness and kurtosis ($\gamma 1 = 1.00$, $\gamma 2 = 3.00$; $\gamma 1 = 1.50$, $\gamma 2 = 5.00$; $\gamma 1 = 2.00$, $\gamma 2 = 6.00$; $\gamma 1 = 0.00$, $\gamma 2 = 25.00$) respectively. Findings were similar to the Nguyen et al. (2012) results with normal populations. Both the Satterthwaite's and conditional *t*-tests provided tremendous improvements in Type I error control compared to the independent means *t*-test when group variances were unequal. However, extreme skewness contaminated the Type I error control for these tests. On the other hand, kurtosis did not seem to have the same effect. Increasing sample size ($n \geqslant 200$) helped improve the Type I error control for the Satterthwaite's and conditional tests, but not for the independent *t*-test.

1.3 The Trimmed *T*-test

Yuen (1974) proposed the *Trimmed t*-test for the independent two-sample case, under unequal population variances. In each sample, the trimmed mean is computed by removing *g* observations from each tail of the distribution:

$$\bar{X}_t = \frac{1}{n-2g}\left(x_{g+1} + x_{g+2} + \ldots + x_{n-g}\right),$$

where
$x_1,\ldots,x_n$ are the ordered values in a sample
g = observations trimmed from each tail of the sample distribution
$n - 2g$ = the number of observations in the trimmed sample.

In addition to the trimmed mean, the Winsorized mean is required to compute the appropriate variance estimate. Instead of "trimming," this method replaces the most extreme *g* observations by the next-most-extreme value.

$$\bar{X}_w = \frac{1}{n}\left(\left[g+1\right]x_{g+1} + x_{g+2} + \ldots + \left[g+1\right]x_{n-g}\right).$$

Given the Winsorized mean, the Winsorized sum-of-squared deviations is computed as:

$$SSD_w = \left[g+1\right]\left[x_{g+1} - \bar{X}_w\right]^2 + \left[x_{g+2} - \bar{X}_w\right]^2 + \ldots + \left[g+1\right]\left[x_{n-g} - \bar{X}_w\right]^2.$$

Note that this is just the regular sum-of-squares approach using the replaced values and the Winsorized mean. From the Winsorized sum-of-squared deviations, the Winsorized variance is obtained as:

$$S_w^2 = \frac{SSD_w}{n - 2g - 1}.$$

Finally, the obtained value of the trimmed *t* is obtained by dividing the difference between the trimmed means by the estimated standard error of the difference:

$$t = \frac{\bar{X}_{t1} - \bar{X}_{t2}}{\sqrt{\frac{S_{w1}^2}{n_1 - 2g} + \frac{S_{w2}^2}{n_2 - 2g}}}$$

The degrees of freedom are obtained from $\frac{1}{df} = \frac{c^2}{n_1 - 2g - 1} + \frac{(1-c)^2}{n_2 - 2g - 1}$

where

$$c = \frac{S_{w1}^2 \Big/ (n_1 - 2g - 1)}{\left[S_{w1}^2 \Big/ (n_1 - 2g - 1)\right] + \left[S_{w2}^2 \Big/ (n_2 - 2g - 1)\right]}$$

Generally, the Welch and Satterthwaite approximate *t*-tests become conservative with leptokurtic distributions (Yuen, 1974). However, there is caveat for its use with small samples. The trimmed *t* (Yuen, 1974) is recommended instead because of its advantages (e.g., easy to compute and critical values from the standard *t* table can be used). Yuen (1974) conducted a study to determine whether the use of trimmed means and Winsorized variances resulted in robust tests for mean equality. Variables manipulated included sample sizes (10 or 20), standard deviation ratios (0.25, 0.5, 2.0 and 4.0), trimming rate (*g*) (from 1 observation to .25n_j observations), and a variety of distribution shapes. For unequal sample sizes, the amount of trimming was in fixed proportions. Ten thousand replications per condition were generated. Results showed a Type I error control for the trimmed means closer to the nominal alpha level than those obtained with the Welch's test, although some still deviated notably from the nominal level. Yuen suggested an adaptive trimming approach; that is, the number of observations trimmed (*g*) should be chosen depending on the degree of leptokurtosis.

2. Method

A crossed factorial mixed design included seven between-subjects factors: (a) total sample size (10, 20, 50, 100, 200, 300, 400), (b) sample size ratio between groups (1:1, 3:2, 4:1, 2:3, 1:4), (c) variance ratio between groups (1:1, 1:2, 1:4, 1:8, 1:12, 1:16, 1:20), (d) population effect size (Δ = 0, .2, .5, .8), (e) alpha for testing treatment effect (α =.01, .05, .10, .15, .20, .25), (f) alpha for testing the homogeneity of variance for Folded *F* test (α = .01, .05, .10, .15, .20, .25, .30, .35 .40, .45, .50), (g) percent trimming (5, 10, 20) for

trimmed *t*-test, and one within-subjects, (h) population distributions with varying kurtosis and skewness values ($\gamma 1 = 1.00$, $\gamma 2 = 3.00$; $\gamma 1 = 1.50$, $\gamma 2 = 5.00$; $\gamma 1 = 2.00$, $\gamma 2 = 6.00$; $\gamma 1 = 0.00$, $\gamma 2 = 25.00$; plus the normal distribution, $\gamma 1 = 0.00$, $\gamma 2 = 0.00$). This crossed factorial design (7 x 5 x 7 x 4 x 6 x 11 x 3 x 5) provided a total of 970,200 conditions. For each data generation condition, 100,000 replications were generated. We majorly examined Type I error and statistical power as the simulation outcomes. For Type I error, we further investigated the robustness of Type I error control using the Bradley's liberal criterion (for example, .between 025 and .075 when the significance level is .05). Also in comparing the performance of the *t*-test, Satterthwaite's *t*-test, and the trimmed *t*-test, the Satterthwaite's *t*-test and the trimmed *t*-test were conducted both unconditionally and conditionally on a preliminary test of variances.

3. Results

3.1 Type I Error Control for the Tests of Means

Figure 1 displays the distribution of Type I error rates of the independent means *t*-test, Satterthwaite's test, and the trimmed *t*-test at alpha = .05. Satterthwaite's test controlled Type I error around the predetermined alpha level. On the other hand, the independent means *t*-test showed a considerable variability in Type I error rates. The Type I error rates of the trimmed *t*-test depended on the degree of trimming. As more observations were trimmed (from 5% to 20%), the overall behavior of the trimmed *t*-test deteriorated showing larger variability in Type I error control. For the trimmed *t*-test, the majority of simulation conditions including large sample and nonnormal conditions yielded Type I error rates over the nominal alpha level.

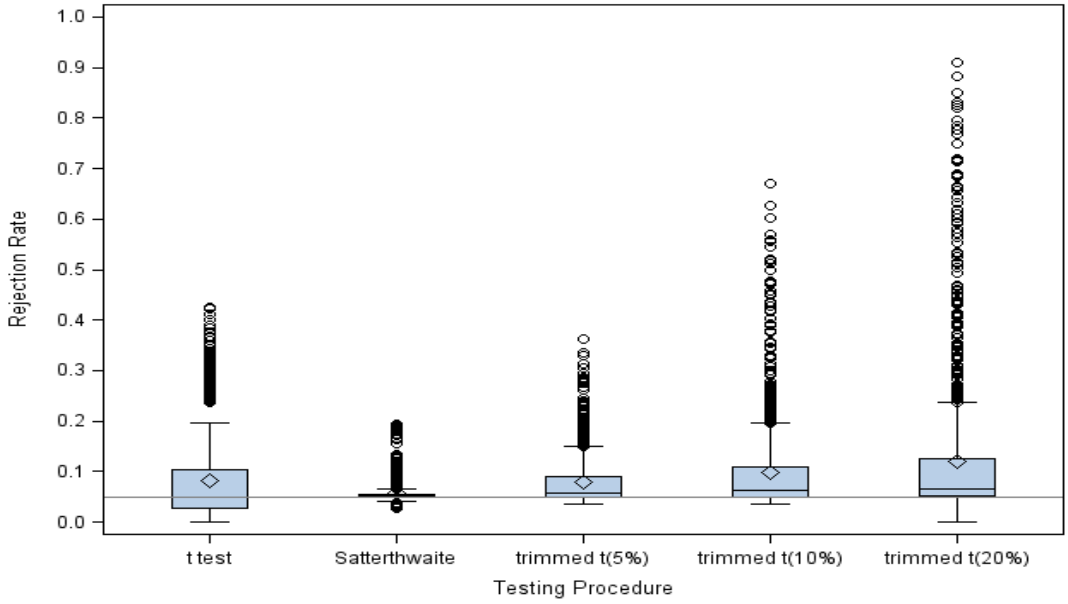

Figure 1: Distribution of estimated Type I error rates of the unconditional independent means *t*-test, Satterthwaite's test, and trimmed *t*-test at $\alpha = .05$

The performance of the two conditional tests was investigated in comparison to the independent means *t*-test and Satterthwaite's *t*-test. Figure 2 shows the performance of

the conditional *t*-test, in which the rejection of homogeneous variance lead to Satterthwaite's test instead of the independent means *t*-test, becoming comparable to that of Satterthwaite's test as the alpha for testing the homogeneity of variance increased over .20.

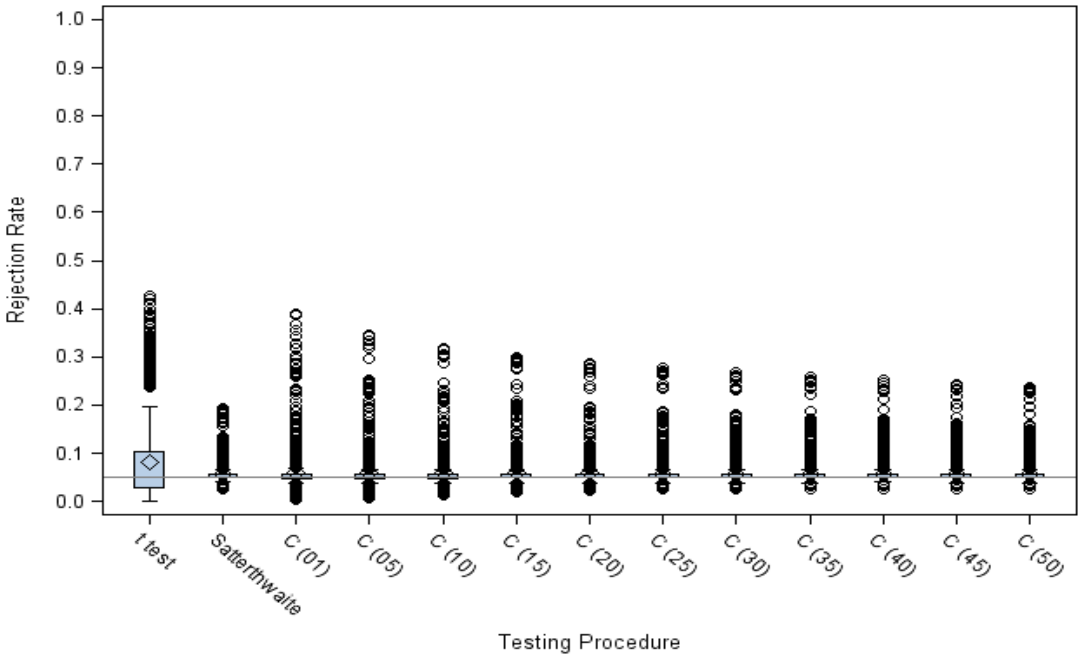

Figure 2: Distribution of estimated Type I error rates of the independent means *t-t*est, Satterthwaite's test, and conditional *t*-test at α = .05. C(01) means the conditional *t*-test when the alpha for testing the homogeneity of variance was .01

Figure 3 shows the performance of the conditional trimmed *t*-test, in which the rejection of homogeneous variance leads to the trimmed *t*-test instead of the independent means *t*-test, was not impacted by the significance level set for the homogeneity of variance test. In other words, conditional decisions between the independent means *t*-test and trimmed *t*- test yielded inflated Type I error rates regardless of the power of Folded F test when 5% of the observations were trimmed. Similar patterns emerged with apparently increased variability in Type I error rates for 10% and 20% trimming.

3.2 Factors that Influence Type I Error Control

The simulation design factors were examined for their influences on Type I error rates for all the tests of means in this study, i.e. the independent means *t*-test, Satterthwaite's test, trimmed *t*-test, conditional Satterthwaite's test and conditional trimmed *t*-test. The eta-squared value associated with each factor and the first-order interactions were computed to analyze the variability in the estimated Type I error rates of these tests.

For the Satterthwaite's test and conditional Satterthwaite's test, the major factor associated with variability in estimated Type I error rates was the interaction of sample size ratio and total sample size. Figure 4 and Figure 5 demonstrate that for the both tests at α =.01, Type I error rate estimate was liberal when total sample size decreased and became highly inflated with very small sample size of 10 and an unequal sample size ratio of 1:4.

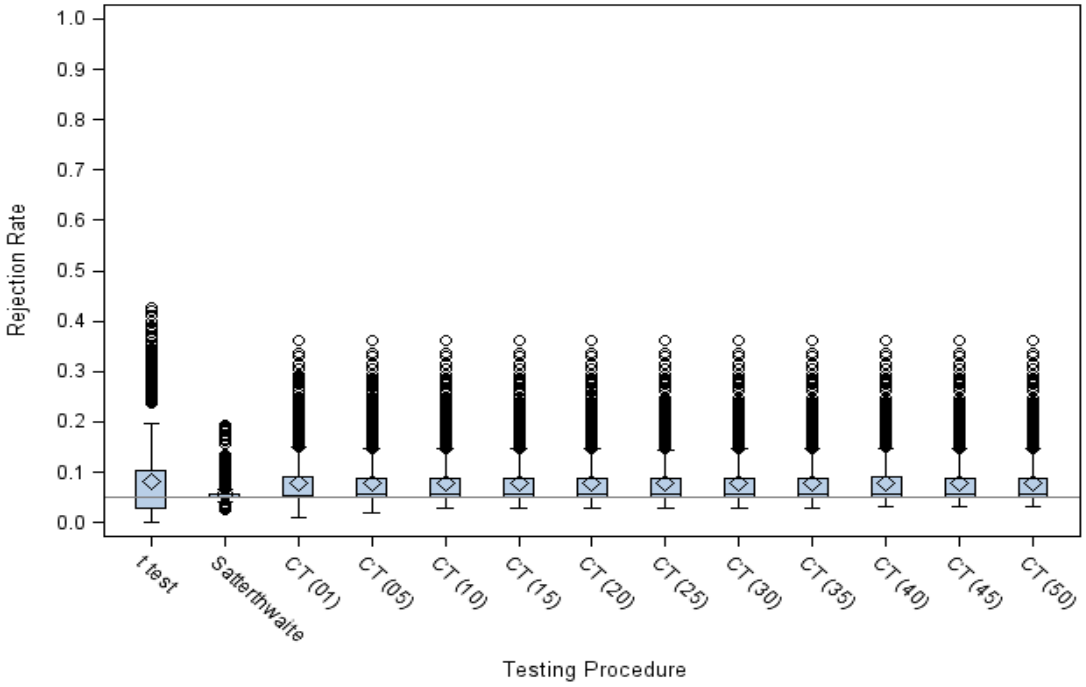

Figure 3: Distribution of estimated Type I error rates of independent means *t*-test, Satterthwaite's test, and conditional trimmed *t*-test with trimming at 5% across all alpha levels. CT (01) means conditional trimmed *t*-test with 5% trimming when the alpha for testing the homogeneity of variance equals .01

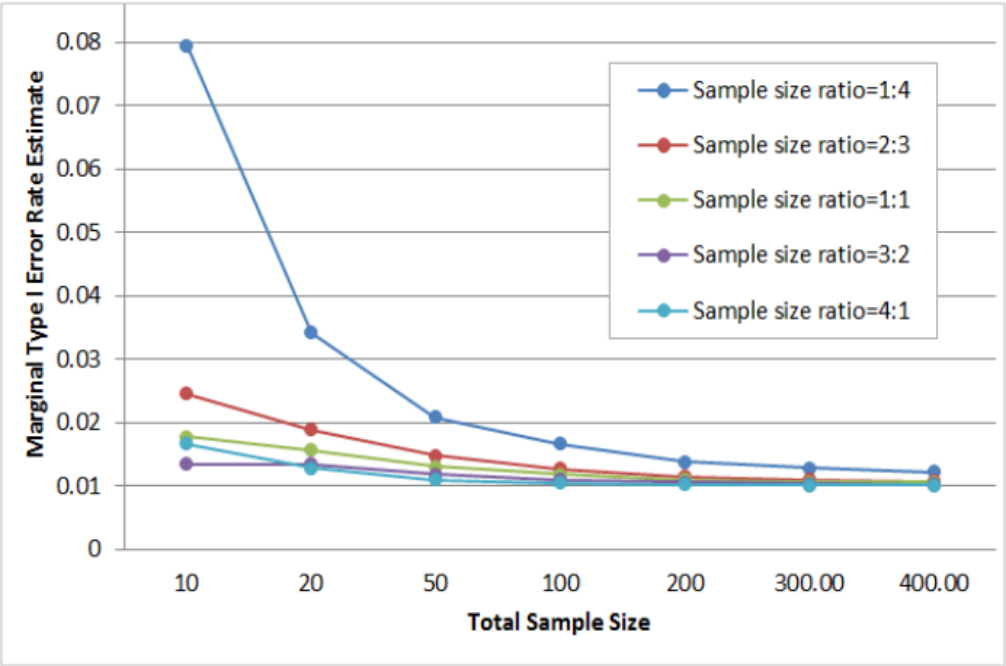

Figure 4: Marginal Type I error rate by sample size ratio and total sample size at α = .01 for the test of means for the Satterthwaite's test

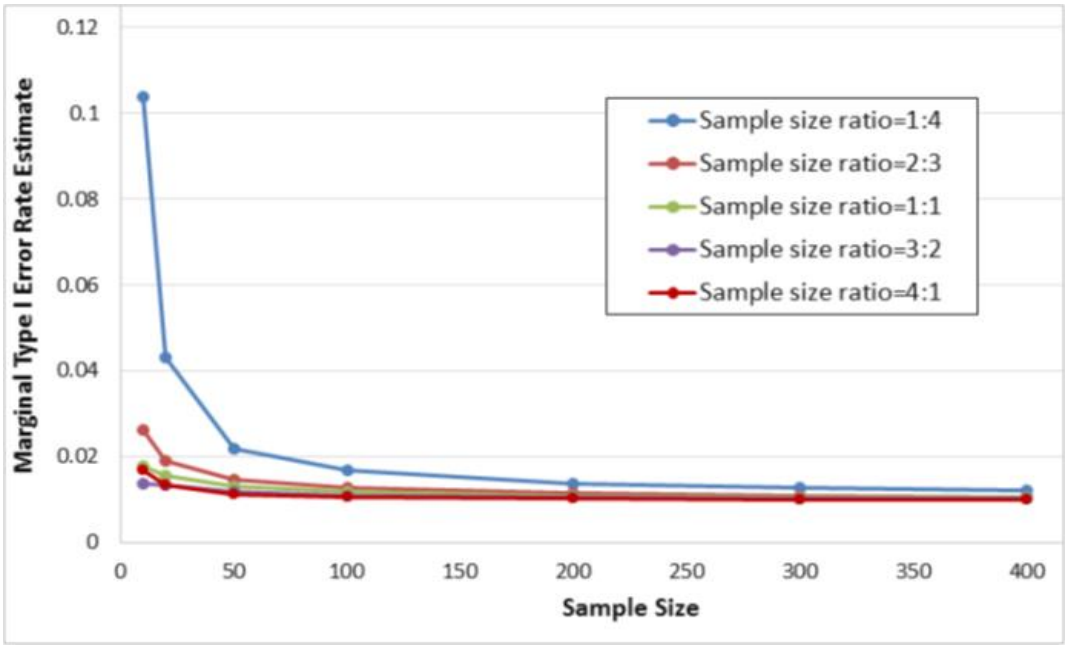

Figure 5. Marginal Type I error rates by sample size and sample size ratio for conditional Satterthwaite's test at α= .01 for the test of means and α=.50 for test of variance

The population shape has the most influence on Type I error control for the trimmed *t*-test. The more data are trimmed, the higher inflated rate of Type I error rates become (see Figure 6).

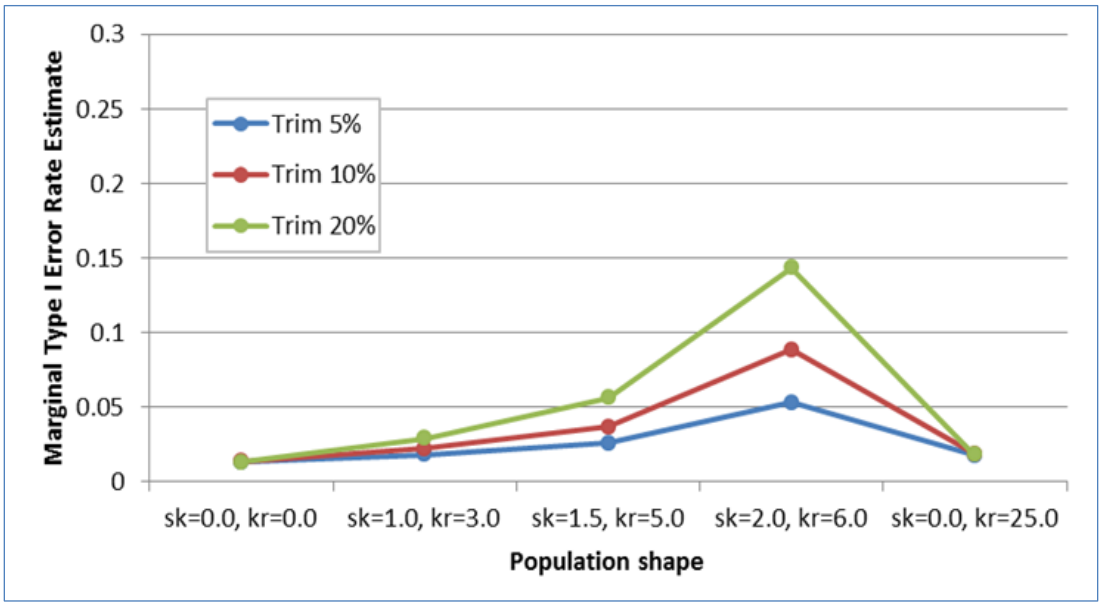

Figure 6: Marginal Type I error rate by shape at α =.01 for the test of means for the regular trimmed *t*-test with trim=5%, 10%, and 20%

Unlike skewness which was found to have a big influence on Type I error rate estimates of the conditional trimmed *t*-test, kurtosis doesn't show to have much effect regarding this type of error control. Under the normal distribution condition, the conditional trimmed *t*-tests were robust regardless of the trimming levels, but with an elevated skewness the more the data were trimmed the more liberal the conditional trimmed *t*-tests were. Figure 7 depicts Type I error rate estimates across five different distribution shapes for the conditional trimmed *t*-test at trimming levels of 5%, 10%, and 20%. All three tests performed well under the normal and not-skewed condition whereas the poorest performance occurred at the highest level of skewness (sk=2.0) in which Type I error rate estimate for the 20%-trimmed *t*-test was double that of the 5%-trimmed *t*-test.

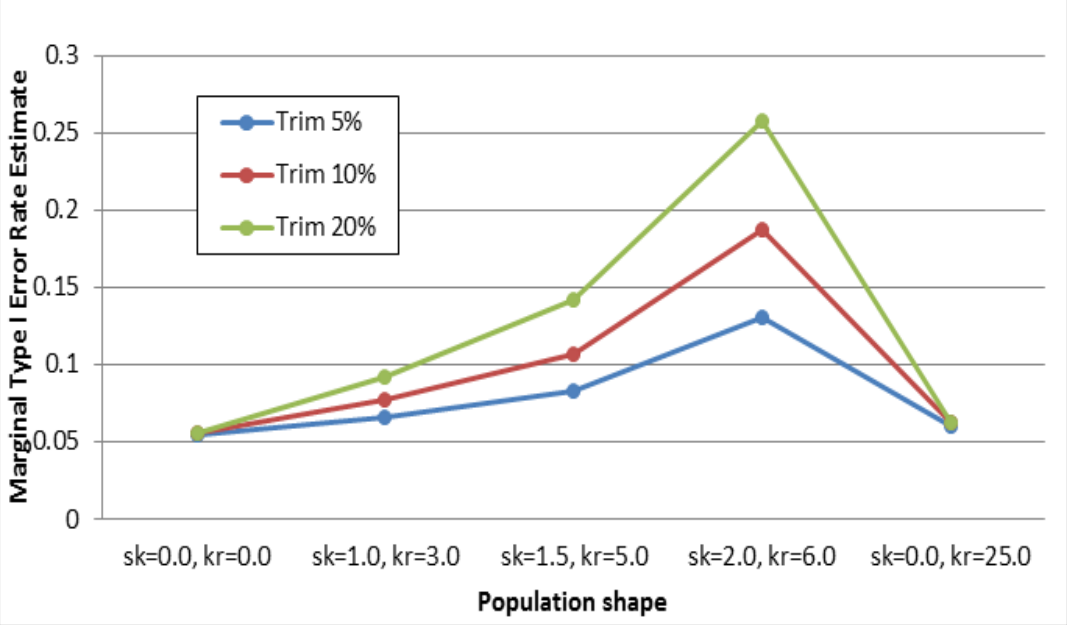

Figure 7: Mean Type I error rates by population shape at α =.05 for the test of means, α=.25 for the test of equal variances

It is found that heterogeneity of variance did not influence Type I error control much as long as the data were normal or had no skewness; however, Type I error was out of control with elevated kurtorsis; that is, Type I error rate increased as the heterogeneity of variance levels increased. Figure 8 illustrates Type I error rate estimate for the conditional 10%-trimmed *t*-test in five different distribution shapes across seven population variance ratios. The same as with the independent *t*-test, the conditional trimmed *t*-test is robust in normal distribution and homogeneity conditions, but those tests with elevated kurtosis were slightly conservative in the homogeneous variance condition and became very liberal as the variance ratio increased.

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

Table 1 presents the proportion of conditions meeting the Bradley's liberal criterion for Type I error control. The Satterthwaite's *t*-test always met the liberal criterion when the total sample size exceeded 200. Unless the total sample size was very small such as 10, Satterthwaite's test showed reasonable performance in controlling Type I error within the

Bradley's criterion. The trimmed *t*-test in general outperformed the independent means *t*-test across the total sample size conditions. Interestingly, the trimmed *t*-test performed better with the total sample size about 50 and 100 than with larger samples. Even when the total sample size was very small (i.e., 10), the proportion of conditions meeting the Bradley's criterion was over 60% except the conditions of large degree of trimming (20%). Similar patterns were observed with the conditional trimmed *t*-test regardless of the alpha set for testing the homogeneity of variance.

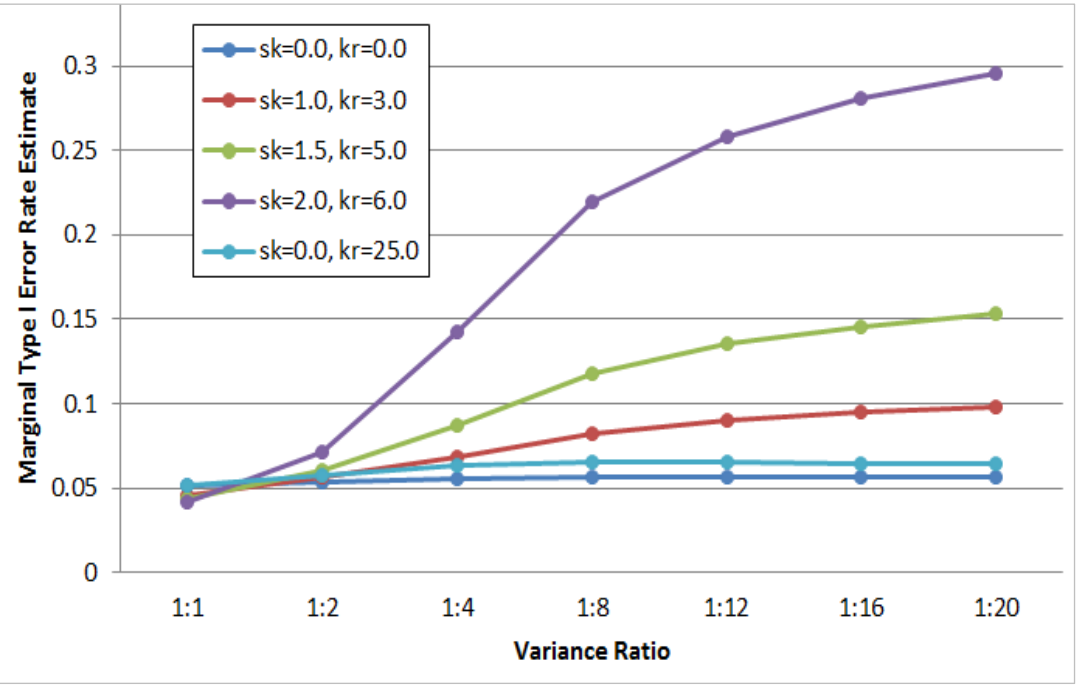

Figure 8: Mean Type I error rates by variance ratio and population shape at α=.05 for the tests of means, α=.25 for the test of equal variances, and trim = 10%.

3.4 Statistical Power Analysis

We conducted power analysis only with the conditions meeting the Bradley's liberal criterion because high power associated with high Type I error is not compelling. Overall, the Satterthwaite's *t*-test on average produced higher power followed by the trimmed *t*-test and the independent means *t*-test. For the trimmed *t*-test, more trimming with the loss of observations possibly leads to lower power as demonstrated in Figure 9.

4. Conclusions and Recommendations

Using a simulation approach and a crossed factorial research design, the purpose of this study was to investigate the Type I error control and statistical power of the independent means *t*-test and alternatives (Satterthwaite's approximate *t*-test, Yuen's Trimmed *t*, as well as conditioning the *t*-test and the trimmed *t* on a preliminary test of variance) under normal population distribution and equal variances, and under violations of the normality and homogeneity of variance assumptions.

In agreement with previous research on the robustness of the *t*-test and alternatives (Nguyen et al. 2013 and Kellermann et al. 2014), this study found that the *t*-test is very sensitive to violations of the normality and homogeneity of variance assumptions. While preliminary tests of the homogeneity of variance assumption do not seem to help to control Type I error rates when sample size between groups differs, conditional tests are recommended when the assumption of equal variances is rejected. The Satterthwaite's approximate *t*-test and conditional trimmed *t* had better control of Type I error rates than the *t*-test and trimmed *t*.

Table 1. Proportions of Cases Meeting the Bradley's Liberal Criterion by Tests and Conditions at $\alpha = .05$

Condition	*t-test*	*Conditional*	*Satterthwaite's*	*Trimmed t-test*			*Conditional Trimmed t-test*		
N				5%	10%	20%	5%	10%	20%
10	0.47	0.68	0.65	0.62	0.62	0.56	0.51	0.51	0.63
20	0.49	0.76	0.81	0.59	0.54	0.45	0.64	0.60	0.50
50	0.45	0.93	0.94	0.77	0.67	0.53	0.75	0.69	0.56
100	0.44	0.97	0.97	0.74	0.64	0.59	0.74	0.65	0.59
200	0.42	1.00	1.00	0.69	0.59	0.55	0.69	0.59	0.54
300	0.41	1.00	1.00	0.63	0.56	0.52	0.63	0.56	0.52
400	0.41	1.00	1.00	0.59	0.56	0.52	0.59	0.55	0.52
N ratio									
1:4	0.14	0.71	0.73	0.56	0.48	0.39	0.57	0.49	0.41
2:3	0.29	0.91	0.91	0.59	0.55	0.56	0.59	0.55	0.49
1	0.92	0.97	0.97	0.69	0.62	0.47	0.66	0.59	0.58
3:2	0.67	0.98	0.98	0.75	0.67	0.62	0.73	0.66	0.60
4:1	0.18	0.98	0.97	0.73	0.65	0.60	0.76	0.70	0.65
Variance ratio									
1:1	1.00	0.95	0.92	0.94	0.91	0.88	0.98	0.99	0.99
1:2	0.62	0.93	0.94	0.88	0.84	0.73	0.91	0.87	0.76
1:4	0.40	0.91	0.93	0.72	0.60	0.52	0.70	0.59	0.52
1:8	0.29	0.89	0.91	0.58	0.49	0.42	0.56	0.47	0.41
1:12	0.27	0.89	0.89	0.53	0.44	0.39	0.52	0.42	0.39
1:16	0.26	0.89	0.89	0.50	0.44	0.39	0.49	0.42	0.39
1:20	0.23	0.89	0.89	0.49	0.44	0.37	0.47	0.42	0.36
Shape									
0,0	0.43	0.96	0.97	0.94	0.94	0.92	0.95	0.95	0.93
1,3	0.44	0.96	0.97	0.77	0.61	0.41	0.77	0.59	0.42
1.5,5	0.45	0.93	0.93	0.44	0.36	0.31	0.42	0.36	0.31
2,6	0.46	0.77	0.77	0.28	0.22	0.19	0.29	0.24	0.23
0,25	0.42	0.91	0.91	0.89	0.84	0.81	0.88	0.85	0.84

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

Among the factors studied, sample size was a determinant factor for sustaining adequate Type I error control. Under the normal distribution, adequate Type I error control was observed for larger sample sizes while smaller sample sizes and unbalanced groups (e.g., sample size ratio 1:4) were associated with greater Type I error rates. The significance level appeared to have little effect on the Type I error control for the conditional trimmed *t* but extreme variance ratios did.

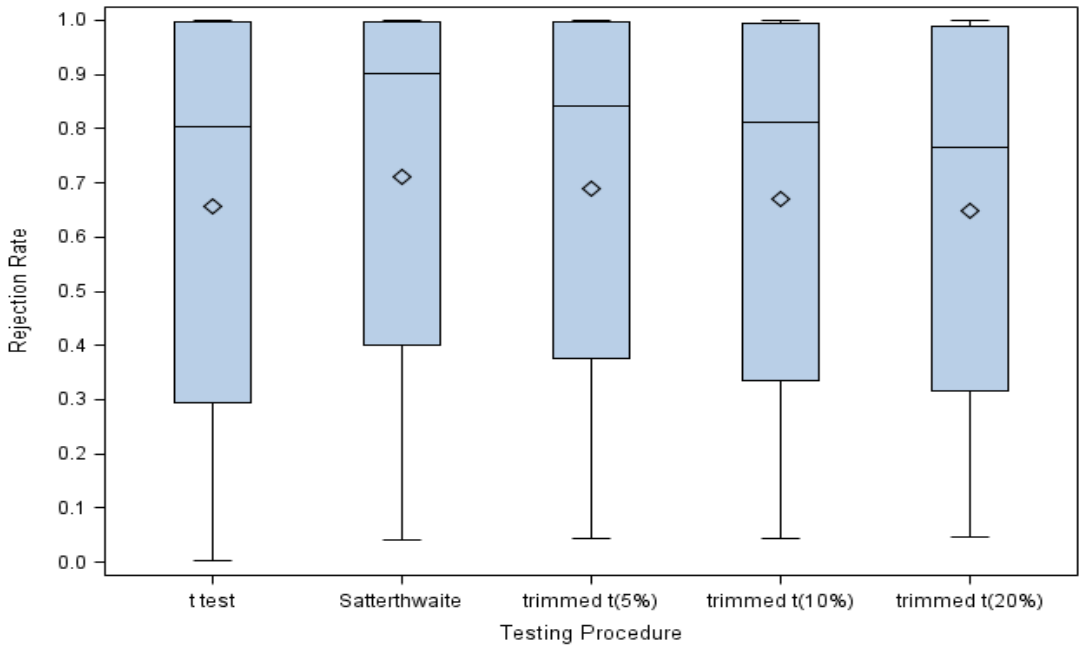

Figure 9: The distribution of power of the independent means *t*-test, Satterthwaite's *t*-test, and trimmed *t*-test at $\alpha = .05$.

While the trimmed *t* test offers the advantage of being easy to compute and the critical values from the standard *t*-test can be used, researchers need to be careful with small sample sizes and the number of observations trimmed should be chosen depending on the degree of skewness.

The conditional *t*-test or Satterthwaite's test could be used without a preliminary test of variance if the samples are normally distributed, regardless heterogeneity of variance and total sample size since even small total samples showed adequate Type I error control. However, careful decisions need to be taken for unbalanced samples. While it seems that unequal variances did not affect the Type I error rates of these tests under the normal distribution, whether the samples are balanced (e.g., N ratio 1:1) had an effect on the performance of these tests and a relationship with variance ratio was also observed since the proportion of conditions meeting the Bradley's criterion for Type I error control decreased as the variance ratio between groups increased.

In conclusion, both the conditional *t*-test and the Satterthwaite's outperformed the *t*-test, trimmed *t* and conditional trimmed *t*, which showed no improvement in their Type I error control regardless sample size, balanced samples or percentage of trimming. However,

the trimmed *t* and *t*-test showed a slightly higher statistical power than the conditional *t*-test. Although the trimmed *t* and conditional trimmed *t* tests showed similar performance as the conditional *t*-test and Satterthwaite's when the equality of variance and normality assumptions were met, the proportion of conditions meeting the Bradley's criterion decreased as the percentage of trimming increased across all factors. Consequently, the use of the conditional *t*-test and Satterthwaite's is recommended over the trimmed *t* and conditional trimmed *t*.

References

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

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.

Kellermann, A. P, Bellara, A. P., Rodriguez de Gil, P. , Nguyen, D., Kim, E. S., Chen, Y-H, & Kromey, J. D. (2013). Variance heterogeneity and Non-Normality: How SAS PROC TTEST® can keep us honest. Proceedings of the Annual SAS Global Forum Conference, Cary, NC: SAS Institute Inc.

Keselman, H. J., Othman, A. R., Wilcox, R. R., and Fradette, K. (2004). The new and improved two-sample t-test. Psychological Science, 15(1), 47-51.

Keselman, H. J., Wilcox, R. R., Algina, J., Fradette, K., & Othman, A. R. (2004). A power comparison of robust test statistics based on adaptive estimators. Journal of Modern Applied Statistical Methods, 3(1), 27-38.

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

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

Keselman, H. J., Wilcox, R. R., Lix, L. M., Algina, J., and Fradette, K. (2007). Adaptive robust estimation and testing. British Journal of Mathematical and Statistical Psychology, 60, 267-293.

Lix, L. M., and Keselman, H. J. (1998). To trim or not to trim: Test of location equality under heteroscedasticity and nonnormality. Educational and Psychological Measurement, 54(3) 409-429.

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.

Nguyen, D., Rodriguez de Gil, P. , Kim, E. S., Bellara, A. P., Kellermann, A. P., Chen, Y-H., & Kromrey, J. D. (2012). PROC TTest® (Old Friend), What are you trying to tell us?. Proceedings of the South East SAS Group Users, Cary, NC.

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.

SAS Institute Inc. (2008). *SAS, release 9.2* [computer program]. Carey, NC: SAS Institute Inc.

Yuen, K. K. (1974). The two-sample trimmed t for unequal population variances. Biometrika, 61(1), 165-170.

Yusof, Z., Abdullah, S., Yahaya, S. S. S., Othman, A. R. (2012). A robust alternative to the t-test. Modern Applied Science, 6(5), 27-33.

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

Zimmerman, D. W. (2010). A simple and effective decision rule for choosing a significance test to protect against non-normality. British Journal of Mathematical and Statistical Psychology, 64, 388-409.