Journal of Modern Applied Statistical Methods

Volume 10 | Issue 1 Article 33

5-1-2011

Type I Error Inflation of the Separate-Variances Welch t test with Very Small Sample Sizes when Assumptions Are Met

Albert K. Adusah
Ohio University

Gordon P. Brooks
Ohio University, brooksg@ohio.edu

Journal of Modern Applied Statistical Methods
May 2011, Vol. 10, No. 1, 362-372

1538 – 9472/11/$95.00

Emerging Scholars

Type I Error Inflation of the Separate-Variances Welch *t* test with Very Small Sample Sizes when Assumptions Are Met

Albert K. Adusah Gordon P. Brooks
Ohio University,
Athens, Ohio USA

This Monte Carlo study shows that the separate-variances Welch *t* test has inflated Type I error rates at very small sample sizes, especially when sample sizes are very small in one group and larger in the second group – even when all assumptions for the statistical test are met.

Key words: Small sample sizes, type I error inflation, Welch *t* test, Student *t* test.

Introduction

It is well known that violations of the homogeneity of variance assumption can severely diminish the confidence we have in the statistically significant results of our statistical tests—in particular, the pooled-variance independent *t* test. For example, the independent *t* test is relatively robust to violations of the homogeneity of variance assumption when sample sizes are equal, or perhaps even just relatively equal. Stevens (1999) indicated that "unequal variances will distort the Type I error rate appreciably *only if* the group sizes are sharply unequal (largest/smallest > 1.5)" (p. 9). But when the sample sizes are not relatively close, Type I error rates can be affected dramatically (Author, et al., 2004). As Mickelson and Ayers (2001) stated, "this implies a real risk of claiming to have generated a new understanding, ostensibly corroborated by a statistical significance test, when in actuality the 'finding' is nothing more than an artifact of violating an assumption of the test" (p. 3).

Just as it is well known that the actual Type I error probability rate of the pooled-variance *t* test, or Student *t* test, is raised or depressed by unequal variances combined with unequal sample sizes, it is also fairly well known that the separate-variances version of the *t* test, often called the Welch *t* test, usually eliminates these effects (Hinkle, Wiersma & Jurs, 2003). That is, the Welch *t* test maintains the nominal Type I error rate (i.e., level of significance or α) no matter how unequal the variances. Because power differences between the tests are relatively small when assumptions are met, and because the Welch *t* test maintains the nominal α even under violations of the homogeneity of variances assumption, some researchers have recommended abandoning both the Student *t* test and the commonly used preliminary tests of variances (e.g., Levene's test of equality of variances) in favor of Welch *t* tests with no preliminary variance tests. For example, Zimmerman (2004a) suggested that "when sample sizes are unequal, it appears that the most efficient strategy is to perform the Welch *t* test or a related separate-variances test unconditionally, without regard to the variability of sample values" (p. 180).

Interestingly results reported – but not interpreted – by Zimmerman (2004a), Gibbons

Albert K. Adusah was a doctoral student in Educational Research and Evaluation at Ohio University. Gordon P. Brooks is an Associate Professor of Educational Research and Evaluation. His research interests include statistics education, power and sample size analysis, and Monte Carlo programming. Email: brooksg@ohio.edu.

Editorial Note: JMASM regrets the passing of Mr. Adusah. This article was accepted for issue 7(2), but was inadvertently omitted.

and Chakraborti (1991), and Penfield (1994) suggested a problem with Type I error rates for the Welch *t* test even when variances are equal. The Welch *t* test appears to exhibit inflated Type I error rates when sample sizes are very small and the homogeneity assumption is met (i.e., both groups have the same variance). For example, Zimmerman found that with $n_1 = 5$ and $n_2 = 25$, actual Welch *t* test Type I error rates were approximately 0.058 in the equal variance condition; as the standard deviation ratio increased from 1.0 to 2.5, however, Type I error rates decreased toward 0.05. Gibbons and Chakraborti calculated a similar result with equal variances when $n_1 = 4$ and $n_2 = 16$: an actual Type I error rate of 0.0582. Curiously, Penfield reported a too-conservative actual Type I error rate for the Welch *t* test with $n_1 = 5$ and $n_2 = 15$.

Unfortunately, because none of these studies sought to examine this problem specifically, they did not include sufficient sets of conditions to confirm whether such results represented systematic bias or were simply artifacts of Monte Carlo sampling error (e.g., result of a particular random number generator seed or a particular random number generation process). For example, Zimmerman (2004a) only used conditions where the sample size combinations were (50, 10), (40, 20), (25, 5), and (20, 10). Penfield (1994) used combinations of (5, 5), (10, 10), (20, 20), (5, 15), and (10, 20). Gibbons and Chakraborti (1991) only used sample size combinations of (10, 10) and (4, 16). However, when taken together, these studies suggest that it may be fruitful to examine the matter further. Therefore, the purpose of this study is to investigate the Type I error rate behavior of the Welch *t* test under very small sample size conditions.

It is commonly understood that the Type I error rates of the Student *t* test and the Welch *t* test differ in respect to how these tests fare when both sample sizes and population variances are unequal across groups. These conditions alter both Type I error rates and power (Author, et al., 2004); that is, when the larger group has the smaller variance, the actual Type I error rate of the Student *t* test is inflated – or higher than the nominal Type I error rate. In other words, researchers would make more Type I errors than they expect to make using their given level of significance. Recall that when Type I error is set to 0.05, a researcher expects to make Type I errors at a rate of 5%; when assumptions are violated and the actual Type I error rate becomes inflated, however, the expected number of actual Type I errors is higher than 5% over a hypothetically large number of samples.

For example, if a researcher conducts (hypothetically) 100 statistical tests where the null hypothesis is true and statistical assumptions are met, 5 of those 100 tests would be wrongly rejected using an actual Type I error rate roughly equal to nominal $\alpha = 0.05$; but if the homogeneity of variance assumption is not met and the actual Type I error rate becomes inflated to 0.14, then roughly 14 of the 100 null hypotheses would be wrongly rejected, not 5 as expected when $\alpha = 0.05$. Conversely, when the larger group has the larger variance, the Type I error rate of the Student *t* test is conservative (i.e., lower than nominal α) and the null hypothesis is rejected less often than it should be (e.g., 2% of the time), which in turn reduces statistical power.

Much research has confirmed that these problematic properties of the Student *t* test can be eliminated by using the Welch *t* test (e.g., Gibbons & Chakraborti, 1991; Glass, Peckham & Sanders, 1974; Zimmerman, 2004a). Numerous studies have found that the Welch *t* method maintains Type I probabilities close to the nominal significance level and also eliminates spurious increases or decreases of Type II error rates and power (Zimmerman, 2004b). Although several studies have investigated unequal samples and unequal variances, no studies could be found that examined the impact of small sample sizes on such results. That is, Monte Carlo studies have included sample size as a variable (e.g., Gibbons and Chakraborti, 1991; Zimmerman, 2004a), but none could be found that systematically studied the effects of sample size itself on Type I error.

Gibbons and Chakraborti (1991) compared the Mann-Whitney *U* test, the Student *t* test, and the Welch *t* test. They used a total sample size of 20 for the two groups, sometimes equal (i.e., $n_1 = n_2 = 10$) and sometimes with $n_1 = 4$ and $n_2 = 16$. Because their focus was on violations of assumptions, they paid little

attention to the inflated Type I error rates of the Welch *t* test for the equal variance but unequal groups condition, where "the largest difference of the average of the three runs was 0.0596 - 0.0500 = 0.0096 for the two-tailed [Welch *t*] test" (p. 261). This is the summary of their results wherein actual Type I error for the equal variance but unequal sample size conditions were consistently beyond Bradley's (1978) fairly stringent criterion of $\alpha \pm 0.1\alpha$ (i.e., 0.045 to 0.055). In the end, Gibbons and Chakraborti recommended that "if the populations can be assumed normal with equal variances, use Student's *t* test for any sample sizes" (p. 266), but "if the populations can be assumed normal but the variances cannot be assumed equal, use the alternate t test for any sample sizes" (p. 266). Gibbons and Chakraborti recommended the Mann-Whitney test for non-normal data and when either (or both) sample size is less than 30.

Also for example, Zimmerman (2004a) compared the unconditional Student *t* test (i.e., no preliminary test of equality of variances), the unconditional Welch *t* test, and the Conditional *t* test (i.e., Levene's test followed by the appropriate *t* test). Zimmerman reported – but did not comment on – the condition where $n_1 = 25$, $n_2 = 5$, and $\sigma_1 / \sigma_2 = 1.0$, in which actual Type I error was 0.058 for the Welch *t* test but a more accurate 0.051 for the Student *t* test. Because Gibbons and Chakraborti (1991) used on 5,000 replications per condition, their results may have been subject to Monte Carlo sampling error issues (e.g., a poor seed choice, a particularly odd set of 5,000 randomly drawn samples). However, Zimmerman's (2004a) results were based on 50,000 replications, thus producing results less likely to be due to Monte Carlo sampling error issues. Further, among the equal variance conditions in both studies, only these results with very small *n* in one group were outside the fairly stringent range (i.e., 0.045 to 0.055).

Small Sample Sizes in Research

Although very small sample sizes are rare when *t* tests are used in actual research, several meta-analyses have been reported to suggest that researchers sometimes, in practice, do use very small sample sizes. For example, Reid, Kenaley and Colvin (2004) completed a meta-analysis of 39 small-group interventions in social work. They found that 15 of these 39 studies (i.e., 38%) had a total sample size of 20 or less; only 10 had total sample sizes over 50. Similarly, Shadish and Baldwin (2005) performed a meta-analysis of marital therapy interventions and found 14 of 30 studies had total sample sizes of 20 or less, while only 2 had total sample sizes over 50. Unfortunately, these studies did not report individual sample sizes, so whether group sizes were equal is unknown without further investigation.

Methodology

A Monte Carlo data generation and analysis program, called MC4G: Monte Carlo Analyses for up to 4 Groups (Author, 2005), was used to simulate data to obtain the appropriate Type I error rates. The rejection rates of both the Student *t* test and the Welch *t* test will be recorded for various combinations of sample sizes, especially with very small sample size in one group. That is, the specific conditions for the study were: (a) both Group 1 and Group 2 means remained constant at 0.0, (b) Group 1 sample size varied from 3 to 150 by 1, (c) Group 2 sample size varied from 3 to 30 by 1, (d) Group 1 standard deviation remained constant at 1.0, and (e) Group 2 standard deviation varied from 1.0 to 4.0 by 0.5.

For the primary research question, only the 3,738 conditions were analyzed where Group 1 sample sizes were larger than Group 2 sample sizes and both standard deviations were 1.0; however, some other conditions were analyzed for specific reasons. All data were generated to follow a univariate normal distribution. There were 100,000 replications performed for each condition in order to minimize the impact of Monte Carlo sampling problems. For each sample generated, appropriate standard error estimates and degrees of freedom were used to calculate both the Student *t* test (Hinkle et al., 2003, p. 240), the Welch *t* test (Hinkle et al., 2003, p. 252), and a Conditional *t* test (either the Student *t* test or the Welch *t* test was calculated appropriately depending on the results of Levene's test of equality of variances). Nominal level of significance was set at $\alpha = 0.05$ for each test performed.

The MC4G program was developed (Brooks, 2005) to perform Monte Carlo analyses for *t* tests and ANOVA in a Windows environment. The MC4G program was written in Delphi Pascal and is available for download from the author's web site (see references). The program was used to create normally distributed data that met the conditions for the study. For these robustness analyses, the number of incorrect rejections of the null hypothesis (i.e., Type I error rate) was stored and reported by the program.

The MC4G program uses the L'Ecuyer (1988) uniform pseudorandom number generator. Specifically, the FORTRAN code of Press, Teukolsky, Vetterling, and Flannery (1992), was translated into Delphi Pascal. The L'Ecuyer generator was chosen because of its large period and because combined generators are recommended for use with the Box-Muller method for generating random normal deviates, as will be the case in this study (Park & Miller, 1988). The computer algorithm for the Box-Muller method used in this study was adapted for Delphi Pascal from the standard Pascal code provided by Press, Flannery, Teukolsky and Vetterling (1989). Extended precision floating point variables were used, providing the maximum possible range of significant digits. Simulated samples were chosen randomly to test program function by comparison with results provided by SPSS.

Results

First, the Type I error rates of the Student *t* test are investigated across the full range of sample size conditions. These results confirmed that Type I error rates for the Student *t* test are robust to variation of all sample sizes tested. Specifically, every one of the 3,738 sample size conditions under equal variances (i.e., both group standard deviations are 1.0) was between 0.0446 and 0.0560, just beyond the most stringent criterion recommended by Bradley (1978). One would not expect Type I error rates of exactly 5% due to the sampling error inherent to the Monte Carlo process. Therefore, Bradley recommended a stringent criterion of $\alpha \pm 0.1\alpha$ to be used for robustness studies; that is, results within 10% of α are considered close enough to α for the statistical test to be considered robust to the conditions being investigated. These results are shown graphically in Figure 1.

A similar examination of the Welch *t* test was performed and an issue with robustness for these results was identified (see Figure 2). In particular, the actual Type I error rates across the 3,738 conditions (100,000 samples per condition) ranged from 0.0424 to 0.0793. Clearly, some of the Type I error rates for the Welch *t* test fell outside Bradley's (1978) stringent criterion range. Further comparison showed that 99% of all Student *t* test Type I error rates were less than 0.0536, but only 88% of the Welch *t* test Type I error rates were below 0.0551, at the top end of Bradley's range. Also, there were only 10 extreme Student *t* test Type I error rates beyond 0.0542 but there were 340 extreme Welch *t* test Type I error rates beyond 0.0569.

In order to investigate further the inflated Type I error rates for the Welch *t* test, an attempt was made to identify the patterns in Figure 2. Observe clear patterns among the scatter that represent Group 2 sample sizes. For example, at the top of the chart, there is a clear pattern of circles, representing a Group 2 sample size of $n_2 = 3$. Because a sample size of $n_2 = 3$ is not practical, we examined further the $n_2 = 5$ condition (while still not terribly practical, it is more reasonable than $n_2 = 3$ and has been studied by several authors cited above). Table 1 shows these results for a subset of the data (only where $n_1 < 45$, but no important differences existed beyond $n_1 = 45$). Figure 3 displays these data for equal variances, Figure 4 illustrates the data where variances were unequal (Group 1 SD = 1.0 and Group 2 SD = 2.0), Figure 5 shows the data where variances were unequal (Group 1 SD = 1.0 and Group 2 SD = 4.0).

The Welch *t* test clearly has inflated Type I error rates when sample sizes are small and unequal; however, note in Figure 3 that the inflation does not emerge until the sample size ratio increases beyond 2:1 (specifically, where $n_1 = 13$ and $n_2 = 5$). Although the inflation is not dangerously high, as is the case with the Student *t* test when both sample size and variances are unequal (e.g., where $n_1 = 44$, $n_2 = 5$, $\sigma_1 = 1.0$, and $\sigma_2 = 4.0$, as shown in Figure 5), it does exist. Interestingly, Figures 4 and 5 show that the Welch *t* test does indeed maintain nominal Type

I error rates when variances are unequal, but Figure 3 shows that when variances are equal the Type I error rates are biased upward. Further investigation beyond the conditions where $n_2 = 5$ suggested that the problem is limited to very small sample sizes. Figure 6 shows that, although there is a clear, upward bias of Type I error beyond a smaller group size of $n = 10$, those rates do fall well within Bradley's (1978) stringent criterion range. Figure 6 also shows that the average inflation of Type I error reduces dramatically as the smaller group size increases. Further note in Figure 6 that the *t* test conditional on the result of Levene's test does not help the matter, because its Type I error rates are inflated even beyond the Welch *t* test once $n_2 > 4$.

Conclusion

Results suggest that the Welch *t* test is indeed inflated, according to Bradley's (1978) fairly stringent criterion, when sample sizes are unequal – even when assumptions for the *t* test are met in the population. The inflation rate seems to be dependent more on the size of the smaller group than on the total sample size, but sample size ratio does seem to play a small role(i.e., with roughly equal sample sizes there was no apparent inflation). Although the Welch *t* test Type I error inflation exposed here is not dangerously high, it is high enough to be considered more than trivial, particularly with the smallest smaller group sample sizes examined. Specifically, Type I error rates are inflated beyond Bradley's stringent criterion

Figure 1*:* Type I Error Rates for the Student *t* test when Homogeneity of Variance Assumption Is Met in the Population

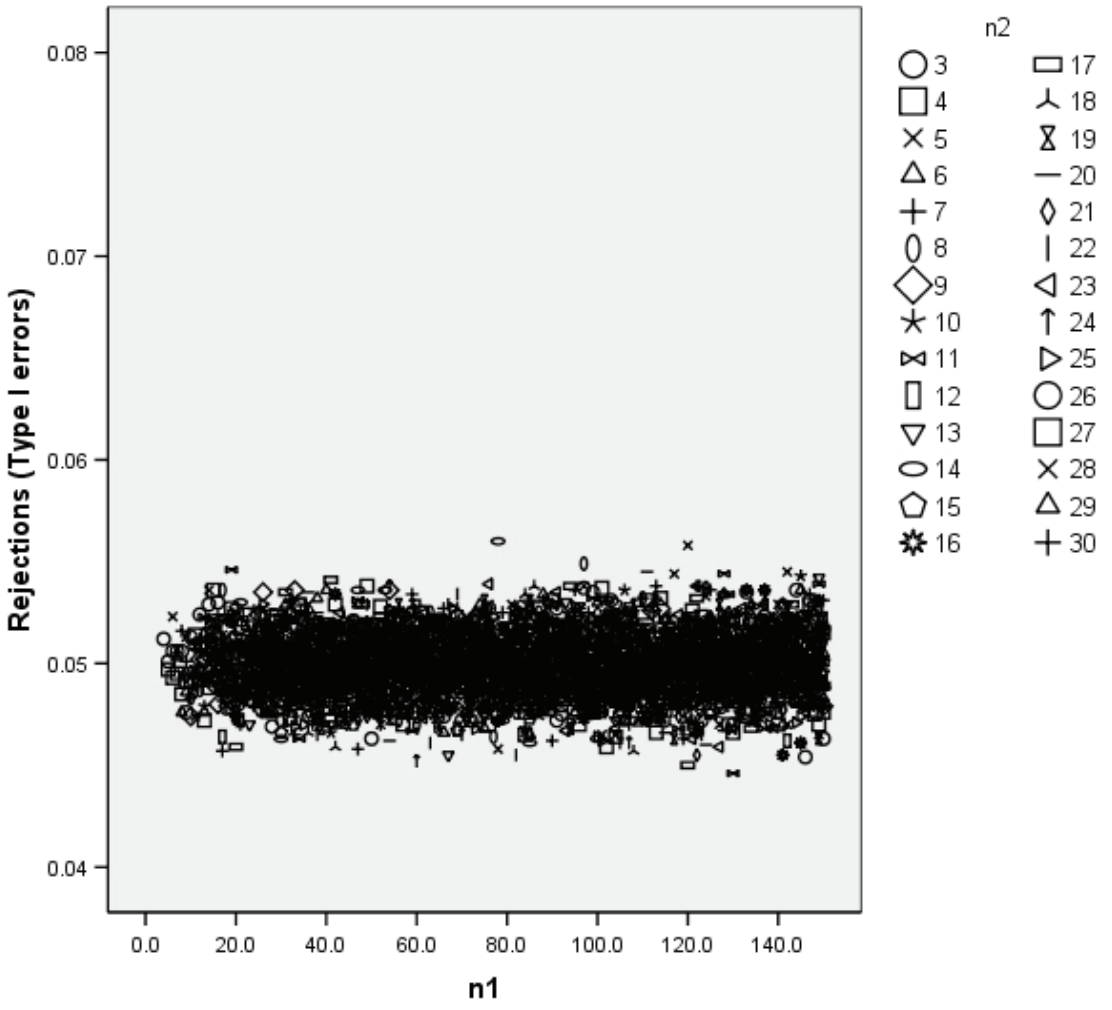


Figure 2: Type I Error Rates for the Welch *t* test when Homogeneity of Variance Assumption Is Met in the Population

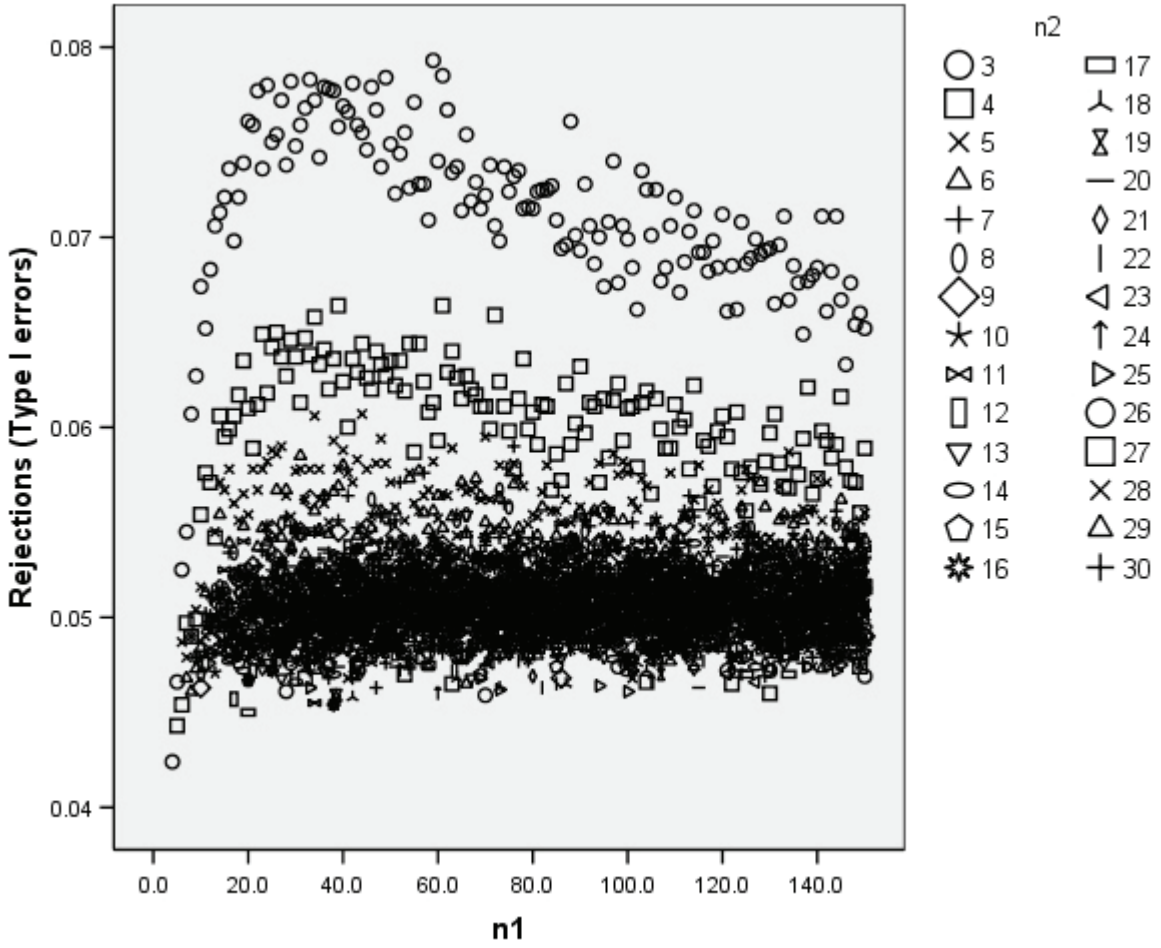


Table 1: Type I error rates of Student t test, Welch t test, and the Conditional t test at $\alpha = 0.05$ Where $n_2 = 5$, $n_1 < 45$, $n_1 > n_2$, and Both Population Standard Deviations are 1.0

n_1	Student *t*	Welch *t*	Conditional *t*
6.0	0.0523	0.0487	0.0514
7.0	0.0507	0.0479	0.0499
8.0	0.0495	0.0490	0.0499
9.0	0.0513	0.0504	0.0519
10.0	0.0493	0.0516	0.0505

Table 1 (continued): Type I error rates of Student t test, Welch t test, and the Conditional t test at $\alpha = 0.05$ Where $n_2 = 5$, $n_1 < 45$, $n_1 > n_2$, and Both Population Standard Deviations are 1.0

n_1	Student *t*	Welch *t*	Conditional *t*
11.0	0.0474	0.0490	0.0488
12.0	0.0492	0.0512	0.0518
13.0	0.0505	0.0545	0.0523
14.0	0.0536	0.0571	0.0557
15.0	0.0498	0.0567	0.0525
16.0	0.0519	0.0578	0.0556
17.0	0.0514	0.0560	0.0542
18.0	0.0508	0.0551	0.0551
19.0	0.0505	0.0565	0.0541
20.0	0.0525	0.0554	0.0554
21.0	0.0499	0.0578	0.0549
22.0	0.0493	0.0567	0.0532
23.0	0.0507	0.0578	0.0558
24.0	0.0501	0.0582	0.0553
25.0	0.0493	0.0588	0.0551
26.0	0.0527	0.0586	0.0586
27.0	0.0494	0.0590	0.0556
28.0	0.0507	0.0564	0.0554
29.0	0.0501	0.0553	0.0558
30.0	0.0512	0.0577	0.0569
31.0	0.0480	0.0578	0.0542
32.0	0.0503	0.0562	0.0561
33.0	0.0532	0.0578	0.0589
34.0	0.0490	0.0606	0.0573
35.0	0.0481	0.0578	0.0547
36.0	0.0485	0.0560	0.0549
37.0	0.0503	0.0550	0.0567
38.0	0.0490	0.0593	0.0567
39.0	0.0498	0.0578	0.0567
40.0	0.0490	0.0588	0.0560
41.0	0.0518	0.0576	0.0565
42.0	0.0485	0.0571	0.0558
43.0	0.0507	0.0583	0.0573
44.0	0.0501	0.0607	0.0578

Figure 3: Graphical Display of Results Where $n_2 = 5$ Across All $n_1 > 5$ and Both Standard Deviations were 1.0

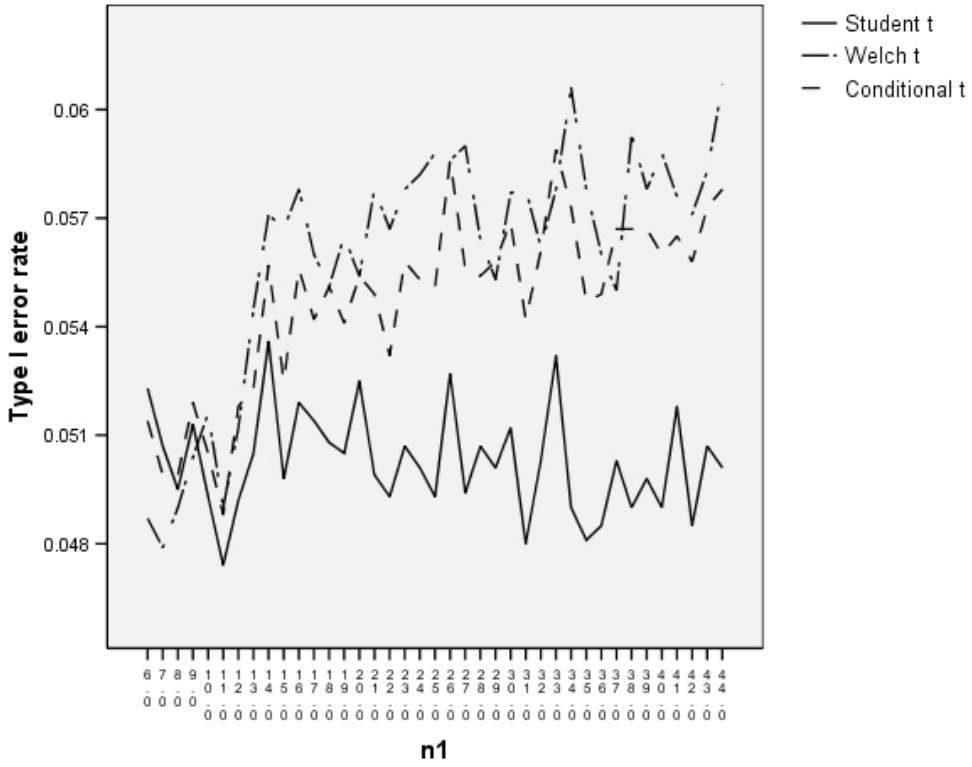


Figure 4: Graphical Display of Results where $n_2 = 5$ Across All $n_1 > 5$, Group 1 Standard Deviation was 1.0, and Group 2 Standard Deviation was 2.0

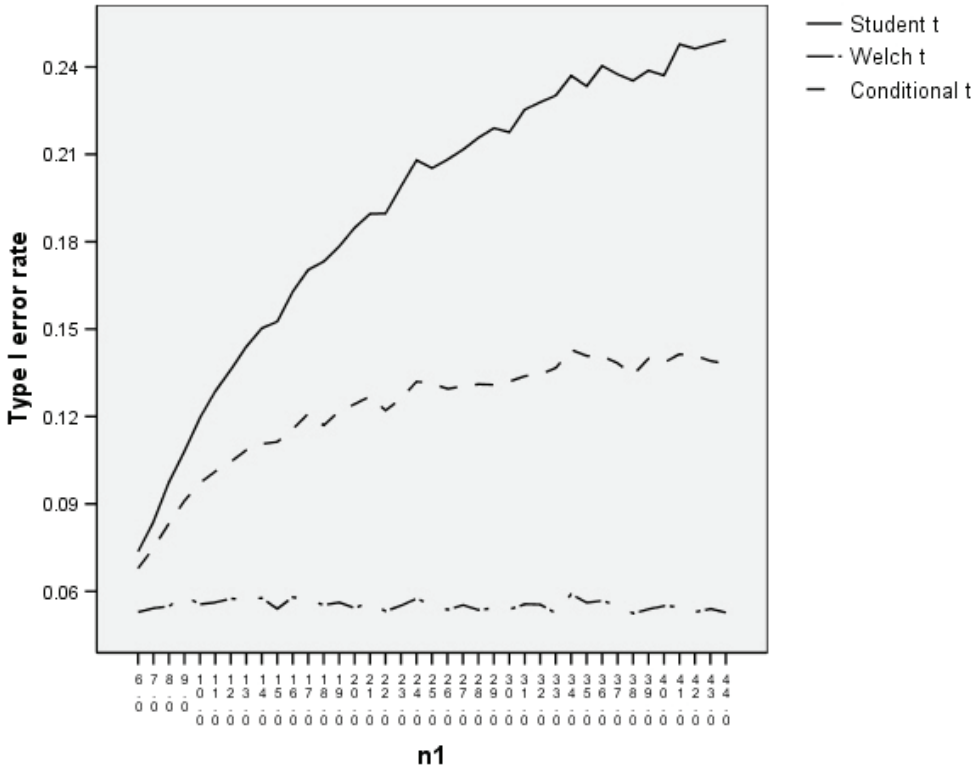


Figure 5: Graphical Display of Results where $n_2 = 5$ Across All $n_1 > 5$, Group 1 Standard Deviation was 1.0, and Group 2 Standard Deviation was 4.0

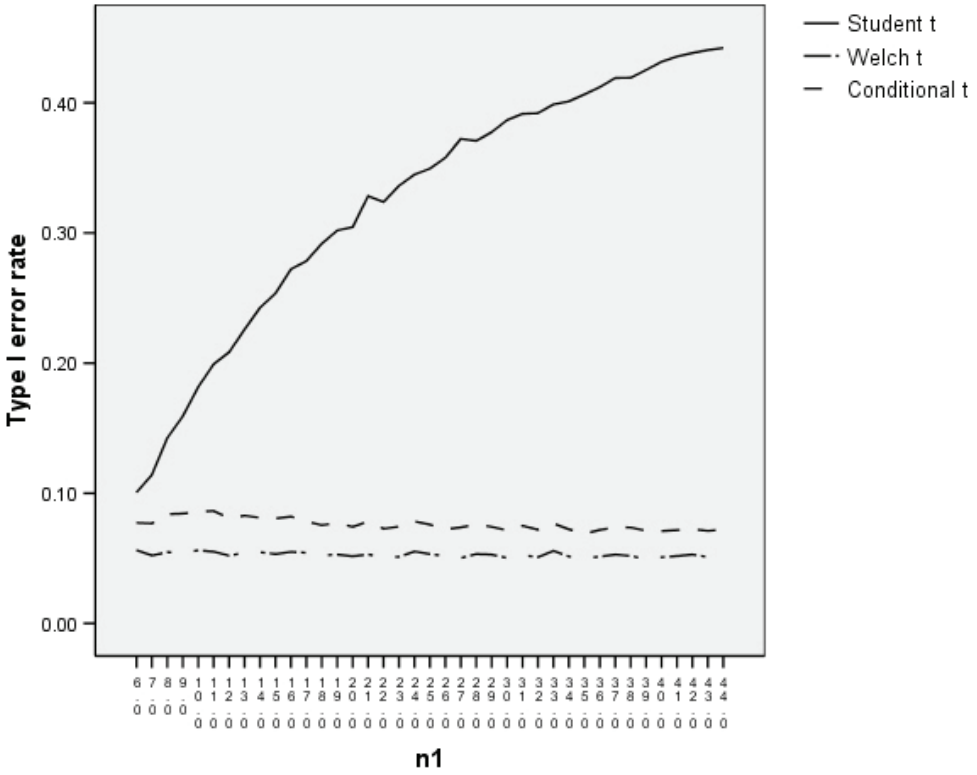
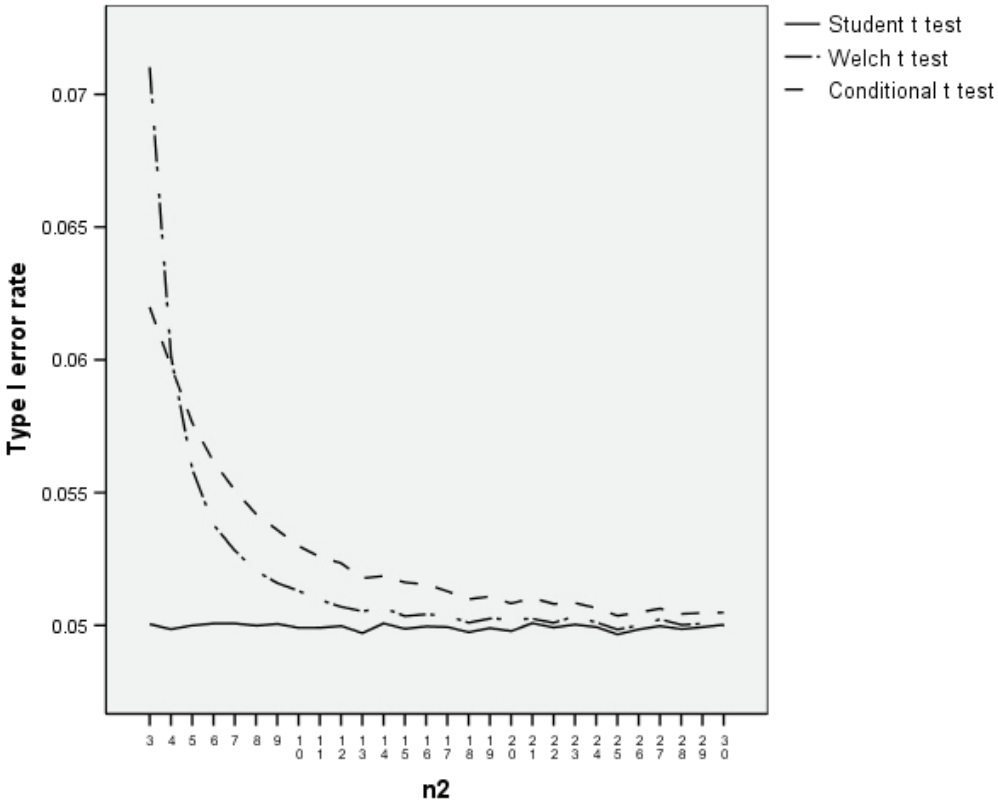


Figure 6: Average Type I Error Rates where Both Standard Deviations are 1.0



when the smaller sample size is less than $n = 6$. We also found that the inflation problem becomes relatively benign once the smaller sample size is greater than $n = 10$; that is, the average actual Type I error rates for the Student *t* test and the Welch *t* test differ by no more than 0.002 when smaller $n > 10$. Finally, we confirmed that the Student *t* test did not exhibit any noticeable problems with Type I error when assumptions are met, no matter the sample size combinations.

There have been a number of studies to ascertain the best statistical test to use for two-group comparison studies under violations of assumptions. Such studies have often also showed that there is not a dramatic difference in statistical power between the Student *t* test and Welch *t* test under many conditions. Consequently, these results have led some scholars (e.g., Zimmerman, 2004a) to recommend using the Welch *t* test unconditionally, so as to minimize the impact of violations of assumptions on Type I error rates. Unfortunately, because it appears that the Welch *t* test may have unexpected problems when one group is very small, this recommendation may lead to problems in studies with very small sample sizes. Indeed, supplemental analyses performed here suggested that the Welch *t* test may be conservative for very small, equal sample sizes (less than 7 in each group) even when variances are equal.

Because the Conditional *t* test did not help the situation, there is no easy solution to the problem. That is, because one does not know whether the homogeneity of variance assumption has been violated, one cannot know which *t* test to choose with small sample sizes. More specifically, if one knew that the populations had unequal variances, one could choose to use the Welch *t* test with little concern for type I error, even with small sample sizes; conversely, if one knew that variances were equal, one could use the Student *t* test. However, the commonly recommended Conditional *t* test using Levene's test also appears to lead to inflated type I error rates with very small sample sizes in one group and with larger sample sizes in the other—even when variances are equal.

The most obvious recommendation, for a variety of reasons both statistical and otherwise, is for researchers to use more than 10 participants per group when comparing means. In situations where there is no choice, based on Gibbons and Chakraborti's (1991) results, it appears that researchers should use the Mann-Whitney *U* test when sample sizes are very small to maintain nominal Type I error rates; their results do not hint at any inflation of Type I error rates at small sample sizes. However, future research must verify this recommendation. Further investigation into type I error rates should include examinations of Analysis of Variance and its alternatives (e.g., Brown-Forsythe, Welch, and Kruskal-Wallis). There is no reason to expect terribly different results when viewed from an ANOVA perspective; such similarities between the Type I error rate properties of the *t* test and ANOVA have been confirmed in the literature (e.g., Glass, Peckham, & Sanders, 1974). Finally, these results relied on the assumption of normality being met; future researchers may want to investigate the problem by violating the normality assumption. Based on work by Gibbons and Chakraborti, and others, there is reason to suspect that the nonparametric tests should be uniformly adopted as the tests of choice when the sample size of at least one group is very small.

References

Brooks, G. P. (2005). MC4G: Monte Carlo analyses for up to 4 groups [computer software]. http://www.ohio.edu/people/brooksg/software.htm.

Brooks, G. P., Fang, H., Heh, V. (2004, October). *Modeling true Type I error rates: A Monte Carlo analysis*. Paper presented at the meeting of the Mid-Western Educational Research Association, Columbus, OH.

Brooks, G. P., Johanson, G. A., & Barcikowski, R. S. (2004, April). *Underappreciated factors that affect statistical power*. Paper presented at the meeting of the American Educational Research Association, San Diego, CA.

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

Gibbons, J. D., & Chakraborti, S. (1991). Comparisons of the Mann-Whitney, Student's *t*, and alternate *t* tests for means of normal distributions. *Journal of Experimental Education*, *59*, 258-267.

Glass, G. V., Peckham, P. D., Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research*, *42*, 237-288.

Hinkle, D. E., Wiersma, W., & Jurs, S. G. (2003). *Applied statistics for the behavioral sciences* (*5th Ed.*). Boston: Houghton Mifflin.

Mickelson, W. T., & Ayers, R. (2001, October). *Testing for group mean difference: What happens when population variances are unequal*. Paper presented at the annual meeting of the Mid-Western Educational Research Association, Chicago, IL.

Reid, W. J., Kenaley, B. D., & Colvin, J. (2004). Do some interventions work better than others? A review of comparative social work experiments. *Social Work Research*, *28*(*2*), 71-81.

Shadish, W. R., & Baldwin, S. A. (2005). Effects of behavioral marital therapy: A meta-analysis of randomized controlled trials. *Journal of Consulting and Clinical Psychology*, *73*, 6-14.

Stevens, J. P. (1999). *Intermediate statistics: A modern approach* (*2nd Ed.*). Mahwah, NJ: Lawrence Erlbaum Associates.

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

Zimmerman, D. W. (2004b). Conditional probabilities of rejecting null hypotheses by pooled and separate variances *t* tests given heterogeneity of sample variances. *Communications in Statistics - Simulation and Computation*, *33*, 69-81.