Practical Assessment, Research, and Evaluation

Volume 19 *Volume 19, 2014* Article 11

2014

# Impact of Sample Size and Variability on the Power and Type I Error Rates of Equivalence Tests: A Simulation Study

Shayna A. Rusticus

Chris Y. Lovato

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.



Volume 19, Number 11, August 2014 ISSN 1531-7714

# Impact of Sample Size and Variability on the Power and Type I Error Rates of Equivalence Tests: A Simulation Study

Shayna A. Rusticus & Chris Y. Lovato

*University of British Columbia*

The question of equivalence between two or more groups is frequently of interest to many applied researchers. Equivalence testing is a statistical method designed to provide evidence that groups are comparable by demonstrating that the mean differences found between groups are small enough that they are considered practically unimportant. Few recommendations exist regarding the appropriate use of these tests under varying data conditions. A simulation study was conducted to examine the power and Type I error rates of the confidence interval approach to equivalence testing under conditions of equal and non-equal sample sizes and variability when comparing two and three groups. It was found that equivalence testing performs best when sample sizes are equal. The overall power of the test is strongly influenced by the size of the sample, the amount of variability in the sample, and the size of the difference in the population. Guidelines are provided regarding the use of equivalence tests when analyzing non-optimal data.

Do students who complete their medical education in a distributed program achieve the same level of academic competence regardless of the location where they complete their education? This is an example of a question concerning the equivalence of two (or more) groups. It is not the same as a question that asks whether two groups of students are achieving different levels of academic competence (e.g., Do students who complete their medical education in an urban location achieve higher marks than students completing their medical education in a rural location?). However, questions of these two types (equivalence and difference) are most commonly analyzed using the same method: a test of the null hypothesis of no significant difference.

ANOVAs and t-tests, often referred to as "difference tests", are designed to provide evidence that groups are different when a statistically significant p-value is calculated. A significant p-value indicates that there is enough evidence to reject the null hypothesis of no difference, thus supporting the alternative hypothesis that there is a difference between the groups. To address questions of equivalence, researchers have commonly used these same tests to conclude that groups are equivalent when a non-significant p-value is found. As several researchers have argued (e.g., Cribbie, Gruman, & Arpin-Cribbie, 2004; Rusticus & Lovato, 2011), this is not the correct method to use if your purpose is to demonstrate that groups are comparable. A non-significant finding, which reflects a failure to reject the null hypothesis of no difference, rather than the acceptance of the null hypothesis, indicates only that there is not enough evidence to support that two groups are statistically different. It does not provide sufficient evidence for the groups being comparable; a non-significant result could indicate that the groups are comparable, but it could also be a reflection of insufficient sample size or unreliable measurements.

To correctly address questions about comparability, equivalence testing is a more appropriate method. Equivalence testing provides evidence of equivalence by demonstrating that any difference that exists between groups is small enough that, for practical purposes, the groups can be treated as equivalent (Blackwelder, 2004; Rogers, Howard, Vessey, 1993). Although still a form of statistical significance testing, the role of the null and alternative hypotheses have been reversed, such that the null hypothesis in an equivalence test asserts that the

difference between two groups is at least as large as a difference specified in advance by the researcher (i.e., the point at which the difference represents a meaningful difference). The alternative hypothesis is that the difference is smaller than the one specified by the researcher. A rejection of the null hypothesis here provides support for the alternative hypothesis that any difference that exists is not of practical importance.

As the usual meanings of the null and alternative hypotheses have been reversed, this means that the interpretations of Type I and Type II errors and power must also be altered. In both difference and equivalence testing, a Type I error occurs when we incorrectly conclude that the null hypothesis is false; a Type II error occurs when we incorrectly conclude that the null hypothesis is true. Power is the ability to correctly reject the null hypothesis. For equivalence testing, the practical interpretation of this is that a Type I error occurs when we conclude equivalence when in fact the groups are not equivalent and a Type II error occurs when we conclude non-equivalence when in fact the groups are equivalent. The power of an equivalence test is its ability to correctly conclude that two groups are equivalent.

While tests of equivalence have been gaining in popularity within fields such as education and psychology, as researchers are becoming more aware of this method, few recommendations currently exist regarding the appropriate use of these tests. The primary concern is related to whether these tests will be able to correctly detect equivalence when the groups are equivalent (i.e., the test has sufficient power) and will not conclude equivalence when the groups truly are different (i.e., a Type I error). Typically, we want the power of our test to be at .80 or greater (i.e., we will correctly conclude equivalence 80% or more of the time). Insufficient power could lead to a conclusion of non-equivalence even if the population means were equivalent (a Type II error). It is also equally important to ensure that the Type I error rates are at an appropriate level. A Type I error rate of .05 is generally considered to be acceptable (i.e., we will incorrectly conclude equivalence around 5% of the time). An inflated Type I error rate is of more concern than a depressed rate in that the former makes a test invalid, while the latter makes the test more conservative (Nordstokke, Zumbo, Cairns, & Saklofske, 2011). Because tests of equivalence and difference tests differ in how they specify the null hypothesis, and have been shown to have different power (Rusticus & Lovato, 2011), it cannot be assumed that the power of these two types of tests will be the same.

The confidence interval approach, also known as Schuirmann's equivalence test (Schuirmann, 1987), was selected because it is a commonly used approach when conducting tests of equivalence and is easy to calculate in popular software programs such as SPSS. Furthermore, this approach has performed better, or nearly as good as, other methods of assessing equivalence (Cribbie et al., 2004; Gruman, Cribbie, Arpin-Cribbie, 2007). Briefly stated, this approach calculates a confidence interval around the difference between two group means using the standard error of the difference. If this confidence interval is within a pre-specified range (the equivalence interval) then the groups are said to be equivalent. Rogers and colleagues (1993) suggest that two groups are different when "the minimum difference between two groups that would be important enough to make the groups non-equivalent" (p. 554). As the difference between two groups could be in either a positive or negative direction, there is both a positive and a negative value to define equivalence, forming an equivalence interval. Equivalence can be concluded if the confidence interval around the mean difference is fully contained within the equivalence interval.

Sample size and variability are two important factors that conceptually should influence the power of an equivalence test. For instance, small sample sizes and high variability result in larger confidence intervals than large sample sizes and low variability. As such, small sample sizes and/or high variability should be more likely to lead to conclusions of non-equivalence while large sample sizes and/or low variability should be more likely to lead to conclusions of equivalence.

Simulation studies by Cribbie and colleagues (Cribbie, Arpin-Cribbie, & Gruman, 2010; Cribbie et al., 2004; Gruman et al., 2007) have compared the performance of the Schuirmann confidence interval approach to the student t difference test, as well as other modified equivalence tests, under combinations of equal and unequal sample sizes, group sizes, equal and unequal population variances, and/or population mean configurations. They found that the power of the confidence interval approach increases with increasing sample sizes and that it is affected by the pattern of sample heterogeneity, with power increasing when variances are positively paired (larger sample size paired with the larger variance) and decreasing when variances


DOI: https://doi.org/10.7275/4s9m-4e81

are negatively paired (larger sample size paired with the smaller variance). In general, it is clear from their work that the confidence interval approach has unacceptably low power when sample sizes are very small.

This study examined the power and type I error rates of the confidence interval approach to equivalence testing under varying conditions of equal and non-equal sample sizes and variances. We were particularly interested in the performance of equivalence tests when dealing with both unequal sample sizes and unequal variances, as this is a common situation we, and likely many others, have faced. For example, our work in higher education often requires conducting equivalence tests where one group can be up to seven times larger than the other group(s), and in many cases we have found that the data violate the assumption of homogeneity of variance. In the present study, we expand on previous studies by examining the power and Type I error rate of the Schuirmann confidence interval approach under conditions of both unequal sample sizes and unequal variances for comparisons involving both two and three groups. We conclude with a discussion of the implications of analyzing non-optimal data and provide recommendations for using tests of equivalence under such conditions.

## Methods

A simulation study was conducted to examine the power of the confidence interval approach (Schuirmann, 1987) to detect population equivalence. The data were simulated to represent academic assessment data; a continuous variable that theoretically ranges from 0 to 100, but more typically ranges between 60 and 100. To define parameters for the simulation, descriptive statistics were calculated on a set of commonly collected assessment variables (e.g., exam scores, end of year grades) for a sample of undergraduate medical school students. These analyses suggested a mean assessment score of 81 (SD = 3) and a mean standard deviation of 6 (SD = 2) were representative of the assessment data collected at our university; thus, all simulations were centered on these two values.

Several variables were manipulated in this study including (1) number of groups (2 or 3), (2) sample size (30, 60, 90, 150, 210), (3) sample standard deviation (4, 6, 8), and (4) population mean difference (0, 2, 4, 5 points). Based upon our previous work, we selected an equivalence interval of ±5% (Rusticus & Lovato, 2011). Thus, the first three population mean differences (0, 2, 4 points) represent data that should be deemed equivalent, allowing identification of the power of the analyses, whereas the latter (5 points) represents non-equivalence and allows us to identify the Type I error rate. In manipulating the standard deviation variable for the three group scenario, three variance conditions were created as follows:

1. Equal variance condition - all the samples sizes had a standard deviation of six.
2. Negatively paired variance condition - each sample size was paired with a specific standard deviation that decreased with increasing sample size: $n$ = 30, SD = 8; n = 60, SD = 7; $n$ = 90, SD = 6; $n$ = 150, SD = 5; $n$ = 210, SD = 4.
3. Positively paired variance condition - each sample size was paired with a specific standard deviation that increased with increasing sample size: $n$ = 30, SD = 4; n = 60, SD = 5; $n$ = 90, SD = 6; $n$ = 150, SD = 7; $n$ = 210, SD = 8.

One thousand normally distributed simulations were conducted for each condition using SPSS version 20. Ninety percent confidence intervals on the mean difference between groups were calculated for each set of simulations via the t-test procedure in the two group scenario and the analysis of variance procedure in the three group scenario. When variances were unequal in the two group scenario, the Welch-Satterhwaite corrected confidence intervals were used (i.e., the confidence intervals were obtained from the row reading "equal variances not assumed". In the three group case, confidence intervals were calculated using both a Games-Howell and Tukey post-hoc test to allow for comparison between these two options. Games-Howell was selected because this method takes unequal group sizes into account, as well as violations of homogeneity of variance (Dunnet, 1980). Tukey was selected as it is one of the most widely used post-hoc tests.

The power of the test was determined via the percentage of the 1000 simulations in each of the 0, 2 and 4 point difference conditions in which the confidence intervals were fully contained within the equivalence interval. The Type I error rate was determined by calculating the percentage of the 1000 simulations whereby equivalence was concluded in the 5 point difference condition when the correct conclusion should have been non-equivalence.

## Results

The results of a representative selection of the simulations for the two group scenario are presented in Figures 1 (power) and 2 (Type I error) and for the three group scenario are presented in Figures 3, 4 (power) and 5 (Type I error).[1] In each of the figures, the sample sizes for each condition are presented along the x-axis. The equal sample size pairings are on the left side of the graph and the unequal sample size pairings are on the right side of the graph; both are ordered from lowest to highest total sample size. Power is presented on the y-axis in Figures 1, 3, and 4 and represents the percentage of the 1000 simulations for which equivalence was *correctly* concluded. In general, power should increase as the sample size increases. Type I error is presented on the y-axis in Figures 2 and 5 and represents the percentage of the 1000 simulations for which equivalence was *incorrectly* concluded. Type I error should not be impacted by sample size. Separate

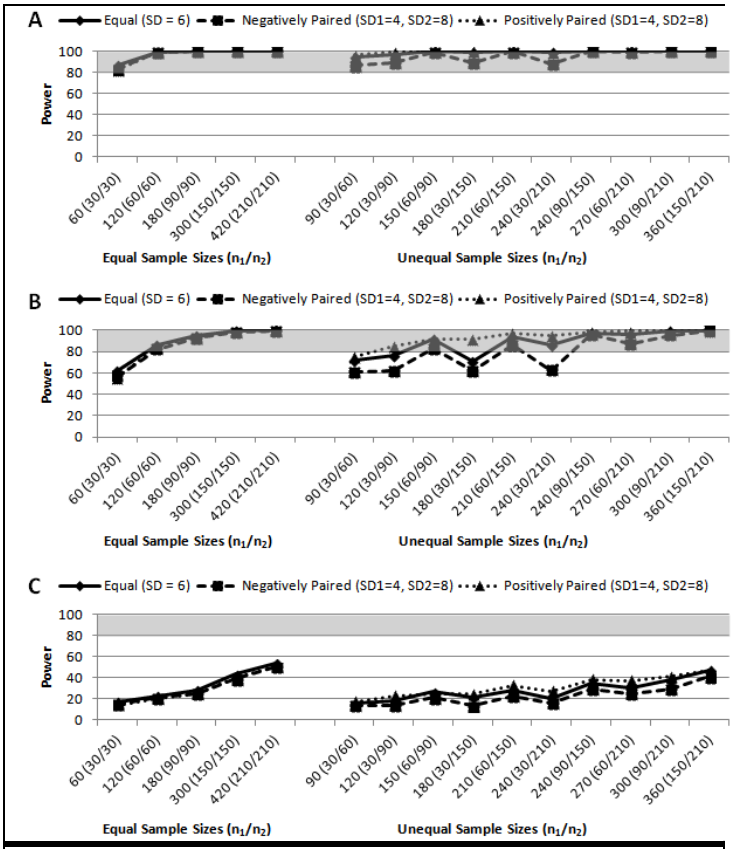


*Figure 1.* Power for declaring two populations equivalent under a representative selection of sample size condition and three selected variability conditions: equal, negatively, and positively paired variances. A: Population mean difference = 0. B: Population mean difference = 2. C: Population mean difference = 4.

*Note.* Data points contained within the shaded area are within acceptable limits for power.

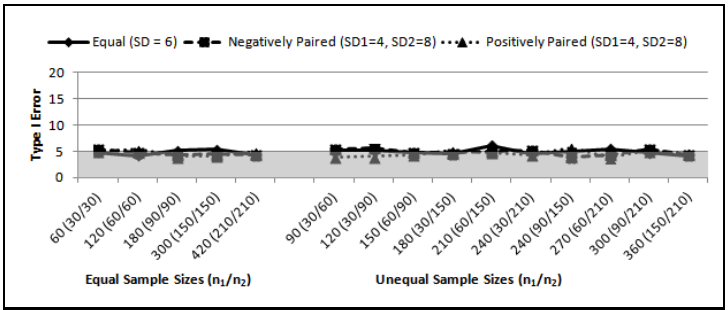


*Figure 2.* Type I error rate for declaring two populations equivalent under a representative selection of sample size condition and three selected variability conditions: equal, negatively, and positively paired variances.

*Note.* Ideally, data points should be maintained at 5%. Data points that fall below 5% indicate a test that is conservative, but still acceptable. Data points that are greater than 5% indicate a test that is too liberal and is not acceptable.

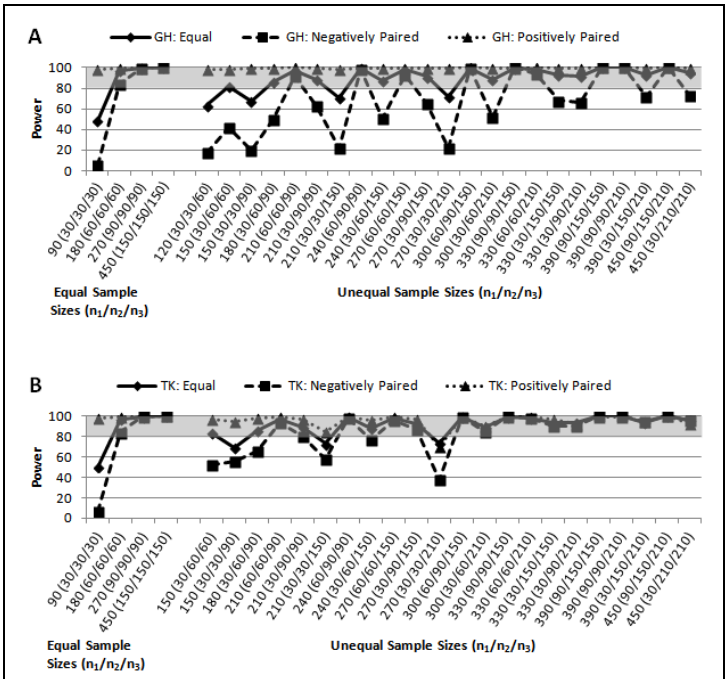


*Figure 3.* Power for declaring three populations equivalent using a (A) Games-Howell (GH) and (B) Tukey (TK) post-hoc test under selected sample sizes and equal, negatively paired, and positively paired variability conditions when the population mean difference is equal to zero.

*Note.* Data points contained within the shaded area are within acceptable limits for power.

[1] Readers interested in the graphical results for all simulation conditions should contact the first author.

lines are used for the three variance conditions: (1). equal variances, (2) negatively paired variances and (3) positively paired variances. The shaded rectangle in each figure represents the area of acceptable power (Figures 1, 3, and 4) or acceptable Type I error (Figure 2 and 5). Ideally, we would like to see the data points contained within the shaded areas

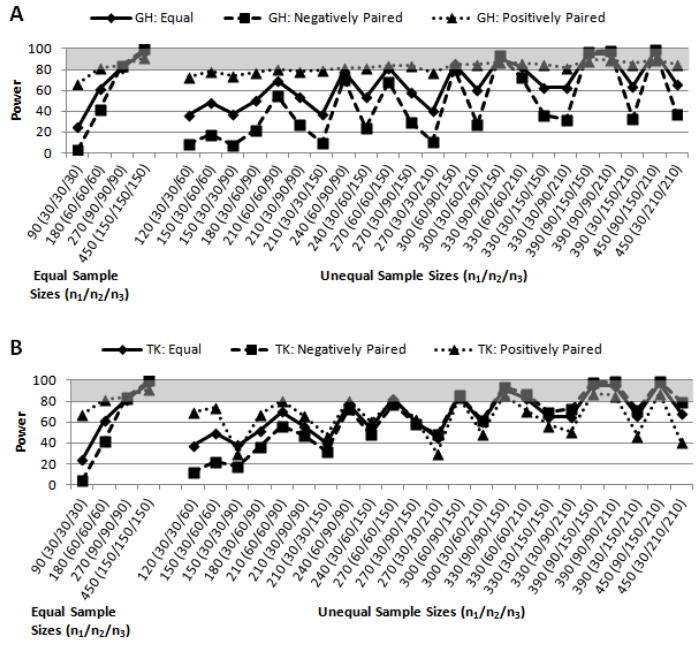


*Figure 4.* Power for declaring three populations equivalent using a (A) Games-Howell (GH) and (B) Tukey (TK) post-hoc test under selected sample sizes and equal, negatively paired, and positively paired variability conditions when the population mean difference is equal to two.

*Note.* Data points contained within the shaded area are within acceptable limits for power.

## Sample Size

In both the two and three group scenarios, when group sample sizes were equal, power increased as the total sample size increased. Equal sample sizes were also more powerful than unequal sample sizes of the same total sample size and the greater the disproportion in sample sizes, the lower the overall power of the test. For example, in looking at Figure 4A and the negative Games-Howell condition, there were four scenarios where the total sample size was 270. When each of the three sample sizes was 90, power was acceptable at 85%, however, as the samples became more unbalanced, power dropped to 69% (sample sizes = 60, 60, 150), then to 30% (sample sizes = 30, 90, 150) and finally to 11% (sample sizes = 30, 30, 210). Power was also lower in the three group scenario over the two group scenario for samples of equal size and variability.

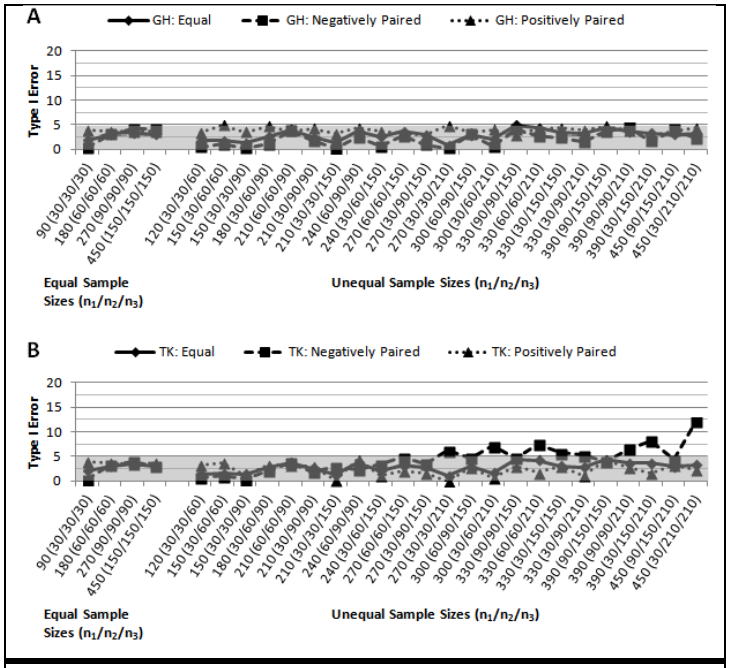


*Figure 5.* Type I error rate for declaring three populations equivalent under varying conditions of sample size and variability.

*Note.* Ideally, data points should be maintained at 5%. Data points that fall below 5% indicate a test that is conservative, but still acceptable. Data points that are greater than 5% indicate a test that is too liberal and is not acceptable.

## Variability

In the two group scenario, power decreased as variance increased from four SDs to six SDs to eight SDs (results not shown). When sample sizes were equal, power was not affected by variance heterogeneity. However, when sample sizes were unequal, positive variance pairings (the larger sample size paired with the larger standard deviation) tended to be the most powerful, negative pairings (the larger sample size paired with the smaller standard deviation) the least powerful, and equal standard deviations falling in the middle. However, because power tended to be high across the board for the condition where the population mean difference was zero (Figure 1A) and low when the population mean difference was four (Figure 1C), this pattern was only clearly seen in the condition where the population mean difference was two (Figure 1B). The greater the imbalance in sample size, the greater the differences in power between the equal and unequal variance conditions.

In the three group scenario, the relationship between variance and sample size differed slightly

depending on whether a Games-Howell or Tukey post-hoc test was used. For the Games-Howell post hoc test, power tended to be highest when the variances were positively paired and lowest when the variances were negatively paired; the greater the sample size imbalance, the greater the differences among the three variance conditions. When the variances were positively paired, power was nearly 100% for all sample size conditions when there was no difference between population means (Figure 3A) and slowly increased from 67% to 92%, with almost no impact from sample size imbalances when there was a two point difference between population means (Figure 4A).When the variances were negatively paired, there were substantial drops in power as samples became more unbalanced. The equal variance conditions showed more modest drops in power for unbalanced sample sizes. For the Tukey post-hoc test, power tended to be fairly similar among the three variance conditions once total sample sizes were over 200. For total sample sizes less than 200, power tended to be highest for positively paired variances and lowest for negatively paired variances; a pattern that was consistent with the Games-Howell test and the two group comparisons.

### *Post-Hoc Test*

When variances were equal, the Games-Howell and Tukey post-hoc tests performed similarly, regardless of whether sample sizes were equal or unequal. When variances were unequal, the Games-Howell test showed greater fluctuations in power among the three variance conditions with higher power than Tukey when the variances were positively paired and lower power than Tukey when the variances were negatively paired.

### *Population Mean Difference*

Overall, when the samples were drawn from a population specified to have a mean difference of zero, the results yielded acceptable levels of power (i.e., ≥ 80%) for nearly all conditions in the two group scenario; only a few of the smallest sample size conditions, in combination with higher standard deviations, fell below the 80% criterion. For the three group condition and a zero population mean difference, power varied depending on the sample size, the sample variance and the post-hoc test used. For both the two and three group conditions, as the population mean difference increased, power decreased, especially for the condition in which the population mean configuration was set to four (results not shown in the three group case). For this latter configuration, there were no conditions for which power reached the 80% criterion.

### *Type I Error*

In the two group scenario, Type I error was maintained at 5% for all simulation conditions (Figure 2). In the three group scenario, Type I error rates were maintained at 5% or less for all conditions for which the variances and/or sample sizes were equal (Figure 5). When variances were negatively paired and sample sizes unbalanced, Type I error rates tended to be conservative for the Games-Howell post-hoc test. Type I error rates also tended to be conservative for the Tukey post-hoc test for samples under 200, but became liberal as sample sizes increased from 300, with increasingly, and unacceptably, high levels for the Type I error rate as the samples became larger and more unbalanced (full range of results not shown). When variances were positively paired, Type I error was acceptable for both post-hoc tests for all sample size conditions, although the Tukey post-hoc test was slightly more conservative.

## Discussion

When seeking to demonstrate that two or more groups are comparable, equivalence testing is the recommended method to use. Equivalence tests provide evidence that any differences that exist between groups are not meaningful and the groups can be treated as equivalent. A key first step in conducting these tests is to operationally define and justify the equivalence interval (i.e., the point at which differences are considered to be meaningful differences). Equivalence tests have been gaining popularity in education and the social sciences; however, there have been few studies that have investigated the statistical properties of this method and few guidelines provided for appropriately using tests of equivalence.

This study examined the statistical power and Type I error rate of the confidence interval approach to equivalence testing under varying conditions of group size, sample size, sample variability, and population mean difference. Knowing what happens to the power of a study when there are unequal sample sizes and/or variances will help to: (1) determine whether is it better to collect data from equal samples or to select a sample from within a larger sample to make the sample sizes more equal and (2) more accurately reach conclusions about the equivalence or comparability of two or more groups and the possible threats to internal validity of a study under non-optimal data conditions.

We demonstrated that group differences in sample size and variance did influence the power of equivalence tests when comparing both two and three groups, and that unequal sample sizes paired with unequal variances interacted to have a large impact on power (See Table 1 for a summary of the results); findings that are consistent with and extend the simulation studies done by Cribbie et al. (2010) and Gruman et al. (2007). As expected, if sample sizes were equal, increasing the total sample size increased power. However, reductions in power were seen as sample sizes became unbalanced, even with a constant total sample size.

Sample variability also impacted power, such that the lower the variance within groups, the greater the power; findings that are consistent with hypothesis testing models. What this study adds is its exploration of the combination of unequal sample sizes with unequal variances for various total sample sizes, thus allowing users of equivalence tests to make informed decisions regarding the appropriate use of these tests.

Table 1. Overall Summary of Simulation Results for Two and Three Group Comparisons

| Condition | Results |
|---|---|
| Sample size | Power increases as sample size increases |
| Equal versus unequal sample size | Equal sample sizes are more powerful than unequal sample sizes<br>Power decreases as samples become increasingly unequal |
| Variability | Power decreases as variability increases |
| Equal versus unequal variability | Impact of unequal variability depends on whether sample sizes are equal or unequal and choice of post-hoc test |
| Sample size and variability | When sample sizes are equal, power tends not to be affected by unequal variability, except for smaller sample sizes in the three group comparisons<br>When sample sizes are unequal, positive pairings tend to be the most powerful and negative pairings the least powerful; the greater the imbalance, the greater the differences in power among the three variance groups<br>Unequal sample sizes paired with unequal variances affected Type I errors for both Games-Howell and Tukey post-hoc tests in the three group comparison; Type I error was not affected in the two group comparison when the Welch-Satterhwaite correction was used |
| Post-hoc test | When sample sizes were equal, the two post-hoc tests performed similarly<br>When sample sizes were unequal, the power of the Games-Howell test was more strongly affected by unequal sample sizes and variances than the Tukey test |
| Population mean difference | Power decreases as the size of the difference in the population increases |

When sample sizes were equal, variance heterogeneity did not have an impact on power (except for the smallest sample sizes). However, when variances were unequal and were paired with unequal sample sizes, there was often a substantial impact on power. The exact nature of this impact depended on the extent of the imbalances in sample size and variability, whether two or three groups were being compared, and, in the three group case, which post-hoc test was being used; in general, positively paired variances (i.e., larger variance paired with larger sample size) were the most powerful, negatively paired variances (i.e., larger variance paired with smaller sample size) the least powerful, and the greater the disparity, the greater the differences in power.

When there was truly no difference between the population means (population mean configuration equal to zero), power tended to be quite high in all of the samples drawn across the conditions tested in this study. It is not surprising that power dropped as the difference in population means increased.

For the three group scenario, this study also sought to examine whether the Tukey or Games-Howell post-hoc test would be the most appropriate to use for equivalence tests involving three (or more) groups. There were differences in the patterns of results that depended on the extent to which sample sizes and variances were equal or unequal. When sample sizes were equal, Tukey and Games-Howell performed similarly. Differences in power were seen between these two tests for the three variance conditions (equal, positive, negative) as sample sizes became more unequal, with the Games-Howell post-hoc test showing larger fluctuations in power. While there were less fluctuations in power levels for the Tukey post-hoc test among the three variance conditions, there were still some large differences. Therefore, in deciding on which post-hoc test to use, it is important to first consider the sample variability. If sample variability is equivalent, either post-hoc test would suffice. If sample variability is positively paired, Games-Howell is the most powerful. If sample variability is negatively paired Tukey tends to be more powerful. However, if the total sample size exceeds around 300 and sample sizes are unequal, the Type I error rate for Tukey under these conditions becomes inflated, making the test invalid. It is not recommended to use the equivalence test under those conditions that produced an inflated Type I error rate. If Type I errors are below 5% this does not invalidate the test, but does make it more conservative.

What are the practical implications of the findings from this study? If a researcher or evaluator has control over sample size, we advise collecting data from groups such that each group is roughly equal in sample size. Equal group sizes will maximize power for a given total sample size and the power and Type I error rate will not be impacted if one were to find sample heterogeneity (i.e., violate the homogeneity of variance assumption).

If controlling sample size is not possible (which is often the case in applied settings), there are two possible options: (1) collect as much data as is possible, even if this results in unequal sample sizes or (2) sample from the larger group(s) to bring all group sizes into alignment. For example, let's say we are interested in comparing two groups where one group has a total possible sample size of 210 and the other group has a maximum sample size of 30. Further, let's say that the variances are negatively paired (this tends to be the more common situation as larger samples tend to provide better estimates of population parameters than smaller samples). If we assume that the groups are truly equivalent (population difference of zero), our power in this situation will be .89. If we were to equate the sample sizes by sampling only 30 from the larger class of 210, power drops slightly to .82, which is still within acceptable limits and requires less data collection. Thinking of this in the reverse, if we had two samples of 30 each, increasing the sample size in one group only results in a small increase in power. If it is difficult to collect data or if student survey burden is an issue, sampling may be an option, as long as there is sufficient power with both samples being at the smaller group size. Type I error is not a concern when comparing two groups via the t-test method, as long as the appropriate correction is made when variances are unequal (i.e., reading results from the "equal variances not assumed" row in the SPSS output).

If we extend this to a three group scenario, and add a third group of 30 students, our example comparison will now be between two groups of 30 and one group of 210. Let's assume the variances are negatively paired and the true difference among the groups is zero. If a Games-Howell post-hoc test is used power will be unacceptably low at .20 and if a Tukey post-hoc test is used power is still too low at .40. Looking at the corresponding Type I error rates, the Games-Howell test is conservative, while the Tukey test is slightly liberal, but both are generally in an acceptable range. Overall, this suggests that it may be inappropriate to conduct an equivalence test on these data because of a lack of power. What this means is that if you were to conduct an equivalence test on this data and found the groups to be equivalent then you can still have confidence in your conclusion of comparable groups because the conclusion is reached in spite of the study being underpowered. However, if you found the groups to be non-equivalent, it would be unclear as to whether this was because of the groups truly being non-equivalent or because sufficient power was lacking to detect equivalence (i.e., a Type II error). It is not until each group has at least a sample size of 60

that we see power at acceptable levels[2]. Building from this, if we were limited by two groups of 60, but had access to a larger third group, would it be better to collect data from the entire group or a subsample of 60? Similar to the two group scenario, collecting more data, even if it does create an unbalanced design, does result in more power; however, in many cases, the increase in power is very little and may or may not be worth the extra effort collecting the additional data. Additionally, one needs to keep in mind that if a Tukey post-hoc test is used, severely unbalanced groups will result in increased Type I error rates.

There are some limitations to this study that are important to consider. As this was a simulation study, the results are specific to the conditions investigated. While we tried to include a range of likely values and variables in conducting the simulations, not all ranges or variables could be modeled. Furthermore, all simulations were modeled as normally distributed. While the assessment data that were used to guide the selection of parameters and variables generally followed a normal distribution, in many cases, data often violate the assumption of normality. Further research is needed to investigate the impact of non-normality on equivalence tests, as well as other conditions that may be relevant to these tests. Finally, equivalence testing is a form of significance testing and is subject to the criticisms and misinterpretations of these types of tests (Thompson, 1994, 1999). However, equivalence tests have an advantage over traditional differences tests due to the use of an equivalence interval over a point-null hypothesis; a concept associated with the good-enough principle (Serlin & Lapsley, 1985).

## Conclusion

Stated broadly, the confidence interval approach to equivalence testing is the most powerful and valid when applied to equal sample sizes. When sample sizes are equal, the inequality of variances across the two groups will not impact the conclusions drawn from these tests. If unequal sample sizes are paired with unequal variances, this can result in dramatic differences in power and inflated or reduced type I error rates.

Inadequately powered studies can result in incorrect conclusions being drawn about the comparability of groups and can lead to a misuse of time and valuable resources. This study explored options for dealing with data that are not ideal. The figures provided can be used by researchers as guidelines for determining the minimum sample sizes needed for an appropriately powered study. Taking both sample size and variance into consideration when planning an analysis that address questions of comparability will result in more reliable, valid and generalizable results.

## References

Blackwelder, W. C. (2004). Current issues in clinical equivalence trials. *Journal of Dental Research, 83,* 113-115.

Cribbie, R. A., Arpin-Cribbie, C. A., & Gruman, J. A. (2010). Tests of equivalence for one-way independent groups designs. *The Journal of Experimental Education, 78,* 1-13.

Cribbie, R. A., Gruman, J. A., Arpin-Cribbie, C. A. (2004). Recommendations for applying tests of equivalence. *Journal of Clinical Psychology, 60*(1), 1-10.

Dunnet, C. W., (1980). Pairwise multiple comparison in the homogenous variance, unequal sample size case. *Journal of the American Statistical Association, 75,* 789-795.

Gruman, J. A., Cribbie, R. A., & Arpin-Cribbie, C. A. (2007). The effects of heteroscedasticity on tests of equivalence. *Journal of Modern Applied Statistical Methods, 6*(1), 133-140.

Nordstokke, D. W., Zumbo, B. D., Cairns, S. L., & Saklofske, D. H. (2011). The operating characteristics of the nonparametric Levene test for equal variances with assessment and evaluation data. *Practical Assessment, Research, & Evaluation, 16*(5). Available online: http://pareonline.net/getvn.asp?v=16&n=5.

Rogers, J. L., Howard, K. I., & Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Pyschological Bulletin, 113,* 553-565.

Rusticus, S. A. & Lovato, C. Y. (2011). Applying tests of equivalence for multiple group comparisons: Demonstration of the confidence interval approach. *Practical Assessment, Research & Evaluation*, *16*(7). Available online: http://pareonline.net/getvn.asp?v=16&n=7.

[2] Keep in mind that these power values are based on specific values for the standard deviations. If standard deviations were smaller, power would be greater and if standard deviations were greater, power would be reduced.

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedures and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics, 15,* 657-680.

Serlin, R. C., & Lapsley, D. K. (1985). The good-enough principle. *American Psychologist, 40*(1), 73-83.

Thompson, B. (1994). The concept of statistical significance testing. *Practical Assessment, Research & Evaluation, 4*(5). Available online: http:.//pareonline.net/getvn.asp?v=4&n=5.

Thompson, B. (1999). If statistical significance tests are broken/misused, what practices should supplement or replace them? *Theory & Psychology, 9*(2), 165-181.

## Acknowledgments:

The authors wish to thank Dr. Kevin Eva for helpful comments on the manuscript and Dr. Tavinder Ark and Dr. Bruno Zumbo for their assistance in writing some of the simulation syntax.

This study was supported by the Evaluation Studies Unit, Faculty of Medicine, University of British Columbia

## Citation:

Rusticus, Shayna A. & Lovato, Chris Y. (2014). Impact of Sample Size and Variability on the Power and Type I Error Rates of Equivalence Tests: A Simulation Study. *Practical Assessment, Research & Evaluation*, 19(11). Available online: http://pareonline.net/getvn.asp?v=19&n=11

## Corresponding Author:

Shayna Rusticus
Evaluation Studies Unit
Diamond Health Care Centre
2775 Laurel Street, 11th Floor
Vancouver, BC, V5Z 1M9
Contact: shayna.rusticus [at] ubc.ca