Paired-Samples Test of Equivalence

Constance Mara & Robert A. Cribbie

Quantitative Methods Program

Department of Psychology

York University

Send correspondence concerning this article to Rob Cribbie (cribbie@yorku.ca)

Abstract

Tests of equivalence are used when a researcher's objective is to find that two or more conditions or groups are nearly equivalent on some outcome variable, such that any difference is inconsequential within the framework of the research. Tests of equivalence are available for several different research designs, including two independent samples designs and one-way independent groups designs. However, paired-samples tests of equivalence that are accessible and relevant to the research performed by psychologists have been understudied. This study evaluated a paired-samples test of equivalence first introduced by Seaman and Serlin (1998) and compared it to a paired-samples test of equivalence developed by Wellek (2003) across several data conditions. Overall, Seaman and Serlin's paired-samples test of equivalence had better Type I error control and increased power over Wellek's test, and no information about the population standard deviation of the differences is required.

Keywords: test of equivalence, paired-samples

Paired-Samples Test of Equivalence

"Absence of evidence is not evidence of absence." – Sagan, 1995, p. 213

Psychologists often investigate differences between the means of two or more conditions or groups on some outcome variable. Traditional tests (e.g., *t*- and *F*-tests) are appropriate when the research question addresses differences. The null hypothesis for these traditional tests is that there is no difference between the group population means (e.g., H_0: $\mu_1 = \mu_2$), and the researcher seeks to reject this null hypothesis. However, when the research is investigating the equivalence of group means, researchers still commonly employ the use of traditional difference tests, and using non-rejection of the null hypothesis as grounds to conclude equivalence. One problem with this logic is that the probability of rejecting the null hypothesis increases as sample sizes increase. If a researcher is interested in demonstrating the equivalence of means, this result will be quite difficult or impossible to find with a statistically powerful study when the traditional difference tests are used. Further, equivalence will usually be found when studies are under-powered. Therefore, recommendations by statisticians since the late 1980's (e.g., Cribbie, Gruman, & Arpin-Cribbie, 2004; Rogers, Howard, & Vessey, 1993; Seaman & Serlin, 1998; Shuirmann, 1987) are to use tests of equivalence when the research question deals with issues of equivalence. However, this recommendation has not been widely adopted as common practice by researchers in psychology (as discussed later). The goals of this paper are to: 1) Inform psychological researchers of the availability of paired samples tests of equivalence and outline the situations for which these tests are recommended; and 2) Compare two available paired samples tests of equivalence under conditions that are common with psychological data.

Tests of equivalence have been used in biopharmaceutical studies for several decades in order to assess the equivalence of different medications (Seaman & Serlin, 1998). For example, a

new drug might be less expensive than a currently recommended drug, but in order to recommend the use of the new drug, its effects must be equivalent to the older, reliably used drug. In other words, the difference between the effects of the drugs must be so small that it is insignificant or unimportant within the context of the research. More recently, tests of equivalence have been introduced into psychological research, as their potential relevance within behavioural research has been recognized (Cribbie et al., 2004; Rogers et al., 1993; Seaman & Serlin, 1998). Researchers would use tests of equivalence, as opposed to the traditional difference tests, to determine if a population mean difference between two or more groups or conditions is small enough to be considered inconsequential. In traditional difference tests, the null hypothesis states (as mentioned previously) that the difference between the group or condition population means is equal to zero. In a test of equivalence, the null and alternative hypotheses are essentially the reverse of the hypotheses in the traditional difference tests. For tests of equivalence, the null hypothesis states that the difference between the group or condition population means falls *outside* a determined equivalence interval (i.e., $\mu_1 - \mu_2 \leq -\delta$ or $\mu_1 - \mu_2 \geq \delta$) and are therefore not equivalent. The equivalence interval is set by the researcher and represents the maximum difference between the population means that would be considered inconsequential in terms of the research conducted. The alternate hypothesis states that the difference between the population means falls *within* the equivalence interval (i.e., $\mu_1 - \mu_2 > -\delta$, or $\mu_1 - \mu_2 < \delta$).

One of the first tests of equivalence for two independent samples was developed by Schuirmann (1987). Schuirmann's two one-sided tests of equivalence uses two simultaneous one-sided *t*-tests to assess equivalence. The first step in this test is to set an equivalence interval that makes sense within the framework of the research. For example, a difference of $\delta = 5$ points

between population means might be considered inconsequential, resulting in an equivalence interval of (-5, 5). Thus, the null hypothesis is laid out as two hypotheses that must both be rejected in order to declare equivalence of the means. Specifically, H_{01}: $\mu_1 - \mu_2 \geq \delta$ states that the difference between the population means is greater than δ (e.g., 5) and H_{02}: $\mu_1 - \mu_2 \leq -\delta$ states that the difference between the population means is less than $-\delta$ (e.g., -5), and thus the means are not considered equivalent. The alternate hypothesis states that the difference between the population means falls within the equivalence interval (i.e., H_{11}: $\mu_1 - \mu_2 > -\delta$; H_{12}: $\mu_1 - \mu_2 < \delta$). Rejecting both of the null hypotheses implies that the difference between the means falls within the equivalence interval of $(-\delta, \delta)$ or (-5, 5), and the population means are therefore equivalent. It is important to note again that *both* of the null hypotheses must be rejected in order to declare the means equivalent. Cribbie et al. (2004) clarified that Schuirmann's test is more powerful at detecting equivalence (i.e., differences small enough to fall within the equivalence bounds) relative to Student's *t*-test as sample sizes increase, whereas Student's *t*-test is more likely to declare small differences significant as sample sizes increase. If the research question is dealing with equivalence, it is more logical to use tests that increase the chance of rejecting the null hypotheses associated with equivalence tests as sample sizes increase.

Using a traditional *t*-test when addressing questions of equivalence will often result in faulty conclusions. Specifically, if one has a large sample size and uses a traditional *t*-test to declare equivalence, too many Type II errors (i.e., declaring the groups not equivalent when they are equivalent) are likely to occur. If one has a small sample size and uses a *t*-test to declare equivalence, too many Type I errors (i.e., declaring the groups equivalent when they are not equivalent) are likely to occur. In essence, compared to a traditional *t*-test, the test of

equivalence's null and alternate hypotheses are reversed and as such, the definitions of Type I and Type II errors are reversed as well (see Figure 1).

Insert Figure 1

Establishing an Equivalence Interval

Establishing an equivalence interval ($-\delta$, δ) is a decision that should be customized by the researcher to their particular research question. A researcher should decide, *a priori*, what difference between the means would be considered insignificant within the context of their research. Because the nature of the outcome variables utilized by psychological researchers varies greatly, a "standard" or recommended equivalence interval is not practical or logical to propose. For example, an equivalence interval of one standard deviation might be inconsequential in one study, but might be a meaningful difference (i.e., not equivalent, non-ignorable) in another study. Essentially, an equivalence interval should define the difference that is of no practical importance for the particular research area. Establishing an equivalence interval requires knowledge of the behaviour or effect in question, and thus, is ultimately determined by the researcher's knowledge of the field.

There are some loose guidelines that can be provided for establishing and equivalence interval. For instance, the equivalence interval can take many forms, such as a raw value, percentage differences, or a percentage of the standard error or pooled standard deviation of the differences. Additionally, the equivalence interval is often symmetrical (see Westlake, 1976, for a discussion), as researchers often cannot predict with certainty which direction the difference between the means will occur. Indeed, Dunnett and Gent (1996) remark that most statisticians

strongly recommend that all tests be two-tailed, and thus the equivalence interval would be symmetrical. However, others argue there is some flexibility in this (Wellek, 2003), specifically when it is known that mean differences in a particular direction cannot occur, or distinguishing that mean differences in the opposite direction is irrelevant to the research goals. The argument for one- versus two-tailed tests and setting symmetrical versus asymmetrical equivalence intervals is beyond the scope of this paper; however Dunnett and Gent (1996) provide a good review of this argument for the interested reader.

Paired-Samples Tests of Equivalence

The traditional paired-samples *t*-test assumes that observations are correlated and removes variability due to inter-subject differences from the error term. Thus, the paired-samples *t*-test is more powerful than the independent samples *t*-test when observations are correlated or non-independent (see Zimmerman, 1997, for a discussion). Consequently, a paired-samples test of equivalence should also assume that observations are correlated. Because Schuirmann's two one-sided tests procedure assumes independence of observations, evaluation of an equivalence test for observations that are non-independent and which is easily accessible and relevant for the research conducted by psychologists is necessary.

Currently, researchers addressing the equivalence of paired-sample means usually look for a nonsignificant paired samples *t*-test. For example, Norlander, Bergman, and Archer (2002) examined the stability of personality traits in athletically-inclined individuals. They measured numerous personality traits at pretest, administered an intensive training over the course of a year designed to alter personality characteristics (e.g., optimism), and then re-measured the same personality traits at the end of the year. It was found that several personality traits were equivalent at pre- and posttest (i.e., impervious to change), given several nonsignificant paired-

samples *t*-tests. However, in order to assert this conclusion, the researchers would be more accurate to use a paired-samples test of equivalence.

In another example, Greig, Nicholls, Wexler and Bell (2004) examined the stability of schizophrenic patients on a number of neuropsychological tests. These researchers compared baseline to posttest on these measures. They used a paired-samples *t*-test to establish no change from baseline to posttest. However, these researchers would have benefitted from using a paired-samples test of equivalence in order to determine the equivalence of baseline and posttest scores.

We would like to highlight that we are not criticizing the statistical decisions made by the authors of these studies, as paired-samples tests of equivalence are currently not widely available to psychological researchers and are not available in popular statistical packages. Further, many of the tests that currently exist are not easily adoptable by psychological researchers.

To our knowledge, there are only a few paired-samples test of equivalence available (e.g., Wellek, 1993; Feng, Liang, Kinser, Newland, & Guilbaud, 2006). Using the same logic as Schuirmann's two one-sided tests procedure, Feng et al.'s (2006) test assesses the equivalence of drug concentration levels across different biopharmaceutical labs that are defined in terms of ratios instead of assessing differences in means. Psychological researchers are typically interested in mean differences or equivalence, and thus a test invoking the use of ratios is usually not practical in behavioural research. Other methods to assess equivalence have been developed in the field of biopharmacy that use binary probabilities to test bioavailability of drugs (see Lui & Zhou, 2004; Tang, 2003; Tang, Tang, & Wang, 2006). Again, these methods are often not practical or relevant for use in behavioural research.

Wellek's Paired-Samples Test of Equivalence

Wellek (2003) developed a test of equivalence that assesses the mean of the difference scores for paired observations, which is more relevant to the work behavioural scientists perform. The null and alternate hypotheses for the test developed by Wellek (2003) are:

$$H_0: \mu_D/\sigma_D \leq \theta_1 ; \mu_D/\sigma_D \geq \theta_2 \text{ vs. } H_1: \theta_1 < \mu_D/\sigma_D < \theta_2,$$

where θ is the specified standardized equivalence interval. To relate θ to δ (i.e., equivalence interval), θ would represent δ/σ_D. The population mean difference score divided by the population standard deviation of the differences is represented by μ_D/σ_D.

Wellek's test compares a *t*-statistic to a critical value in order to determine equivalence. The *t*-statistic can be obtained with the normal paired-samples *t*-test formula:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{sd_{Diff}/\sqrt{n}}$$

The test statistic is distributed as *t* with *n*-1 degrees of freedom and $\bar{x}_1 - \bar{x}_2$ is the mean of the difference scores. In order to determine the critical value of *t*, a noncentrality parameter (ncp) is determined using the equivalence interval, and can be defined as:

$$ncp = \frac{\delta}{\sigma_{Diff}/\sqrt{n}}$$

where δ is the equivalence interval. If |*t*| exceeds the critical value, $C = t_{\alpha,n-1,ncp}$, one would reject the null hypothesis and declare the means equivalent. Although the Wellek test is designed to evaluate hypotheses framed in standardized units, as one of the only paired samples tests of equivalence available it is conceivable that researchers would also utilize the test for hypotheses relating to raw mean differences by simply making an estimate of the population standard

deviation of the differences. Therefore, although psychological researchers rarely have info about the population standard deviation of the differences, we felt it was important to evaluate this procedure in situations in which researchers would make an estimate of the population's standard deviation.

Seaman and Serlin's (1998) Paired-Samples Test of Equivalence

Another paired-samples test of equivalence was introduced by Seaman and Serlin (1998), which borrows logic from Schuirmann's two one-sided tests procedure and the paired-samples *t*-test. This test frames the hypotheses in terms of raw mean differences, not standardized mean differences. Specifically, the null hypothesis states that the population mean difference score ($\mu_1 - \mu_2$) falls outside a determined equivalence interval (δ), and are therefore not equivalent (H_{01}: $\mu_1 - \mu_2 \geq \delta$; H_{02}: $\mu_1 - \mu_2 \leq -\delta$). Consequently, the alternate hypothesis states that the mean difference score is small enough to fall within the determined equivalence interval, and the population means are thus equivalent (i.e., H_{11}: $\mu_1 - \mu_2 < \delta$; H_{12}: $\mu_1 - \mu_2 > -\delta$). This differs from Schuirmann's two one-sided tests procedure for independent samples because the numerator contains the mean of the difference scores (rather than the difference between independent sample means) and assumes that the two samples are correlated. The null hypothesis is defined by two simultaneous predictions that both must be rejected in order to declare the mean differences in paired observations equivalent (where 'equivalent' is defined in terms of the established equivalence interval). H_{01} would be rejected if $t_1 \leq -t_{\alpha,n-1}$ and H_{02} would be rejected if $t_2 \geq t_{1-\alpha,n-1}$, where:

$$t_1 = \frac{\bar{x}_1 - \bar{x}_2 - \delta}{sd_{diff}/\sqrt{n}} \quad \text{and} \quad t_2 = \frac{\bar{x}_1 - \bar{x}_2 - (-\delta)}{sd_{diff}/\sqrt{n}}$$

$\bar{x}_1 - \bar{x}_2$ are the sample means, δ is the specified equivalence interval, and sd_{Diff} is the standard deviation of the difference scores. For simplicity, this test will be referred to as the Seaman-Serlin (SS) procedure in the remainder of the paper.

In order to be able to evaluate the properties of the Wellek and SS procedures, the next section of the paper will utilize a simulation study to evaluate how each test performs under data conditions thought to be common in psychological studies.

Method

A Monte Carlo simulation study was used to compare the Wellek test of equivalence to the SS procedure. Several variables were manipulated in this study, including sample size, correlation between paired observations, mean differences, distribution shapes, and the relationship between the true and the estimated population variance (see Table 1).

Insert Table 1

We compared the SS and Wellek tests using four sample sizes: 10, 25, 50 and 200. The equivalence interval was held constant at 1 for all simulations and the difference between the means was varied in order to examine power and Type I error control. The sets of means used in this study can be found in Table 1. Note that because the equivalence interval was set to 1 point, setting the populations mean difference ($\mu_1 - \mu_2$) equal to 1 represents a Type I error condition, and differing the population means by less than 1 point represents a power condition. The estimated population variance and the true population variance were also manipulated in order to determine how Wellek's test would perform if a researcher were to inaccurately estimate the value of the population variance. For example, simulations were conducted with the estimated

population variance and the true population variance both set to 1 (i.e., a correct estimation of the population variance), the estimated population variance set to 1.1 and true population variance set to 1 (i.e., overestimating the population variance), and with the estimated population variance set to 0.9 and the true population variance set to 1 (i.e., underestimating the population variance). The correlational structure between paired observations was also manipulated in order to determine what effects, if any, different magnitudes of correlation would have on both the Wellek and the SS paired-samples tests of equivalence. In particular, we ran simulations with the correlation between observations set at .5 and .8. The above conditions were investigated when the underlying distributions were normal as well as when the distributions were positively skewed. Given that distributions in psychology are frequently non-normal (Micceri, 1989), it is important that we investigate these procedures under common conditions of non-normality as well as optimal conditions of normal distributions. To generate a non-normal distribution with kurtosis = 4 and skewness = 1.63 (a moderately skewed distribution), the method recommended by Headrick and Sawilowsky (1999) using polynomial transformations was employed. The populations simulated for this study were specified in terms of their marginal, rather than their difference score, distributions because typically researchers are more likely to investigate the characteristics marginal distributions of their data than they are the difference score distributions. The alpha level was set to .05, and 20,000 simulations were conducted for each of the 120 conditions tested in this study. In order to evaluate the Type I error rates of the procedures, the bounds of $\pm 0.2\alpha$ was used. Therefore, with an alpha level of .05 a procedure would be considered to have an accurate empirical Type I error in a specific condition if the rate fell between .04 and .06.

Results

A complete summary of the results of the Monte Carlo simulations for all the conditions for N = 10, 25, 50, and 200 are presented in Tables 2 through 5, respectively.

Type I Error Control

Normal distribution. For normally distributed data, the SS test of equivalence performs consistently across conditions, maintaining the Type I error rate within the conservative bounds of .04 and .06. Further, Wellek's test performs well if the population standard deviation of the differences is accurately estimated. However, Wellek's paired samples test is does not perform well when the population standard deviation of the differences is not accurately estimated. Specifically, the empirical Type I error rates are smaller than the nominal alpha level when the population standard deviation of the differences is underestimated, and the empirical Type I error rates are inflated when it is overestimated.

Non-normal distribution. For non-normality, again, the SS procedure consistently maintains the Type I error rate at the nominal level within conservative bounds of .04 and .06. However, Wellek's test is inconsistent. Even if the variance estimation is accurate, the Type I error rate becomes inflated with mild non-normality. Interestingly, underestimating the variance in combination with non-normality creates an accurate Type I error rate, although this is obviously not a situation that researchers should strive for.

Power

Normal distribution. Given that Wellek's test performs poorly with regard to Type I error rates when inaccurately estimating the variance of the population, the power estimates are inaccurate for these conditions and are thus meaningless. Specifically, power is misleadingly increased when the population variance is overestimated, and power is reduced when the

population variance is underestimated (compared to the accurate variance estimation condition). However, even if the variance estimation is accurate, for $N = 10$, no power comparisons can be made between the two procedures because Wellek's test does not provide accurate Type I error rates.

With accurate variance estimation, $N = 25$, and a moderate correlation ($r = .50$) between the paired data, Wellek's test has a slight power advantage. However, as the correlation between paired data increases ($r = .80$), the advantage of Wellek's test diminishes and the SS test performs on par with Wellek's test. For larger sample sizes ($N = 50, 200$), the SS test gains an advantage over Wellek's test even in accurate variance estimation conditions. Additionally, as the correlation between paired data increases, the SS test of equivalence becomes increasingly powerful.

Non-normal distribution. Wellek's test demonstrated poor Type I error control with non-normal distributions across conditions and thus it is meaningless to interpret the power of Wellek's test for non-normal data in general and thus, the SS test cannot be compared to Wellek for this condition. Nevertheless, the power of the SS procedure in the non-normal distribution condition is similar to the power obtained for the SS procedure in the normal distribution condition.

Insert Tables 2-5

Calculating Power for the SS Paired-Samples Test of Equivalence

Calculating power is an important consideration for many researchers in psychology, so it is logical to include instructions on power calculations for paired samples tests of equivalence as

part of the current research (specifically for the SS procedure studied in this paper). First, a researcher would calculate two concurrent effect sizes in the normal way, but adding the equivalence interval into the equation, as demonstrated here:

$$d_1 = \left| \frac{\mu_1 - \mu_2 - \delta}{\sigma_{Diff}} \right| \qquad d_2 = \left| \frac{\mu_1 - \mu_2 - (-\delta)}{\sigma_{Diff}} \right|$$

It should be evident that the power for the tests of equivalence can only be calculated when the equivalence interval exceeds the expected difference in the means (i.e., if the equivalence interval is smaller than the expected mean differences, then the null hypothesis is true and power is irrelevant). The researcher would then choose the effect size that had the smallest absolute value from the equations calculated above:

$$d = \min(d_1, d_2)$$

Once an effect size (d) has been established, δ is calculated with the following formula and power is determined from a power table:

$$\delta = d\sqrt{n}$$

For clarity, we provide a brief example of how to calculate power for a paired samples test of equivalence. Specifically, for a sample size of 50, a researcher might find that a difference of 3 points (i.e., $\mu_1 - \mu_2 = 3$) on a questionnaire is a reasonable expectation, and that 4 is the typical standard deviation of the differences for this questionnaire from Time 1 to Time 2. An equivalence interval of 5 would adequately define the largest difference between the paired samples means that would be practically unimportant. Using the formulas provided above, the researcher would calculate the following effect sizes:

$$d_1 = \frac{30 - 27 - 5}{4} \qquad d_2 = \frac{30 - 27 - (-5)}{4}$$

$$d_1 = \frac{-2}{4} = |.5| \qquad d_2 = \frac{8}{4} = |2|$$

The smallest absolute value from the calculations above is .5, and this value is used to calculate δ:

$$\delta = .5\sqrt{25}$$

$$\delta = 2.5$$

Using a table that expresses power as a function of δ (e.g., Howell, 2009), a power estimate of .71 would be determined.

Discussion

It is important that researchers use the correct statistical tests for the research questions they address. As equivalence tests become more popular in psychological research, recommendations and guidelines for their appropriate use should be established. Generally, it is inappropriate to use non-rejection of the null hypothesis (in traditional tests) as grounds to conclude the equivalence of means. The current study examined the paired samples tests of equivalence developed by Wellek (2003) and Seaman and Serlin (1998). Generally, the SS test outperformed Wellek's test across most of the data conditions investigated. More specifically, the SS test maintained accurate Type I error rates across all conditions, whereas the Type I error rates for the Wellek test were not well controlled when the population standard deviation of the differences was not accurately estimated, or if the distributions demonstrated non-normality. Additionally, as the correlation between the paired data increases, the power of the SS test exceeded that of the Wellek test. Although Wellek's test performs almost as well as the SS test with large sample sizes and normal distributions, the Wellek test is still at a disadvantage because researchers must correctly identify the population standard deviation of the differences in order to calculate the equivalence interval. As mentioned previously, this information is

typically not available to researchers in psychology. The results of the current study suggest that the SS paired samples test of equivalence is the most apposite procedure available to psychological researchers.

Further research in this area could focus on expanding the current research to designs where is it desirable to establish equivalence over multiple time points. For example, researchers might be interested in demonstrating that mean depression scores do not differ over multiple follow up investigations (e.g., 6 months, 1 year, 2 years) following a clinical intervention. Additional research might also work towards development of new tests of equivalence that will evaluate the equivalence of group means in factorial designs. Specifically, a researcher might be interested in evaluating whether the effect of one variable is equivalent across all levels of another variable. A test of equivalence would be required to answer this question. Just as there are many different approaches to testing for differences under specific conditions, so too is it necessary to develop appropriate tests of equivalence for specific conditions.

To summarize, there are a wide range of equivalence tests available to researchers. For instance, if a researcher's purpose is to evaluate the equivalence of two independent group means, they could use the equivalence test developed by Schuirmann (1987; discussed previously) or Dannenberg, Dette, and Munk (1994). If a researcher would like to evaluate the equivalence of more than two means simultaneously, they could use a test developed by Wellek (2003). A researcher could also establish that there is no relationship between two continuous variables (Goertzen & Cribbie, in press) using a lack of association test. There is an increasing assortment of options available to the psychological researcher who wishes to establish equivalence or a lack of association in their research, although more research into these methodologies is essential.

References

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

Dannenberg, O., Dette, H., & Munk, A. (1994). An extention of Welch's approximate *t*-solution to comparative bioequivalence trials. *Biometrika, 81,* 91-101.

Dunnett, C.W. & Gent, M. (1996). An alternative to the use of two-sided tests in clinical trials. *Statistics in Medicine, 15,* 1729-1738.

Feng, S., Liang, Q., Kinser, R. D., Newland, K., & Guilbaud, R. (2006). Testing equivalence between two laboratories or two methods using paired-sample analysis and interval hypothesis testing. *Analytical and Bioanalytical Chemistry, 385,* 975-981.

Goertzen, J. R. & Cribbie, R. A. (in press). Detecting a lack of association: An equivalence testing approach. *British Journal of Mathematical and Statistical Psychology*.

Greig, T.C., Nicholls, S.S., Wexler, B.E., & Bell, M.D. (2004). Test-retest stability of neuropsychological testing and individual differences in variability in schizophrenic outpatients. *Psychiatry Research, 129,* 241-247.

Headrick, T.C. & Sawilowsky, S.S. (1999). Simulating correlated multivariate nonnormal distributions: Extending the Fleishman power method. *Psychometrika, 64,* 25-35.

Howell, D. C. (2009). *Statistical methods for psychology, 7th Edition,* Belmont, CA: Wadsworth.

Lui, K. & Zhou, X. (2004). Testing non-inferiority (and equivalence) between two diagnostic procedures in paired-samples ordinal data. *Statistics in Medicine, 23,* 545-559.

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

Norlander, T., Bergman, H., & Archer, T. (2002). Relative constancy of personality characteristics and efficacy of a 12-month training program in facilitating coping strategies. *Social Behavior and Personality, 30,* 773-784.

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

Sagan, C. (1995). *The demon-haunted world*. New York: Random House.

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

Seaman, M. A. & Serlin, R. C. (1998). Equivalence confidence intervals for two-group comparisons of means. *Psychological Methods, 3,* 403-411.

Tang, M. (2003). Matched-pair noninferiority trials using rate-ratio: A comparison of current methods and sample size refinement. *Controlled Clinical Trials, 24,* 364-377.

Tang, N., Tang, M., & Wang, S. (2006). Sample size determination for matched-pair equivalence trials using rate ratio. *Biostatistics, 0,* 1-7.

Wellek, S. (2003). *Testing statistical hypotheses of equivalence.* New York: Chapman & Hall/CRC

Westlake, W.J. (1976). Symmetrical confidence intervals for bioequivalence trials. *Biometrics, 32,* 741-744.

Zimmerman, D.W. (1997). Teacher's corner: A note of interpretation of the paired-samples t test. *Journal of Educational and Behavioural Statistics, 22,* 349-360.

Table 1

Conditions for the Monte Carlo simulation study.

Condition	Levels
N	n = 10
	n = 25
	n = 50
	n = 200
Distribution Shape	$\lambda_3 = 0$, $\lambda_4 = 0$ (normal)
	$\lambda_3 = 1.63$, $\lambda_4 = 4$ (positively skewed)
Population Means	$\mu_1 - \mu_2 = 1$ (Type 1 Error)
	$\mu_1 - \mu_2 = .8$ (Power)
	$\mu_1 - \mu_2 = .6$ (Power)
σ_{Diff}*	Actual = 1; Estimated = 1 (correct estimation)
	Actual = 1; Estimated = .9 (underestimation)
	Actual = 1; Estimated = 1.1 (overestimation)

* Standard deviation of the differences

Table 2.

Type I error rates and power, N = 10, and equivalence interval = 1.

Conditions		Type I error ($\mu_1 - \mu_2 = 1$)		Power ($\mu_1 - \mu_2 = .8$)		Power ($\mu_1 - \mu_2 = .6$)	
Vars**	*r**	Wellek	SS	Wellek	SS	Wellek	SS
Normal Distribution							
Equal	.5	.0695	.0443	.184	.125	.376	.285
Underestimated	.5	.0509	.0436	.153	.137	.348	.309
Overestimated	.5	.0880	.0418	.217	.122	.405	.264
Equal	.8	.0772	.0419	.253	.215	.554	.540
Underestimated	.8	.0492	.0441	.205	.226	.512	.583
Overestimated	.8	.1069	.0424	.311	.208	.597	.514
Non-normal Distribution							
Equal	.5	.0801	.0407	.183	.141	.344	.326
Underestimated	.5	.0639	.0401	.153	.152	.315	.356
Overestimated	.5	.0977	.0415	.209	.134	.374	.305
Equal	.8	.1015	.0408	.237	.256	.485	.600
Underestimated	.8	.0814	.0405	.200	.268	.432	.631
Overestimated	.8	.1249	.0410	.279	.236	.521	.568

Note. 'SS' refers to the paired samples test of equivalence introduced by Seaman and Serlin (1998). **'Vars' refers to the true population variance versus estimated population variance (applies only to the Wellek test and does not affect the SS test). *'*r*' refers to the correlation between paired data.

Table 3.

Type I error rates and power, N = 25, and equivalence interval = 1.

Conditions		Type I error ($\mu_1 - \mu_2 = 1$)		Power ($\mu_1 - \mu_2 = .8$)		Power ($\mu_1 - \mu_2 = .6$)	
Vars**	*r**	Wellek	SS	Wellek	SS	Wellek	SS
Normal Distribution							
Equal	.5	.0599	.0465	.263	.241	.602	.605
Underestimated	.5	.0358	.0488	.195	.259	.548	.647
Overestimated	.5	.0866	.0458	.314	.224	.644	.569
Equal	.8	.0661	.0480	.365	.446	.812	.921
Underestimated	.8	.0346	.0468	.274	.474	.749	.942
Overestimated	.8	.1068	.0473	.460	.415	.851	.897
Non-normal Distribution							
Equal	.5	.0801	.0423	.253	.256	..558	.624
Underestimated	.5	.0550	.0458	.201	.271	.497	.664
Overestimated	.5	.1087	.0472	.300	.244	.599	.597
Equal	.8	.1074	.0480	.351	.467	.731	.910
Underestimated	.8	.0736	.0462	.271	.503	.671	.931
Overestimated	.8	.1421	.0436	.416	.441	.780	.891

Note. 'SS' refers to the paired samples test of equivalence introduced by Seaman and Serlin (1998). **'Vars' refers to the true population variance versus estimated population variance (applies only to the Wellek test and does not affect the SS test). *'*r*' refers to the correlation between paired data.

Table 4.

Type I error rates and power, N = 50, and equivalence interval = 1.

Conditions		Type I error ($\mu_1 - \mu_2 = 1$)		Power ($\mu_1 - \mu_2 = .8$)		Power ($\mu_1 - \mu_2 = .6$)	
Vars**	*r**	Wellek	SS	Wellek	SS	Wellek	SS
Normal Distribution							
Equal	.5	.0559	.0477	.371	.399	.818	.872
Underestimated	.5	.0272	.0449	.274	.424	.763	.900
Overestimated	.5	.0999	.0489	.459	.373	.857	.840
Equal	.8	.0560	.0465	.529	.709	.960	.997
Underestimated	.8	.0251	.0476	.389	.745	.926	.998
Overestimated	.8	.1193	.0485	.651	.671	.974	.995
Non-normal Distribution							
Equal	.5	.0848	.0479	.351	.405	.777	.869
Underestimated	.5	.0485	.0464	.275	.440	.720	.899
Overestimated	.5	.1213	.0484	.427	.378	.819	.838
Equal	.8	.1051	.0467	.485	.713	.918	.995
Underestimated	.8	.0597	.0488	.373	.756	.868	.998
Overestimated	.8	.1619	.0475	.582	.682	.947	.992

Note. 'SS' refers to the paired samples test of equivalence introduced by Seaman and Serlin (1998). **'Vars' refers to the true population variance versus estimated population variance (applies only to the Wellek test and does not affect the SS test). *'*r*' refers to the correlation between paired data.

Table 5.

Type I error rates and power, N = 200, and equivalence interval = 1.

Conditions		Type I error ($\mu_1 - \mu_2 = 1$)		Power ($\mu_1 - \mu_2 = .8$)		Power ($\mu_1 - \mu_2 = .6$)	
Vars**	*r**	Wellek	SS	Wellek	SS	Wellek	SS
Normal Distribution							
Equal	.5	.0516	.0462	.787	.879	.999	.999
Underestimated	.5	.0119	.0485	.605	.910	.998	1.00
Overestimated	.5	.1476	.0499	.894	.851	.999	.999
Equal	.8	.0524	.0488	.939	.998	1.00	1.00
Underestimated	.8	.0078	.0503	.808	.999	1.00	1.00
Overestimated	.8	.1863	.0523	.984	.996	1.00	1.00
Non-normal Distribution							
Equal	.5	.0816	.0488	.744	.878	.999	.999
Underestimated	.5	.0306	.0483	.571	.907	.994	1.00
Overestimated	.5	.1801	.0501	.857	.848	.999	.999
Equal	.8	.1112	.0500	.882	.997	1.00	1.00
Underestimated	.8	.0352	.0483	.734	.999	.999	1.00
Overestimated	.8	.2435	.0484	.951	.994	1.00	1.00

Note. 'SS' refers to the paired samples test of equivalence introduced by Seaman and Serlin (1998). **'Vars' refers to the true population variance versus estimated population variance (applies only to the Wellek test and does not affect the SS test). *'*r*' refers to the correlation between paired data.

Figure 1. Traditional t-test versus tests of equivalence with respect to the conclusions of null hypothesis tests.

		POPULATION	
		Not Different/ Equivalent	**Different/ Not Equivalent**
SAMPLE	**Not Different/ Equivalent**	*t-test* – correct decision *Equiv* – Power	*t-test* – Type II error *Equiv* – Type I error
	Different/ Not Equivalent	*t-test* – Type I error *Equiv* – Type II error	*t-test* – Power *Equiv* – correct decision