Communications in Statistics—Simulation and Computation®, 41: 1928–1943, 2012

ISSN: 0361-0918 print/1532-4141 online
DOI: 10.1080/03610918.2011.626545

Taylor & Francis
Taylor & Francis Group

Paired-Samples Tests of Equivalence

CONSTANCE A. MARA AND ROBERT A. CRIBBIE

Department of Psychology, York University, Toronto, Ontario, Canada

Equivalence tests are used when the objective is to find that two or more groups are nearly equivalent on some outcome, such that any difference is inconsequential. Equivalence tests are available for several research designs, however, paired-samples equivalence tests that are accessible and relevant to the research performed by psychologists have been understudied. This study evaluated parametric and nonparametric two one-sided paired-samples equivalence tests and a standardized paired-samples equivalence test developed by Wellek (2003). The two one-sided procedures had better Type I error control and greater power than Wellek's test, with the nonparametric procedure having increased power with non normal distributions.

Keywords Test of equivalence; Paired-samples.

Mathematics Subject Classification 91Cxx; 91Exx; 62-07.

1. Equivalence Testing

Psychologists often investigate differences between the means of two or more conditions or groups on some outcome variable. Traditional tests of differences (e.g., t- and F-tests) are appropriate when the research question addresses differences. The null hypothesis for these difference-based tests is that the population means are equal ($H_0 : \mu_1 = \mu_2$), 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-based tests, using non-rejection of the null hypothesis as grounds to conclude equivalence. One problem with employing a traditional difference-based test for assessing equivalence is that the probability of rejecting the null hypothesis (that the population means are equal) increases as sample size increases. 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-based tests are used. Further, when using traditional difference-based tests, equivalence will usually be found when studies are under-powered. Therefore, recommendations by statisticians since the late 1980's (e.g., Berger and Hsu, 1996; Bross, 1985; Cribbie et al., 2004; Rogers et al., 1993; Schuirmann, 1987; Seaman

Received September 20, 2010; Accepted September 20, 2011
Address correspondence to Constance A. Mara, Department of Psychology, York University, Toronto, M3J 1P3 Ontario, Canada; E-mail: cmara@yorku.ca

and Serlin, 1998) are to use tests of equivalence when the research question deals with the similarities of groups or conditions. However, this recommendation has not been widely adopted as common practice by researchers in psychology (as discussed later). The goals of this article are to inform psychological researchers of the availability of paired samples tests of equivalence, outline the situations for which these tests are recommended, and compare three paired-samples tests of equivalence and the traditional Student's paired-samples t-test for differences 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 and 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 and Serlin, 1998)). Researchers would use tests of equivalence, as opposed to the traditional difference-based tests, to determine if the population mean difference between two or more groups or conditions is small enough to be considered inconsequential. In traditional difference-based 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-based 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. It is important to point out that the equivalence interval does not need to be symmetric (i.e., the interval $[-\delta, \delta]$ could be expressed as $[\delta_1, \delta_2]$, where $|\delta_1| \neq |\delta_2|$), however in most cases the interval is symmetric (Dunnett and Gent, 1996; Westlake, 1976). 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 for an equivalence test 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 $(-5, 5)$. 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 and H_{02}: $\mu_1 - \mu_2 \leq -\delta$ states that the difference between the population means is less than $-\delta$, and thus the means are not considered equivalent (for now the selection of δ is completely arbitrary, but below we discuss in more detail the selection of appropriate equivalence intervals). 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 $(-\delta, \delta)$, 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.

Using a traditional difference-based t-test when addressing questions of equivalence will often result in faulty conclusions (Cribbie et al., 2004). Specifically, if one has a large sample size and uses a traditional difference-based t-test to evaluate equivalence, too often the groups will be declared not equivalent when they are equivalent. If one has a small sample size and uses a t-test to declare equivalence, too often the groups will be declared equivalent when they are not equivalent. In essence, compared to a traditional difference-based t-test, a test of equivalence's null and alternate hypotheses are reversed.

Establishing an Equivalence Interval

Establishing an equivalence interval $(-\delta, \delta)$ 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.

1.1. *Paired-Samples Tests of Equivalence*

Currently, researchers addressing the equivalence of paired-sample means usually look for a nonsignificant paired-samples t-test. For example, Norlander et al. (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 pretest 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 et al. (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.

It is also important to highlight that a paired-samples test of equivalence should take into account that observations across conditions are correlated. For example, the traditional difference-based 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 take into account the non-independence of the observations in order to have a more powerful test of equivalence.

1.2. *Wellek's Paired-Samples Test of Equivalence*

Although other paired samples tests of equivalence have been formulated, very few apply to the type of research conducted in the behavioral sciences. Using the same logic as Schuirmann's two one-sided tests procedure, Feng et al. (2006) developed a test that 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 and Zhou, 2004; Tang, 2003; Tang et al., 2006). Again, these methods are often not relevant for use in behavioural research. Further, recent articless discussing more powerful methods for conducting tests of equivalence (e.g., Ennis and Ennis, 2010) are still controversial and may be excessively liberal with common psychological data (Bi, 2010; Castura, 2010).

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 $(-\theta,\ \theta)$ is the specified standardized equivalence interval. To relate θ to δ, 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:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sigma_{diff}/\sqrt{n}}.$$

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 non

centrality parameter (ncp) is determined using the equivalence interval, and can be defined as:

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

If $|t|$ is less than 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 this 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 standard deviation.

1.3. *Two One-Sided Test of Equivalence for Paired-Samples (TOST-P)*

An alternative paired-samples test of equivalence is based on Schuirmann (1987) two one-sided tests procedure (Seaman and Serlin, 1998). The two one-sided tests procedure for paired-samples (TOST-P) 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 ($-\delta, \delta$), 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., $H_{11} : \mu_1 - \mu_2 < \delta$; $H_{12} : \mu_1 - \mu_2 > -\delta$). 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:

$$t_1 = \frac{\bar{x}_1 - \bar{x}_2 - \delta}{\frac{s_{Diff}}{\sqrt{n-1}}} \quad \text{and} \quad t_2 = \frac{\bar{x}_1 - \bar{x}_2 - (-\delta)}{\frac{s_{Diff}}{\sqrt{n-1}}}.$$

$\bar{x}_1 - \bar{x}_2$ are the sample means, δ is the specified equivalence interval, and s_{Diff} is the standard deviation of the difference scores. It is important to highlight that this test can also be formulated as a one-sample test of equivalence, where $\bar{x}_1 - \bar{x}_2$ is replaced by the difference score.

1.4. *Nonparametric Two One-sided Test of Equivalence for Paired Samples (NPAR)*

Borrowing logic from the Wilcoxon (1945) signed ranks procedure, a nonparametric two, one-sided test of equivalence for paired samples will also be evaluated (NPAR), which is expected to be less susceptible to outliers than the Wellek and TOST-P procedures, which are based on the mean differences. First, signed ranks are computed separately for the observations $x_1 - x_2 - \delta$ and $x_1 - x_2 - (-\delta)$. Signed ranks are computed by ranking the observations (this is done separately for $x_1 - x_2 - \delta$ and $x_1 - x_2 - (-\delta)$) regardless of sign, and then attaching the original

sign to the computed ranks. Let sr_1 represent the absolute value of the sum of the negative ranks associated with $x_1 - x_2 - \delta$, and let sr_2 represent the sum of the positive ranks associated with $x_1 - x_2 - (-\delta)$. Then, $H_{01} : M_1 - M_2 \geq \delta$, where M represents the population median, is rejected if $z_1 \geq z_{1-\alpha}$, where:

$$z_1 = \frac{sr_1 - \left(\frac{N(N+1)}{4}\right)}{\sqrt{\frac{N(N+1)(2N+1)}{24}}},$$

and $H_{02} : M_1 - M_2 \leq -\delta$ is rejected if $z_2 \geq z_{1-\alpha}$, where:

$$z_2 = \frac{sr_2 - \left(\frac{N(N+1)}{4}\right)}{\sqrt{\frac{N(N+1)(2N+1)}{24}}}.$$

Rejection of H_{01} and H_{02} implies that difference between the medians falls within the equivalence interval.

In order to be able to evaluate the properties of the Wellek, TOST-P, and NPAR procedures, the next section of the article will utilize a simulation study to evaluate how each test performs under data conditions thought be common in psychological studies. The traditional Student's paired-samples *t*-test for differences will also be included in the current simulation study. It is important to note that Student's *t*-test does not test that same hypotheses as the equivalence tests, so direct comparisons are not logical. Instead, we will demonstrate the properties of this test when used to evaluate questions of equivalence (as discussed previously). For instance, it is expected that the null hypothesis of Student's *t*-test ($H_0 : \mu_1 = \mu_2$) will rarely be rejected with small sample sizes and nearly always be rejected with large samples.

2. Method

A Monte Carlo simulation study was used to compare the Type I error and power of the Wellek, TOST-P, and NPAR tests of equivalence. Only power was evaluated for Student's *t*-test because the difference in the means was never equal to zero. We are not aware of any simulation studies that investigate the properties of the Wellek, TOST-P, or NPAR tests. Further, it does not appear that Student's paired-samples *t*-test has yet been evaluated with respect to questions of equivalence.

Several variables were manipulated in this study, including the correlation between paired observations, mean differences, distribution shapes, and the relationship between the true and the estimated population variance (see Table 1).

For all simulations the equivalence interval was set at (−1, 1) and the sample size was set at $N = 10$, 25, 50, and 200. 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, 1), setting the population mean difference ($\mu_1 - \mu_2$) equal to 1 represents a Type I error condition (only for the equivalence procedures; not for Student's *t*-test), and differing the population means by less than 1 point represents a power condition.

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)

Note. λ_3 = skewness, λ_4 = kurtosis
*Standard deviation of the differences which only applies to Wellek's test.

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 the tests. In particular, we ran simulations with the correlation between observations set at .5 and .8.

The above conditions were investigated when the underlying distributions for the pretest and posttest variables 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 alpha level was set to .05, and 20,000 simulations were conducted for each of the conditions. 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 rate in a specific condition if the value fell between .04 and .06. The simulations were conducted with the open-source statistical software *R* (R Development Core Team, 2009).

3. Results

A summary of the results of the Monte Carlo simulations for all the conditions are presented in Tables 2–4 for $N = 10$, 25, and 50, respectively. Results for $N = 200$ were excluded because the conclusions mirrored those for $N = 50$.

3.1. *Type I Error Control*

3.1.1. Normal Distribution. For normally distributed data, the TOST-P and NPAR tests of equivalence performed consistently across all conditions, maintaining the Type I error rate within the bounds of .04 and .06 in all sample size conditions. Further, Wellek's test performed well if the population standard deviation of the differences was accurately estimated across all sample sizes. However, Wellek's paired samples test did not perform well when the population standard deviation of the differences was not accurately estimated. Specifically, the empirical Type I error rates were deflated relative to the nominal alpha level when the population standard deviation of the differences was underestimated, and the empirical Type I error rates were inflated when the population standard deviation of the differences was overestimated.

3.1.2. Non Normal Distribution. For non normality, again, the TOST-P and NPAR procedures consistently maintained the Type I error rate at the nominal level within conservative bounds of .04 and .06. However, Wellek's test often had empirical Type I error rates that exceeded the nominal level. Even if the variance estimation was accurate, the Type I error rate was inflated with mild non normality.

The Type I error rates for Student's paired-samples *t*-test could not be evaluated in this study since the differences between the means was not set to zero in any condition (which is necessary to test the point null hypothesis associated with Student's *t*-test). Further, all "power" conditions were actually "power" to detect equivalence, or, in the case of Student's *t*-test, for non rejection of the null hypothesis (since the null hypothesis for the paired-samples *t*-test is always false, this is measuring a Type II error).

3.2. *Power*

3.2.1. Normal Distribution. Given that Wellek's test performed poorly with regard to Type I error rates when inaccurately estimating the variance of the population, the power estimates were inaccurate for these conditions and are thus meaningless. Specifically, power was misleadingly increased when the population variance was overestimated, and power was reduced when the population variance was underestimated (compared to the accurate variance estimation condition).

With accurate variance estimation, there was very little difference between the power rates of the Wellek, TOST-P, and NPAR procedures across all sample size conditions.

3.2.2. 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. As expected, the NPAR procedure was more powerful than the TOST-P when the distributions were skewed.

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	ρ	Wellek	TOST-P	NPAR	Student[1]	Wellek	TOST-P	NPAR	Student[1]	Wellek	TOST-P	NPAR	Student[1]
						Normal Distribution							
Equal	.5	.0677	.0443	.0541	.1193	.182	.126	.148	.271	.374	.286	.318	.489
Under	.5	.0509				.146				.343			
Over	.5	.0880				.212				.411			
Equal	.8	.0748	.0449	.0545	.0026	.250	.208	.236	.028	.555	.546	.576	.153
Under	.8	.0492				.205				.508			
Over	.8	.1069				.308				.594			
						Non normal Distribution							
Equal	.5	.0811	.0405	.0534	.1267	.184	.146	.191	.259	.346	.328	.396	.441
Under	.5	.0634				.149				.315			
Over	.5	.0987				.198				.371			
Equal	.8	.1001	.0408	.0536	.0116	.238	.252	.316	.043	.474	.596	.670	.153
Under	.8	.0827				.204				.436			
Over	.8	.1245				.277				.516			

Note. Vars = relationship between the true population variance and the estimated population variance, which only affects the Wellek procedure; ρ refers to the correlation between paired data; TOST-P = the two one-sided testing procedure for equivalence introduced by Schuirmann (1981) applied to paired observations; NPAR = the non-parametric TOST procedure for equivalence applied to paired observations, Student = traditional Student's paired-samples t-statistic.

[1]This is actually measuring Type II errors, or "power" to detect equivalence.

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	ρ	Wellek	TOST-P	NPAR	Student[1]	Wellek	TOST-P	NPAR	Student[1]	Wellek	TOST-P	NPAR	Student[1]
						Normal Distribution							
Equal	.5	.0595	.0501	.0508	.0006	.259	.241	.246	.013	.598	.608	.607	.111
Under	.5	.0344				.196				.537			
Over	.5	.0907				.307				.645			
Equal	.8	.0646	.0501	.0516	.0000	.362	.437	.440	.000	.810	.917	.913	.002
Under	.8	.0336				.277				.749			
Over	.8	.1031				.457				.851			
						Non normal Distribution							
Equal	.5	.0774	.0500	.0524	.0021	.255	.260	.314	.020	..559	.623	.714	.118
Under	.5	.0548				.204				.499			
Over	.5	.1062				.294				.603			
Equal	.8	.0998	.0505	.0503	.0000	.348	.470	.573	.000	.733	.911	.954	.005
Under	.8	.0676				.275				.664			
Over	.8	.1421				.417				.777			

Note. Vars = relationship between the true population variance and the estimated population variance, which only affects the Wellek procedure; ρ refers to the correlation between paired data; TOST-P = the two one-sided testing procedure for equivalence introduced by Schuirmann (1981) applied to paired observations; NPAR = the non-parametric TOST procedure for equivalence applied to paired observations, Student = traditional Student's paired-samples t-statistic.

[1]This is actually measuring Type II errors, or "power" to detect equivalence.

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	ρ	Wellek	TOST-P	NPAR	Student[1]	Wellek	TOST-P	NPAR	Student[1]	Wellek	TOST-P	NPAR	Student[1]
						Normal Distribution							
Equal	.5	.0559	.0512	.0523	.0000	.363	.393	.390	.000	.821	.870	.864	.006
Under	.5	.0296				.276				.762			
Over	.5	.0974				.453				.863			
Equal	.8	.0560	.0470	.0484	.0000	.526	.706	.695	.000	.958	.997	.996	.000
Under	.8	.0251				.390				.932			
Over	.8	.1165				.650				.975			
						Non normal Distribution							
Equal	.5	.0807	.0488	.0497	.0000	.356	.402	.499	.001	.782	.867	.933	.009
Under	.5	.0485				.272				.720			
Over	.5	.1198				.432				.829			
Equal	.8	.1089	.0470	.0495	.0000	.492	.720	.833	.000	.915	.995	.999	.000
Under	.8	.0608				.375				.874			
Over	.8	.1607				.577				.944			

Note. Vars = relationship between the true population variance and the estimated population variance, which only affects the Wellek procedure; ρ refers to the correlation between paired data; TOST-P = the two one-sided testing procedure for equivalence introduced by Schuirmann (1981) applied to paired observations; NPAR = the non-parametric TOST procedure for equivalence applied to paired observations, Student = traditional Student's paired-samples *t*-statistic.

[1]This is actually measuring Type II errors, or "power" to detect equivalence.

As expected, Student's paired-samples *t*-test was more likely to suggest equivalence (via non-rejection of the null hypothesis) at small sample sizes ($N = 10$), but as sample sizes increased, the null hypothesis was always rejected, indicating differences between the paired-sample means. It is important to note that this is not incorrect, as the point null ($H_0 : \mu_1 - \mu_2 = 0$) was never correct in any of the conditions of this study, so it should always have been rejected. The inclusion of Student's paired-samples *t*-test for difference highlights the fact that using a test of differences to test for equivalence is inappropriate.

3.3. *Calculating Power for the TOST-P 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 TOST-P procedure studied in this articles). First, a researcher would calculate two concurrent effect sizes in the normal way, but adding the equivalence interval into the equation, as demonstrated here:

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

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 25, 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:

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

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

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, a power estimate of .71 would be determined.

3.4. *Empirical Examples*

In order to clarify the nature of the paired-samples equivalence tests, and to demonstrate the inappropriateness of the paired-samples t-test for questions of equivalence, we present two empirical examples. The first example uses a sample size for $N = 15$ and the second example uses a sample size of $N = 100$. In both cases, we use the TOST-P procedure as the test of equivalence.

Example 3.1. A researcher is interested in demonstrating that dyads composed of husbands and wives will have similar satisfaction with life scores ($N = 15$ pairs). The range of scores in this example is quite small, and thus it could be expected that anything larger than a one point difference in the means would be important, so the equivalence interval is set to $(-1, 1)$. After conducting a paired-samples equivalence test on these data, it is found that the dyads are not equivalent, $t_1 = -5.02$, $p_1 < .001$ and $t_2 = 1.61$, $p_2 > .05$. Thus, given that the t_2 statistic is not significant, the null hypothesis that the mean difference falls outside the equivalence interval cannot be rejected.

If Student's paired-samples t-test had been used to evaluate the equivalence of the means on the same data, we fail to reject the null hypothesis that the means are different, $t = -1.7$, $p > .05$. In other words, the paired-samples t-test is unable to detect a difference in the means and a researcher might be tempted to conclude that the means are therefore equivalent.

Example 3.2. A researcher is interested in the stability of personality traits. This researcher measures $N = 100$ participants at Time 1, and then measures the same individuals a year later at Time 2. The research hypothesis is that optimism is a stable personality trait, and thus optimism scores are expected to be similar from Time 1 to Time 2. Given that optimism scores are expected to be quite stable, the equivalence interval is set to a fairly narrow range of $(-1, 1)$. If the researcher uses a paired-samples equivalence test, they would reject the null hypothesis that the difference in the means falls outside the equivalence interval and conclude that the difference in the means is small enough to be considered inconsequential, $t_1 = -20.19$, $p_1 < .001$, $t_2 = 8.11$, $p_2 < .001$.

If the researcher had opted to use Student's paired-samples t-test, again, quite a different conclusion would be made. In this case, the researcher would reject the null hypothesis for the paired-samples t-test ($H_0 : \mu_1 = \mu_2$), and instead conclude that the means are different, $t = -6.04$, $p < .001$. It is important to remind the reader that since the research hypothesis relates to tbhe equivalence of the means, Student's paired-samples t-test is not appropriate.

4. 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 difference-based tests) as grounds to conclude the equivalence of means. The current study examined paired samples tests of equivalence developed by Wellek (2003) and alternative parametric and nonparametric versions of the two one-sided test procedure proposed by Schuirmann (1987). Generally, the TOST-P and NPAR tests outperformed Wellek's test across most of the data conditions investigated. More specifically, the TOST-P and NPAR tests 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 nonnormality. Although Wellek's test performs similarly to TOST-P and NPAR procedures with 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 NPAR test, as expected, was more powerful than the TOST-P test when the distributions were nonnormal, as the outlying cases have little effect on the NPAR procedure but they can increase the variability of the difference scores for the TOST-P procedure. To summarize, the results of the current study suggest that the TOST-P or NPAR paired samples tests of equivalence are most appropriate procedures with normal distributions, and the NPAR procedure is most appropriate with nonnormal distributions.

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 (i.e., lack of interaction). 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 et al. (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 and Cribbie, 2010) 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.

Acknowledgment

This research was supported in part by the Social Sciences and Humanities Research Council.

References

Berger, R. L., Hsu, J. C. (1996). Bio-equivalence trials, intersection-union test and equivalence confidence sets. *Statistical Science* 11:283–302.

Bi, J. (2010). Comments on D. M. Ennis' presentation on equivalence testing. *Food Quality and Preference* 21:259–260.

Bross, I. D. (1985). Why proof of safety is much more difficult than proof of hazard. *Biometrics* 41:785–793.

Castura, J. C. (2010). Equivalence testing: A brief review. *Food Quality and Preference* 21:257–258.

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.

Ennis, D. M., Ennis, J. M. (2010). Equivalence hypothesis testing. *Food Quality and Preference* 21:253–256.

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. (2010). Detecting a lack of association: An equivalence testing approach. *British Journal of Mathematical and Statistical Psychology* 63:527–537.

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.

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.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. URL: http://www.R-project.org.

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.

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* 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.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics* 1:80–83.

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.