British Journal of Mathematical and Statistical Psychology (2014)

The British Psychological Society

www.wileyonlinelibrary.com

Equivalence tests for comparing correlation and regression coefficients

Alyssa Counsell* and Robert A. Cribbie

York University, Toronto, Ontario, Canada

Equivalence tests are an alternative to traditional difference-based tests for demonstrating a lack of association between two variables. While there are several recent studies investigating equivalence tests for comparing means, little research has been conducted on equivalence methods for evaluating the equivalence or similarity of two correlation coefficients or two regression coefficients. The current project proposes novel tests for evaluating the equivalence of two regression or correlation coefficients derived from the two one-sided tests (TOST) method (Schuirmann, 1987, *J. Pharmacokinet. Biopharm*, *15*, 657) and an equivalence test by Anderson and Hauck (1983, *Stat. Commun.*, *12*, 2663). A simulation study was used to evaluate the performance of these tests and compare them with the common, yet inappropriate, method of assessing equivalence using non-rejection of the null hypothesis in difference-based tests. Results demonstrate that equivalence tests have more accurate probabilities of declaring equivalence than difference-based tests. However, equivalence tests require large sample sizes to ensure adequate power. We recommend the Anderson–Hauck equivalence test over the TOST method for comparing correlation or regression coefficients.

1. Introduction

Researchers are often interested in comparing population correlation (ρs) or regression coefficients (βs), and in many cases the interest is in demonstrating that the coefficients are equivalent. For example, Pillemer, Thomsen, Kuwabara, and Ivcevic (2013) examined memories about the self and whether these memories focused on achievement or interpersonal themes in participants from Denmark and the United States. One of their hypotheses was that 'relationships between emotional valence and thematic content should be consistent across cultures' (Pillemer *et al.*, 2013, p. 213). Clogg, Petkova, and Haritou (1995) describe comparing regression coefficients as the most common method in social science research for comparing two explanations of a particular predictor variable. In some instances, researchers would like to demonstrate that one model is better than another or simply that the models are different from one another. In other cases, researchers would like to conclude equivalence between the regression slopes of the two groups. Paternoster, Brame, Mazerolle, and Piquero (1998) give an example where researchers would like to determine whether there is a similar treatment effect from a correctional programme for first-time offenders compared to repeat offenders. They discuss this relationship within the context of traditional difference-based null hypothesis significance testing (NHST) but present research questions consistent with

*Correspondence should be addressed to Alyssa Counsell, Department of Psychology, York University, Toronto, ON M3J 1P3, Canada (email: counsell@yorku.ca).

DOI:10.1111/bmsp.12045

determining equivalence (e.g., does β_1 equal β_2?). These examples highlight research comparing independent groups' correlation or regression coefficients. Other research involves comparing correlation coefficients that are not independent. For example, a researcher may seek to validate a novel depression scale by demonstrating that the correlation between the depression scale and a measure of anxiety is equivalent to the correlation of a previously validated depression scale with the same measure of anxiety (i.e., $\rho_{12} = \rho_{13}$).

2. Current equivalence testing approaches

An issue in psychology is that behavioural researchers do not have access to methods in statistical software that evaluate whether correlation or regression coefficients are equivalent. In the examples given above, there were no readily available equivalence tests for the researchers to use in order to appropriately demonstrate equivalence between groups. Their strategy (and the most prevalent method in psychology) was to test for equivalence using traditional difference-based NHST, whereby a non-significant test statistic is deemed an indicator of equivalence. This approach is flawed for two main reasons. The first issue is theoretical: the purpose of difference-based tests is to detect a difference or relationship, not lack thereof. To quote the title of Altman and Bland's (1995) paper, 'absence of evidence is not evidence of absence'. In other words, accepting the null hypothesis is inappropriate for establishing equivalence because researchers can never statistically determine that the null hypothesis is true. The second issue is of practical concern for researchers: using difference-based tests with a null hypothesis of no difference has undesirable statistical properties if one's goal is equivalence. Non-rejection of the traditional null hypothesis (e.g., no difference) would essentially guarantee equivalence with small sample sizes because of low power to find differences. On the other hand, it would be difficult to conclude equivalence with large sample sizes since the null hypothesis would almost always be rejected due to high power to detect differences. As such, statistically non-significant results could be a function of insufficient sample size or poor research design. It is our opinion that researchers are unintentionally using an incorrect analysis because appropriate methods have not yet been developed, or the procedures have not been popularized in psychology. Below we discuss available equivalence testing methods before introducing methods for comparing correlation or regression coefficients.

3. Equivalence testing

Equivalence tests have been developed and tested over the past several decades in pharmacokinetics, where researchers often want to determine whether the effects of two drugs are equivalent. Numerous methods are available, although some are more widely used than others (e.g., Anderson & Hauck, 1983; Brown, Hwang, & Munk, 1997; Ennis & Ennis, 2009; Hsu, Hwang, Liu, & Ruberg, 1994; Schuirmann, 1987; Westlake, 1972, 1976). The first mention of equivalence testing for psychology was by Rogers, Howard, and Vessey (1993), who discussed the methods outlined by Schuirmann (1987) and Westlake (1972, 1976). Their paper highlighted how and why psychology should adopt the equivalence-based methods used in other disciplines. Since Rogers *et al.*'s (1993) paper, numerous researchers have discussed the utility of equivalence testing for psychological research (e.g., Cribbie, Gruman, & Arpin-Cribbie, 2004; Kendall, Marrs-Garcia, Nath, &

Sheldrick, 1999; Quertemont, 2011; Seaman & Serlin, 1998). Specifically, equivalence tests may be used to answer primary research questions in psychology, but also provide statistical analyses to justify other research or statistical decisions. For example, an equivalence test can justify pooling together groups or establish that two groups are equal at baseline before a treatment takes place (Rogers *et al.*, 1993). Equivalence testing could be also used in validation research. To test for discriminant validity, one could use an equivalence test for a lack of association (e.g., Goertzen & Cribbie, 2010) to demonstrate that a scale is not correlated with another scale measuring an unrelated construct. Equivalence tests also have applications for procedures such as meta-analysis, where data from studies rather than groups are pooled according to a similar criterion.

In equivalence testing, equivalence is not defined as a strict difference of zero, because with sampling error, it is difficult to find mathematical equivalence even if the true difference between population parameters is zero. Equivalence implies that the parameters are similar enough that there is no practical consequence to assuming that they are equal. To determine whether the parameters are equivalent, the researcher must choose an interval such that a parameter difference within their chosen interval can be considered inconsequential. This interval is called the equivalence interval $(-\delta, \delta)$, where δ represents the distance from zero to the edge of the interval in either direction. Choosing an appropriate equivalence interval will be elaborated in Section 9.4. Following the framework of NHST, equivalence tests employ null and alternative hypotheses. However, the null hypothesis is, for example, that there is a non-trivial difference between the population means – that is, the difference lies outside the prespecified equivalence interval. The alternative hypothesis is that the difference in the means falls within the equivalence interval. To tie these ideas into the primary purpose of the current paper, equivalence of regression and correlation coefficients involves demonstrating that the difference between the coefficients is so small that any differences can be considered trivial.

4. Equivalence tests comparing regression coefficients

We propose two equivalence tests for comparing independent regression coefficients. The first (TOST-β) is an equivalence-based version of the *t*-test of independent regression coefficients, based on the popular two one-sided tests (TOST) method originally developed to evaluate the equivalence of two group means (Schuirmann, 1987; Westlake, 1972).The first null hypothesis, H_{01}: $\beta_1 - \beta_2 \leq -\delta$, is rejected if $t_1 \geq t_{(1-\alpha, N-4)}$, and the second null hypothesis, H_{02}: $\beta_1 - \beta_2 \geq \delta$, is rejected if $t_2 \leq t_{(\alpha, N-4)}$, where

$$t_1 = \frac{b_1 - b_2 - (-\delta)}{s_{b_1 - b_2}}, \quad t_2 = \frac{b_1 - b_2 - \delta}{s_{b_1 - b_2}},$$

β is the population regression coefficient, b is the sample regression coefficient, and t is distributed on $N-4$ degrees of freedom. The standard error can be calculated by

$$S_{b_1 - b_2} = \sqrt{\frac{s^2_{Y \cdot X_1}}{s^2_{X_1}(n_1 - 1)} + \frac{s^2_{Y \cdot X_2}}{s^2_{X_2}(n_2 - 1)}},$$

where $s^2_{X_1}$ and $s^2_{X_2}$ are the variances of the independent variable for each group, and $s^2_{Y \cdot X_1}$ and $s^2_{Y \cdot X_2}$ are the error variances for each group. Rejection of both H_{01} and H_{02} is necessary

to conclude that the difference in regression coefficients is within the equivalence interval (i.e., $-\delta < \beta_1 - \beta_2 < \delta$).

Our second equivalence test of regression coefficients (AH-β) is based on Anderson and Hauck's (1983) equivalence test. Their procedure approximates a non-central *t* distribution to determine a *p*-value for the test. This test has only one null hypothesis, H_0: $\beta_1 - \beta_2 \leq -\delta$ or $\beta_1 - \beta_2 \geq \delta$, which is rejected when $p \leq \alpha$, where *p* is approximated by

$$p = \Phi\left[\frac{|b_1 - b_2| - \delta}{s_{b_1 - b_2}}\right] - \Phi\left[\frac{-|b_1 - b_2| - \delta}{s_{b_1 - b_2}}\right],$$

in which Φ represents the standard normal probability function. If $p \leq \alpha$, the regression coefficients are considered equivalent.

5. Comparing correlation coefficients

5.1. Sampling distribution of the population correlation coefficient

Research using correlation coefficients is affected by the sampling distribution of the population correlation parameter, ρ, because the distribution becomes increasingly skewed as ρ approaches ±1.00. With higher values of ρ, the sampling distribution will not be normal and the standard error cannot easily be estimated. This has important implications for comparing two independent correlation coefficients when running a *t*-test on the difference (Howell, 2009). Fisher (1921) demonstrated that by transforming *r* using the formula

$$r' = (0.5)\log_e\left|\frac{1+r}{1-r}\right|,$$

r' will become approximately normally distributed around ρ′ (the transformed ρ) with a standard error of $s_{r'} = 1/\sqrt{N-3}$ After transforming *r* to r', the issue of the skewed sampling distribution can be avoided by obtaining a standardized score of the difference between the independent *r*s. Since the value of δ is specified on a linear scale, δ' must be calculated so that the interval is on the same scale as the transformed *r*s. Transforming the equivalence interval was not found to be effective because the slight deviations in the difference between $r'_1 - r'_2$ and δ' created large discrepancies in the test statistic. Instead, δ was modified by the following statistic based on Fisher's *r* to *z* transformation:

$$\delta' = \left|(0.5)\log_e\left\{\frac{1+\left(\frac{r_1+r_2}{2}-\frac{\delta}{2}\right)}{1-\left(\frac{r_1+r_2}{2}-\frac{\delta}{2}\right)}\right\} - (0.5)\log_e\left\{\frac{1+\left(\frac{r_1+r_2}{2}+\frac{\delta}{2}\right)}{1-\left(\frac{r_1+r_2}{2}+\frac{\delta}{2}\right)}\right\}\right|.$$

Given the complication of transforming r_1 and r_2, we examined a few versions of equivalence-based *t*-tests comparing correlation coefficients. Our TOST-ρ used the transformation above, whereas the other two equivalence tests comparing correlations (KTOST-ρ, and AH-ρ) used a modified standard error instead (described in next section).

5.2. Equivalence tests comparing independent correlation coefficients

The first proposed test for correlation coefficients applied the TOST to the *t*-test of transformed correlation coefficients (TOST-ρ). Similar to the TOST comparing regression

coefficients, the two null hypotheses, H_{01}: $\rho_1 - \rho_2 \leq -\delta$ and H_{02}: $\rho_1 - \rho_2 \geq \delta$, are rejected when $z_1 \geq z_{1-\alpha}$ and $z_2 \leq z_\alpha$, where

$$z_1 = \frac{r'_1 - r'_2 - (-\delta')}{\sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}}, \quad z_2 = \frac{r'_1 - r'_2 - \delta'}{\sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}}.$$

Rejection of both null hypotheses implies that the correlation coefficients between the two groups are equivalent.

The second test we examined (KTOST-ρ) was the untransformed TOST proposed by Kraatz (2007). The null and alternative hypotheses are the same as the transformed tests of independent *ρs*. However, she used the following standard error described by Olkin and Finn (1995):

$$S_{r_1 - r_2} = \sqrt{\frac{(1 - r_1^2)^2}{(n_1 - 2)} + \frac{(1 - r_2^2)^2}{(n_2 - 2)}}.$$

Her test statistics are then

$$z_1 = \frac{r_1 - r_2 - (-\delta)}{s_{r_1 - r_2}}, \quad z_2 = \frac{r_1 - r_2 - \delta}{s_{r_1 - r_2}},$$

where the null hypotheses are rejected when $z_1 \geq z_{1-\alpha}$ and $z_2 \leq z_\alpha$, and rejection of both null hypotheses implies that the correlation coefficients are equivalent between the two groups.

The final equivalence test of independent correlation coefficients that we propose (AH-ρ) was based on Anderson and Hauck's (1983) procedure. We used the same standard error as Kraatz (2007) and after applying the Anderson–Hauck approximation, our proposed statistic was

$$p = \Phi\left[\frac{|r_1 - r_2| - \delta}{s_{r_1 - r_2}}\right] - \Phi\left[\frac{-|r_1 - r_2| - \delta}{s_{r_1 - r_2}}\right],$$

where Φ represents the standard normal probability function. When $p \leq \alpha$, the null hypothesis is rejected, and the researcher can conclude that the correlation coefficients are equivalent.

5.3. Equivalence tests comparing dependent correlation coefficients

A number of difference-based *t*-tests comparing dependent correlation coefficients exist in the literature (e.g., Dunn & Clark, 1969; Hotelling, 1931; Olkin, 1967; Williams, 1959). Our equivalence tests for comparing dependent correlations are based on Williams's modification to Hotelling's test for comparing overlapping dependent correlations. This test was used as the basis for the proposed equivalence tests because it has been compared to the other methods and has been recommended for its overall statistical properties (Boyer, Palachek, & Shucany, 1983; Hittner, May, & Silver, 2003; Steiger, 1980). The first proposed equivalence test comparing dependent correlation coefficients is based on the

TOST (TOST-ρ-D). Its two null hypotheses, H_{01}: $\rho_{12} - \rho_{13} \leq -\delta$ and H_{02}: $\rho_{12} - \rho_{13} \geq -\delta$, are rejected when $t_1 \geq t_{1-\alpha,N-3}$ and $t_2 \geq t_{1-\alpha,N-3}$, where

$$t_1 = [r_{12} - r_{13} - (-\delta)] \sqrt{\frac{(N-1)(1+r_{23})}{2\left(\frac{N-1}{N-3}\right)|R| + \frac{(r_{12}+r_{13})^2}{4}(1-r_{23})^3}},$$

$$t_2 = (r_{12} - r_{13} - \delta) \sqrt{\frac{(N-1)(1+r_{23})}{2\left(\frac{N-1}{N-3}\right)|R| + \frac{(r_{12}+r_{13})^2}{4}(1-r_{23})^3}},$$

in which $|R| = (1 - r_{12}^2 - r_{13}^2 - r_{23}^2) + (2r_{12}r_{13}r_{23})$. t_1 and t_2 are distributed on $N - 3$ degrees of freedom. When both t statistics are statistically significant, r_{12} and r_{13} are considered equivalent since the difference between the correlations falls within the equivalence interval.

Our second proposed equivalence test for dependent correlation coefficients (AH-ρ-D) uses Anderson and Hauck's (1983) formula. When $p \leq \alpha$, the null hypothesis, H_0: $\rho_{12}-\rho_{13} \leq -\delta$ or $\rho_{12}-\rho_{13} \geq \delta$, is rejected, and the researcher can conclude that the correlation coefficients are equivalent. The p-value of the tests is approximated as

$$p = \Phi\left[(|r_{12} - r_{13}| - \delta) \sqrt{\frac{(N-1)(1+r_{23})}{2\left(\frac{N-1}{N-3}\right)|R| + \frac{(r_{12}+r_{13})^2}{4}(1-r_{23})^3}}\right] - \Phi\left[(-|r_{12} - r_{13}| - \delta) \sqrt{\frac{(N-1)(1+r_{23})}{2\left(\frac{N-1}{N-3}\right)|R| + \frac{(r_{12}+r_{13})^2}{4}(1-r_{23})^3}}\right],$$

where Φ represents the standard normal probability function.

6. Study goals

There are four main goals of the current study: to demonstrate that traditional difference-based null hypothesis significance tests for comparing correlation coefficients or regression coefficients are inappropriate when the goal is to determine that these coefficients are equivalent; to evaluate and compare the power and Type I error rates of the equivalence tests using a range of sample sizes, effect sizes, and population correlation coefficients; to make recommendations for behavioural researchers about which of these novel tests are most practical for use; and to make available open-source software for conducting the recommended analyses (R functions to conduct the equivalence procedures discussed in this paper are available at http://www.psych.yorku.ca/cribbie/rfunctions.html).

7. Method

A two-part Monte Carlo study was used to evaluate the Type I error rates and power of the variations of equivalence tests for comparing regression and correlation coefficients. In

part one (independent ρs/βs), multivariate normal data were simulated using the `rmvnorm` function of the `mvtnorm` package in the open source statistical software program R (R Development Core Team, 2010). In each condition, values for r_1 and b_1 were simulated from a specified correlation matrix (ρ_1) and r_2 and b_2 were simulated from a second correlation matrix (ρ_2). Across all of the conditions, the first population correlation matrix remained the same while the second population correlation matrix differed from the first by a specified effect size. The following equivalence test statistics were utilized: TOST of independent βs (TOST-β), Anderson–Hauck test of independent βs (AH-β), TOST of independent ρs (TOST-ρ), Kraatz's test of independent ρs (KTOST-ρ), and Anderson–Hauck test of independent ρs (AH-ρ). Part two (dependent correlations) involved one population correlation matrix using the same data generation method described in part one. Here, the correlation matrix reflected three population correlations for one group, ρ_{12}, ρ_{13}, and ρ_{23}. Part two examined the TOST of dependent ρs (TOST-ρ-D) and the Anderson–Hauck test of dependent ρs (AH-ρ-D).

The study's manipulated variables were sample size, effect size, and values in the population correlation matrices (i.e., ρ_1 and ρ_2 for independent correlations and ρ_{12}, ρ_{13}, and ρ_{23} for the dependent correlations). The equivalence interval was set at (−.1, .1) for all of the conditions investigated because .1 is the smallest correlation acknowledged in Cohen's cut-offs for small, medium and large correlations (Cohen, 1962). In other words, a difference in correlations of .1 is deemed minimally important. Although the establishment of δ in this study was set somewhat arbitrarily since there was no predefined research question, one should note that an appropriate value of δ will vary depending on the nature of the study. The effect size naturally differed for the Type I error and power conditions. For part one (independent groups), a Type I error condition was created for the equivalence tests by setting the difference between ρ_1 and ρ_2 exactly equal to δ. This is where one would expect to see the highest probability of false rejections for the equivalence tests (i.e., α) since the difference is at the bounds of δ. In the power condition for the equivalence tests, the population parameters were either exactly equal ($\rho_1 - \rho_2 = 0$) or differed by .05 ($\rho_1 - \rho_2 = .05$). For each of the replications, the values of ρ ranged from .1 to .85. Both equal and unequal sample size conditions were investigated.

The simulation conditions were similar for part two (dependent correlations). The Type I error condition included a difference between ρ_{12} and ρ_{13} exactly equal to δ, and the two power conditions were $\rho_{12} = \rho_{13}$ or $\rho_{12} - \rho_{13} = .05$. For each of the replications, the values of ρ_{12} and ρ_{13} ranged from .1 to .85, and the values of ρ_{23} were .2, .4, and .6. We chose these parameter values to reflect values commonly encountered by behavioural science researchers.

The nominal α level was set at .05 for all analyses and 5,000 replications were conducted per condition. More detailed information about the simulation conditions and manipulated parameters for both parts of the simulation study are provided in Table 1.

Traditional difference-based tests comparing r_1 and r_2, b_1 and b_2, or r_{12} and r_{13} were also conducted in order to provide a comparison between the equivalence tests and difference-based methods when the goal is to find similarity between two variables. Note that when comparing equivalence and difference-based tests the Type I error and power terminology is reversed between the two methods (e.g., a Type I error condition for equivalence tests is a power condition in difference-based tests). As such, the number of rejections ($p \leq \alpha$) for equivalence tests was compared to the number of non-rejections ($p > \alpha$) in difference-based methods so that both tests reflect the probability of concluding equivalence. Bradley (1978) proposed a liberal robust interval for Type I error rates of $.5\alpha \leq p \leq 1.5\alpha$. Thus, in the current study, a test statistic's Type I error rates

Table 1. Summary of simulation parameters

Parameter	Values
δ	.1
Independent groups	
ρ_1	.1, .2, .3, .4, .5, .6, .7, .8
ρ_2	.1, .2, .3, .4, .5, .6, .7, .8, .15, 25, .35, .45, .55, .65, .75, .85
Effect size ($\rho_1 - \rho_2$)	0 (power), .05 (power), .1 (Type I error)
n (per group)	25, 50, 75, 100, 125, 150, 250, 375, 500, 750, 1,000, 1,500
Dependent	
ρ_{12}	.1, .2, .3, .4, .5, .6, .7, .8
ρ_{13}	.1, .2, .3, .4, .5, .6, .7, .8, .15, 25, .35, .45, .55, .65, .75, .85
ρ_{23}	.2, .4, .6
Effect size ($\rho_{12} - \rho_{13}$)	0 (power), .05 (power), .1 (Type I error)
N	50, 100, 250, 500, 1,000

will be deemed appropriate if the empirical Type I error rate falls within these bounds (.025, .075).

8. Results

Given that the pattern of results was similar across many conditions, only a subset of the results is presented below. Specifically, results for the conditions where $\rho_1 - \rho_2 = .05$, $\rho_{12} - \rho_{13} = .05$, and results for the unequal sample size conditions are omitted because they mirror the results for $\rho_1 = \rho_2$, $\rho_{12} = \rho_{13}$, and the equal sample size conditions.

8.1. Independent correlation and regression coefficients

8.1.1. Probability of falsely concluding equivalence

The probability of falsely concluding equivalence for each of the equivalence tests ($p \leq \alpha$) when the difference between ρs or βs falls at or outside $(-\delta, \delta)$ and the difference-based tests ($p > \alpha$) are presented in Table 2. The results were generally similar for the various equivalence tests of independent ρs and βs, although some important differences emerged.

At the smaller sample sizes, differences in Type I error rates were observed for the TOST-β and AH-β. Specifically, the Type I error rates for the TOST-β were too conservative at lower sample sizes; the null hypothesis was never rejected for sample sizes below 250 per group. This was not the case for the AH-β. Its Type I error rates were stable around α regardless of sample size. At the maximum sample size the error rates for the TOST-β and AH-β were similar. The same pattern of results was observed for the equivalence tests of independent ρs. However, error rates for all correlation-based equivalence tests increased as ρ approached 1.00. At the lower range of ρ, the results of the TOST-ρ were similar to the TOST-β, but empirical Type I error rates increased as the values of ρ increased. Error rates of the TOST-ρ were well maintained at α for group sample sizes of 100 for mid-ranged values of ρ but were too liberal at the highest values of ρ. The results from the KTOST-ρ were virtually identical to those of the TOST-ρ and the empirical Type I error rates for AH-ρ were consistently accurate and minimally affected by the value of ρ.

Table 2. Probability of falsely declaring equivalence for independent groups

ρ_1	ρ_2	*n* (per group)	Diff-β	Diff-ρ	TOST-β	AH-β	TOST-ρ	KTOST-ρ	AH-ρ
.1	.2	50	.912	.918	0	**.047**	0	0	**.052**
		100	.894	.894	0	**.048**	0	0	**.045**
		250	.799	.800	0	**.048**	0	0	**.044**
		500	.631	.630	0	**.050**	0	0	**.049**
		1,000	.378	.369	**.046**	**.048**	**.047**	**.047**	**.048**
.2	.3	50	.914	.912	0	**.052**	0	0	**.043**
		100	.883	.878	0	**.049**	0	0	**.050**
		250	.800	.785	0	**.053**	0	0	**.047**
		500	.631	.614	.003	**.047**	.009	.009	**.045**
		1,000	.364	.334	**.046**	**.047**	**.048**	**.047**	**.049**
.3	.4	50	.924	.915	0	**.050**	0	0	**.050**
		100	.882	.872	0	**.049**	0	0	**.052**
		250	.779	.753	0	**.047**	0	0	**.051**
		500	.603	.556	.008	**.054**	**.027**	**.027**	**.055**
		1,000	.322	.266	**.046**	**.047**	**.045**	**.045**	**.045**
.4	.5	50	.910	.902	0	**.050**	0	0	**.047**
		100	.880	.859	0	**.054**	0	0	**.049**
		250	.756	.709	0	**.052**	0	0	**.05**
		500	.573	.493	.022	**.052**	**.042**	**.041**	**.041**
		1,000	.300	.198	**.044**	**.045**	**.044**	**.044**	**.044**
.5	.6	50	.916	.895	0	**.049**	0	0	**.050**
		100	.862	.833	0	**.050**	0	0	**.048**
		250	.728	.633	0	**.048**	.004	.004	**.047**
		500	.525	.363	**.032**	**.048**	**.047**	**.047**	**.051**
		1,000	.243	.107	**.050**	**.050**	**.054**	**.052**	**.052**
.6	.7	50	.906	.872	0	**.054**	0	0	**.051**
		100	.850	.775	0	**.051**	0	0	**.052**
		250	.688	.505	0	**.051**	**.043**	**.041**	**.055**
		500	.448	.209	**.039**	**.044**	**.053**	**.053**	**.052**
		1,000	.156	**.025**	**.050**	**.050**	**.053**	**.051**	**.051**
.7	.8	50	.894	.809	0	**.052**	0	0	**.057**
		100	.817	.638	0	**.050**	.020	.016	**.056**
		250	.612	.263	.009	**.044**	**.055**	**.052**	**.052**
		500	.340	**.052**	**.052**	**.052**	**.064**	**.059**	**.059**
		1,000	.079	.001	**.050**	**.050**	**.056**	**.050**	**.050**
.8	.9	50	.845	.557	0	**.049**	**.072**	**.055**	**.049**
		100	.727	.259	0	**.049**	.081	**.066**	**.066**
		250	.438	.014	0	**.048**	.078	**.057**	**.057**
		500	.142	0	**.044**	**.050**	.079	**.057**	**.057**
		1,000	.01	0	**.048**	**.048**	.082	**.059**	**.059**

Notes. α = .05; 5,000 replications; Diff-β = difference-based regression test, Diff-ρ = difference-based correlation test, TOST-β = two one-sided tests regression equivalence method, AH-β = Anderson–Hauck regression equivalence test, TOST-ρ = two one-sided tests correlation equivalence method, KTOST-ρ = Kraatz's untransformed two one-sided tests correlation equivalence method, AH-ρ = Anderson–Hauck correlation equivalence test. Type I error condition for equivalence tests; power condition for difference-based tests. Numbers in bold indicate that the rates fall within Bradley's liberal robust interval (.025–.075).

The difference-based methods (Diff-ρ and Diff-β) displayed inappropriate rates of declaring equivalence when the group population values of ρ were not equal. The results of the Diff-ρ were affected more by the values of ρ than Diff-β. As seen in Table 2, the probability of falsely declaring equivalence was as high as .924 for sample sizes of 50 per group, with this number decreasing as sample size increased, and as the values of ρ_1 and ρ_2 increased. For sample sizes of 1,000 per group and ρs as high as .7 or .8, the rates decreased down to .01 or even exactly zero (indicating that the test now had sufficient power to detect a difference of $\rho_1 - \rho_2 = .1$).

8.1.2. Probability of correctly concluding equivalence

The results for the probability of concluding equivalence are presented in Table 3. When higher values of ρ were combined with large sample sizes, acceptable levels of power are observed (greater than .80) for all of the equivalence tests. However, power for the TOST-β test was zero for sample sizes less than 250 per group when values of ρ were less than .5. The AH-β displayed low power at smaller sample sizes, although it consistently had higher power than the TOST-β in these conditions. The TOST-ρ produced similar results to the TOST-β when ρ is around .5. At the higher range of ρ with the largest sample size, the power of the TOST-ρ reached 1.00. Similar results were observed for the KTOST-ρ. The AH-ρ also displayed problems with power for smaller sample sizes, especially at the lower range of ρ, but consistently had higher power than the TOST-ρ and KTOST-ρ. At sample sizes of 1,000 per group, the power of the AH-ρ, TOST-ρ, and KTOST-ρ were all similar.

Because the Diff-β and Diff-ρ are designed to find differences, the probability of concluding equivalence is maintained at approximately $100(1 - \alpha)\%$ regardless of sample size when the population correlations or regression coefficients are exactly equal. When there are slight differences between the groups (e.g., $\rho_1 - \rho_2 = .05$), rates of concluding equivalence expectedly decreased as sample size increased because the traditional tests' power to find a difference increases when *N* increases.

8.2. Dependent correlation coefficients

8.2.1. Probability of falsely concluding equivalence

Rates of falsely concluding equivalence when $\rho_{23} = .4$ are presented in Table 4. Results were similar for the other two tested values of ρ_{23}, although smaller values resulted in more conservative rates at small sample sizes for the TOST-ρ-D and larger values resulted in slightly more liberal error rates. Generally, the same pattern of results emerged for tests of dependent correlations as was seen when comparing independent groups' ρs. Specifically, the TOST-ρ-D displayed Type I error rates that were too conservative for sample sizes less than 500 per group, whereas the AH-ρ-D demonstrated accurate rates for all sample sizes investigated. The traditional test's (Diff-ρ-D) probability of falsely declaring equivalence was much too high at small sample sizes and decreased as sample size increased.

8.2.2. Probability of correctly concluding equivalence

Rates of concluding equivalence when $\rho_{23} = .4$ are presented in Table 5. Decreasing the value of ρ_{23} demonstrated a drop in power for both the equivalence tests, and increasing it

Table 3. Probability of declaring equivalence for independent groups

ρ_1	ρ_2	*n* (per group)	Diff-β	Diff-ρ	TOST-β	AH-β	TOST-ρ	KTOST-ρ	AH-ρ
.1	.1	50	.942	.942	0	.057	0	0	.056
		100	.948	.946	0	.061	0	0	.067
		250	.950	.952	0	.085	0	0	.095
		500	.949	.948	.004	.181	0	0	.183
		1,000	.950	.950	.458	.470	.466	.464	.476
.2	.2	50	.951	.951	0	.053	0	0	.058
		100	.948	.949	0	.065	0	0	.071
		250	.956	.956	0	.101	0	0	.101
		500	.953	.952	.008	.196	.009	.009	.202
		1,000	.948	.947	.476	.487	.512	.512	.518
.3	.3	50	.949	.950	0	.055	0	0	.055
		100	.945	.948	0	.065	0	0	.071
		250	.948	.949	0	.097	0	0	.103
		500	.945	.943	.021	.193	.069	.069	.212
		1,000	.949	.950	.517	.524	.573	.573	.575
.4	.4	50	.949	.956	0	.062	0	0	.059
		100	.953	.952	0	.073	0	0	.069
		250	.951	.956	0	.101	0	0	.115
		500	.947	.950	.059	.210	.188	.187	.268
		1,000	.948	.954	.575	.578	.702	.699	.700
.5	.5	50	.946	.949	0	.059	0	0	.061
		100	.947	.948	0	.070	0	0	.081
		250	.951	.950	0	.113	0	0	.152
		500	.955	.956	.138	.252	.359	.356	.383
		1,000	.954	.954	.651	.653	.821	.820	.820
.6	.6	50	.949	.945	0	.059	0	0	.069
		100	.953	.947	0	.075	0	0	.105
		250	.950	.951	0	.128	.086	.084	.222
		500	.953	.951	.254	.313	.593	.588	.591
		1,000	.950	.952	.744	.744	.935	.934	.934
.7	.7	50	.947	.954	0	.054	0	0	.082
		100	.945	.953	0	.076	.005	.004	.129
		250	.949	.947	.003	.176	.413	.403	.427
		500	.952	.946	.413	.433	.846	.841	.841
		1,000	.950	.947	.859	.859	.993	.993	.933
.8	.8	50	.946	.945	0	.071	.023	.018	.137
		100	.949	.948	0	.100	.265	.242	.316
		250	.941	.947	.167	.266	.859	.843	.843
		500	.947	.945	.670	.672	.992	.991	.991
		1,000	.957	.949	.969	.969	1.000	1.000	1.000

Notes. α = .05; 5,000 replications; Diff-β = difference-based regression test, Diff-ρ = difference-based correlation test, TOST-β = two one-sided tests regression equivalence method, AH-β = Anderson–Hauck regression equivalence test, TOST-ρ = two one-sided tests correlation equivalence method, KTOST-ρ = Kraatz's untransformed two one-sided tests correlation equivalence method, AH-ρ = Anderson–Hauck correlation equivalence test. Power condition for equivalence tests; Type I error condition for difference-based tests.

Table 4. Probability of falsely declaring equivalence when $\rho_{23} = .4$

ρ_{12}	ρ_{13}	*N*	Diff-ρ-D	TOST-ρ-D	AH-ρ-D
.1	.2	50	.906	0	**.046**
		100	.853	0	**.052**
		250	.699	0	**.053**
		500	.459	**.045**	**.051**
		1,000	.183	**.060**	**.061**
.2	.3	50	.906	0	**.055**
		100	.859	0	**.055**
		250	.695	0	**.053**
		500	.445	**.046**	**.053**
		1,000	.170	**.055**	**.055**
.3	.4	50	.892	0	**.046**
		100	.851	0	**.055**
		250	.677	0	**.048**
		500	.422	**.046**	**.051**
		1,000	.152	**.054**	**.055**
.4	.5	50	.892	0	**.049**
		100	.828	0	**.050**
		250	.657	0	**.049**
		500	.381	**.047**	**.050**
		1,000	.116	**.053**	**.053**
.5	.6	50	.887	0	**.053**
		100	.810	0	**.052**
		250	.611	.01	**.053**
		500	.319	**.044**	**.046**
		1,000	.077	**.048**	**.048**
.6	.7	50	.872	0	**.055**
		100	.770	0	**.051**
		250	.532	**.027**	**.047**
		500	.256	**.054**	**.056**
		1,000	.038	**.053**	**.054**
.7	.8	50	.848	0	**.055**
		100	.743	0	**.053**
		250	.425	**.052**	**.057**
		500	.135	**.051**	**.052**
		1,000	.011	**.052**	**.052**

Notes. $\alpha = .05$; 5,000 replications; Diff-ρ-D = difference-based test comparing dependent correlations, TOST-ρ-D = two one-sided tests dependent correlation equivalence method, AH-ρ-D = Anderson–Hauck equivalence test of dependent correlation coefficients. Type I error condition for equivalence tests; power condition for difference-based test. Bolded numbers indicate that the rates fall within Bradley's liberal robust interval (.025–.075).

resulted in higher power for both the TOST-ρ-D and AH-ρ-D. The same pattern of results was observed for different values of ρ_{23}. Both equivalence tests demonstrated inadequate power for sample sizes less than 500. However, a power advantage was observed for the AH-ρ-D at smaller sample sizes with similar results to the TOST-ρ-D at the largest sample size. Again, the rates of concluding equivalence for the Diff-ρ-D did not deviate regardless of sample size when $\rho_{12} - \rho_{13} = 0$, but decreased as sample size increased for the $\rho_{12} - \rho_{13} = .05$ condition.

Table 5. Probability of declaring equivalence when $\rho_{23} = .4$

ρ_{12}	ρ_{13}	*N*	Diff-ρ-*D*	TOST-ρ-*D*	AH-ρ-*D*
.1	.1	50	.948	0	.060
		100	.952	0	.078
		250	.945	0	.136
		500	.949	.316	.360
		1,000	.951	.786	.788
.2	.2	50	.957	0	.059
		100	.950	0	.073
		250	.946	0	.141
		500	.947	.308	.347
		1,000	.947	.782	.783
.3	.3	50	.953	0	.062
		100	.948	0	.076
		250	.951	0	.146
		500	.953	.356	.390
		1,000	.948	.812	.813
.4	.4	50	.950	0	.062
		100	.954	0	.084
		250	.949	.003	.166
		500	.951	.391	.415
		1,000	.947	.833	.835
.5	.5	50	.961	0	.071
		100	.954	0	.088
		250	.952	.024	.194
		500	.953	.492	.507
		1,000	.955	.895	.895
.6	.6	50	.951	0	.064
		100	.949	0	.090
		250	.952	.092	.228
		500	.944	.582	.585
		1,000	.949	.941	.942
.7	.7	50	.949	0	.074
		100	.954	0	.117
		250	.957	.241	.312
		500	.948	.745	.747
		1,000	.949	.979	.980
.8	.8	50	.946	.002	.088
		100	.950	.024	.164
		250	.945	.536	.551
		500	.948	.914	.915
		1,000	.949	.999	.999

Notes. $\alpha = .05$; 5,000 replications; Diff-ρ-D = difference-based test comparing dependent correlations, TOST-ρ-D = two one-sided tests dependent correlation equivalence method, AH-ρ-D = Anderson–Hauck equivalence test of dependent correlation coefficients. Power condition for equivalence tests; Type I error condition for difference-based tests.

9. Discussion

Equivalence testing has many useful applications in psychology. However, few behavioural researchers use equivalence tests, despite having research goals that are congruent with finding equivalence. One area where equivalence tests for comparing independent regression or correlation coefficients is especially relevant is research focusing on cultural similarities. Researchers often seek to demonstrate that relationships between variables (e.g., social support and depression) are consistent cross-nationally or cross-culturally. Several examples of other research designs in psychology that would benefit from equivalence testing are presented in previous papers (e.g., Cribbie *et al.*, 2004; Kendall *et al.*, 1999; Rogers *et al.*, 1993; Stegner, Bostrom, & Greenfield, 1996). This study provides researchers with appropriate methods to answer their research questions when the interest is in demonstrating equivalence or similarity of correlation or regression coefficients.

9.1. Inappropriateness of difference-based tests

Since equivalence testing methods are rarely employed in psychology, behavioural researchers continue to use difference-based methods whereby a non-rejection of the null hypothesis is interpreted as equivalence. When used to demonstrate similar relationships, traditional difference-based null hypothesis significance tests demonstrate inappropriate rates of concluding equivalence. For example, the probability of finding the effect associated with the alternative hypothesis should be directly related to sample size – increasing sample size should increase the chances of finding the effect. Since traditional difference-based methods are designed to test for differences, finding equivalence is contrary to their purpose. This results in a reverse relationship with sample size and finding an effect consistent with the researcher's hypothesis. In smaller sample sizes, equivalence is concluded approximately $100(1 - \alpha)\%$ of the time using the difference-based methods. Thus, when the values of ρ are exactly equivalent, the power to find equivalence did not change as sample size increased. When there were slight differences between ρs (but still within the equivalence interval), the probability of declaring equivalence was high with small sample sizes, but decreased as sample size increased. Observing a decrease in power to find the researcher's effect of interest when sample size increases is contrary to research practice. Researchers aim for a large sample size to ensure adequate power and generalizability of results. This issue suggests that difference-based methods are not valid for examining equivalence. Instead, researchers should use equivalence tests if they would like to find equivalence of two correlation or regression coefficients. These tests are theoretically justified and have desirable statistical properties (e.g., power to find the effect of interest increases as sample size increases; null hypotheses are rejected, not accepted).

9.2. Differences between the equivalence tests

As has been noted in the literature, the TOST procedure can be too conservative when power is low or variance increases (Berger & Hsu, 1996; Brown *et al.*, 1997). This simulation study confirmed this finding for the TOST-β, TOST-ρ, KTOST-ρ, and TOST-ρ-D. In contrast, the AH-based equivalence tests had accurate Type I error rates across all sample sizes and values of ρ. These findings are contrary to the literature that has reported that the Anderson–Hauck procedure's error rates can be overly liberal (Brown *et al.*,

1997; Ennis & Ennis, 2009; Nam & Munk, 1994). The power of each of the equivalence tests was almost identical for higher sample sizes. In the lower sample size conditions the Anderson–Hauck tests consistently had higher power than the TOST tests, although it was still unsatisfactory by the standards of behavioural researchers.

9.3. Low power

As has been mentioned, researchers ideally aim for power of at least .80. Looking at the power results of the equivalence tests clearly demonstrates inadequate power for sample sizes less than 1,000 per group for independent groups or 500 in the dependent correlation design. The power for all of the equivalence tests is considered to be low for the types of sample sizes that would typically be seen in psychology. However, current *t*-tests comparing regression or correlation coefficients display lower power in comparison to other popular statistical analyses such as analyses of mean differences (Howell, 2009). Since ρ must fall between -1.00 and $+1.00$, the difference between the two ρs relative to their standard error will typically be lower in comparison to a strict mean difference with no bounded scale. In other words, the difference-based null hypothesis significance tests comparing regression or correlations also display low power when used for their proper purpose of finding a difference. Researchers whose goal is to find either equivalence or difference between regression/correlation coefficients should be aware that this research design requires large sample sizes to ensure sufficient power. Another point worth mentioning is that power is also related to the size of the equivalence interval. If the specified equivalence interval is quite strict or small, power will be lower than if one had used a larger equivalence interval. Given that the study used simulations and did not include a substantive hypothesis regarding equivalence of regression or correlation coefficients, it is possible that the equivalence interval we used may have resulted in lower power than applied researchers would observe.

9.4. Choosing an equivalence interval

The magnitude of the equivalence interval is an important topic when discussing equivalence tests. Our equivalence interval, δ, was set at .1. We chose this value because we decided that it represents the smallest meaningful difference between two correlation or regression coefficients within the context of our simulation study. Values larger than .1 may represent an important effect for the relationship between two correlation coefficients. A smaller value for δ, such as .05, might be too stringent an interval to declare equivalence, especially given that *t*-tests comparing regression or correlation coefficients have lower power than other analyses such as *t*-tests comparing means. Cohen's standards for effect sizes also deem .1 a small effect size among correlation differences (Cohen, 1962). While these standards may provide a useful guide in choosing an equivalence interval, the most important aspect when selecting δ is to choose, *a priori*, a value that can be considered the largest difference between coefficients that a researcher would consider inconsequential. The equivalence interval could take the form of specific values, as was done in the simulation study, or the difference could correspond to a particular standardized effect size (e.g., an r^2 of .01 corresponds to 1% of the shared variance and an *r*-value of .1). While a common criticism is that setting δ introduces bias because a researcher may choose any value they like for δ, this is not a valid concern as δ must be theoretically justified. Rogers *et al.* (1993, p. 564) wisely noted that 'as with any statistical analysis, equivalency procedures must involve thoughtful planning by the

investigator'. As long as the researcher chooses a value for δ before collecting data, and the value is appropriate for the research problem being addressed, the researcher is in no way biasing his or her results.

9.5. Recommendations

Based on the results of the current simulation study, traditional difference-based *t*-tests for regression and correlation coefficients are not recommended when the researcher's goal is to establish equivalence. Their rates of falsely concluding equivalence are too high, and the best way to find equivalence is to have a small sample size. Thirty years ago Blackwelder (1982, p. 346) concisely stated that '*p* is a measure of the evidence against the null hypothesis, not for it, and insufficient evidence to reject the null hypothesis does not imply sufficient evidence to accept it'. Many have argued that researchers should never accept the null hypothesis, but researchers continue to do so when they use non-significant results to justify equality.

In conclusion, difference-based NHST is never a valid statistical method if the researcher's goal is to demonstrate equivalence. Equivalence testing procedures are the appropriate methods for finding equivalence of population parameters such as ρ or β. Specifically, we recommend the Anderson–Hauck equivalence test for comparing regression or correlation coefficients. This procedure was found to maintain accurate Type I error rates and demonstrated higher power than the TOST at smaller sample sizes. Given the lower power of statistical tests comparing regression and correlation coefficients, it is also recommended that researchers collect a large amount of data before running these statistical procedures to ensure adequate power for their hypotheses.

9.6. Limitations and future directions

One potential limitation is that all of the data used were simulated. Here, the data were all normally distributed, whereas the data typically found in psychology often demonstrate some skewness and kurtosis. An important task would then be to examine the performance of equivalence tests using data more typical of that encountered in psychological studies. Also, other conditions could have been tested such as different values of δ or using other sample sizes. However, these results would likely be predictable based on the results of the current study and are probably unnecessary. One future direction is to explore techniques for improving the power of equivalence tests when comparing regression and correlation coefficients for their employment in psychology, given their low power.

References

Altman, D. G., & Bland, J. M. (1995). Absence of evidence is not evidence of absence. *British Medical Journal*, *311*, 485. doi:10.1136/bmj.311.7003.485

Anderson, S., & Hauck, W. W. (1983). A new procedure for testing equivalence in comparative bioavailability and other clinical trials. *Statistics and Communications – Theory and Methods*, *12*, 2663–2692. doi:10.1080/03610928308828634

Berger, R. L., & Hsu, J. C. (1996). Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science*, *11*, 283–319. doi:10.1214/ss/1032280304

Blackwelder, W. C. (1982). 'Proving the null hypothesis' in clinical trials. *Controlled Clinical Trials*, *3*, 345–353. doi:10.1016/0197-2456(82)90024-1

Boyer, J. E., Palachek, A. D., & Shucany, W. R. (1983). An empirical study of related correlation coefficients. *Journal of Educational and Behavioral Statistics*, *8*, 75–86. doi:10.3102/10769986008001075

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*, 144–152. doi:10.1111/j.2044-8317.1978.tb00581.x

Brown, L. D., Hwang, J. T. G., & Munk, A. (1997). An unbiased test for the bioequivalence problem. *Annals of Statistics*, *25*, 2345–2367. doi:10.1214/aos/1030741076

Clogg, C., Petkova, E., & Haritou, A. (1995). Statistical methods for comparing regression coefficients between models. *American Journal of Sociology*, *100*, 1261–1293. doi:10.1086/230638

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, *65*, 145–153. doi:10.1037/h0045186

Cribbie, R. A., Gruman, J. A., & Arpin-Cribbie, C. A. (2004). Recommendations for applying tests of equivalence. *Journal of Clinical Psychology*, *60*, 1–10. doi:10.1002/jclp.10217

Dunn, O. J., & Clark, V. (1969). Correlation coefficients measured on the same individuals. *Journal of the American Statistical Association*, *64*, 366–377. doi:10.1080/01621459.1969.10500981

Ennis, D. M., & Ennis, J. M. (2009). Hypothesis testing for equivalence defined on symmetric open intervals. *Communications in Statistics – Theory and Method*, *38*, 1792–1803. doi:10.1080/03610920802460787

Fisher, R. A. (1921). On the probable error of a coefficient of correlation deduced from a small sample. *Metron*, *1*, 3–32. Retrieved from http://digital.library.adelaide.edu.au/dspace/bitstream/2440/15169/1/14.pdf

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. doi:10.1348/000711009X475853

Hittner, J. B., May, K., & Silver, N. C. (2003). A Monte Carlo evaluation of tests for comparing dependent correlations. *Journal of General Psychology*, *130*, 149–168. doi:10.1080/00221300309601282

Hotelling, H. (1931). The generalization of Student's ratio. *Annals of Mathematical Statistics*, *2*, 360–378. doi:10.1214/aoms/1177732979

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

Hsu, J. C., Hwang, J. T. G., Liu, H., & Ruberg, S. J. (1994). Confidence intervals associated with tests for bioequivalence. *Biometrika*, *81*, 103–114. doi:10.1093/biomet/81.1.103

Kendall, P. C., Marrs-Garcia, A., Nath, S. R., & Sheldrick, R. C. (1999). Normative comparisons for the evaluation of clinical significance. *Journal of Consulting and Clinical Psychology*, *3*, 285–299. doi:10.1037/0022-006X.67.3.285

Kraatz, M. (2007). *Correlational equivalence testing*. Unpublished master's thesis, Vanderbilt University, Nashville, TN.

Nam, Y. W., & Munk, A. (1994). On a method of combining double t-test and Anderson-Hauck test. *Biometrics*, *50*, 884–886.

Olkin, I. (1967). Correlations revisited. In J. C. Stanley (Ed.), *Improving experimental design and statistical analysis* (pp. 102–128). Chicago: Rand McNally.

Olkin, I., & Finn, J. D. (1995). Correlations redux. *Psychological Bulletin*, *118*, 155–164. doi:10.1037/0033-2909.118.1.155

Paternoster, R., Brame, R., Mazerolle, P., & Piquero, A. (1998). Using the correct statistical test for the equality of regression coefficients. *Criminology*, *36*, 859–866. doi:10.1111/j.1745-9125.1998.tb01268.x

Pillemer, D. B., Thomsen, D., Kuwabara, K. J., & Ivcevic, Z. (2013). Feeling good and bad about the past and future self. *Memory*, *2*, 210–218. doi:10.1080/09658211.2012.720263

Quertemont, E. (2011). How to statistically show the absence of an effect. *Psychologica Belgica*, *51*, 109–127. doi:10.5334/pb-51-2-109

R Development Core Team (2010). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing [computer software manual]. Retrieved from 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. doi:10.1037/0033-2909.113.3.553

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. doi:10.1007/BF01068419

Seaman, M. A., & Serlin, R. C. (1998). Equivalence confidence intervals for two-group comparisons of means. *Psychological Methods*, *3*, 403–411. doi:10.1037/1082- 989X.3.4.403

Stegner, B. L., Bostrom, A. G., & Greenfield, T. K. (1996). Equivalence testing for use in psychosocial and services research: An introduction with examples. *Education and Program Planning*, *19*, 193–198. doi:10.1016/0149-7189(96)00011-0

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87*, 245–251. doi:10.1037/0033-2909.87.2.245

Westlake, W. J. (1972). Use of confidence intervals in analysis of comparative bioavailability trials. *Journal of Pharmaceutical Sciences*, *61*, 1340–1341. doi:10.1002/jps.2600610845

Westlake, W. J. (1976). Symmetrical confidence intervals for bioequivalence trials. *Biometrics*, *32*, 741–744. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/1009222

Williams, E. J. (1959). The comparison of regression variables. *Journal of the Royal Statistical Society, Series B*, *21*, 396–399. Retrieved from http://www.jstor.org/stable/2983809

Received 24 February 2014; revised version received 26 July 2014