Recommendations for Applying Tests of Equivalence

Robert A. Cribbie
York University

Jamie A. Gruman
University of Windsor

Chantal A. Arpin-Cribbie
York University

Researchers in psychology reliably select traditional null hypothesis significance tests (e.g., Student's *t* test), regardless of whether the research hypothesis relates to whether the group means are equivalent or whether the group means are different. Tests of equivalence, which have been popular in biopharmaceutical studies for years, have recently been introduced and recommended to researchers in psychology for demonstrating the equivalence of two group means. However, very few recommendations exist for applying tests of equivalence. A Monte Carlo study was used to compare the test of equivalence proposed by Schuirmann with the traditional Student *t* test for deciding if two group means are equivalent. It was found that Schuirmann's test of equivalence is more effective than Student's *t* test at detecting population mean equivalence with large sample sizes; however, Schuirmann's test of equivalence performs poorly relative to Student's *t* test with small sample sizes and/or inflated variances. J Clin Psychol 60: 1–10, 2004.

Keywords: null hypothesis testing; equivalence tests; variance heterogeneity; Student's *t*

Researchers in psychology who are interested in comparing the means of two groups on a dependent measure reliably select traditional null hypothesis significance tests (e.g., Student's *t* test), in which the null hypothesis relates to the equivalence of the population

Correspondence concerning this article should be addressed to: Robert A. Cribbie, Department of Psychology, York University, 4700 Keele Street, Toronto, ON M3J 1P3, Canada; e-mail: cribbie@yorku.ca.

Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/jclp.10217

means. Further, traditional null hypothesis tests have reliably been applied regardless of whether the research hypothesis relates to whether the group means are different or are equivalent. For example, a clinical researcher may be interested in evaluating the research hypothesis that binge/purge anorexics and restricting anorexics have similar levels of resistance to treatment. This research hypothesis is distinctly different from the hypothesis that binge/purge anorexics and restricting anorexics have different levels of resistance to treatment, and, as discussed later, may have important implications for the statistical procedure that is adopted.

Anderson and Hauck (1983), Rouanet (1996), Schuirmann (1987), Selwyn and Hall (1984), Westlake (1976), and others have proposed statistical methods for determining if two groups are equivalent on a specific dependent measure, where the researcher determines an a priori minimal difference that is acceptable for declaring group means equivalent. These methods have recently been introduced to researchers in psychology through influential articles by Rogers, Howard, and Vessey (1993) and Seaman and Serlin (1998). Both articles focus on the test of equivalence proposed by Schuirmann (although, as Seaman & Serlin point out, Rogers et al. give Westlake, 1976, credit for this method). The Schuirmann test of equivalence has been extremely popular in biopharmaceutical studies for demonstrating bioequivalence although, until recently, tests of equivalence were rarely adopted by researchers in psychology even though research hypotheses dealing with equivalence are often investigated (Rogers et al., 1993).

There are at least two primary motivations for recommending tests of equivalence. First, the purpose of a study may not be to show that treatments are identical but only that the differences between the treatments are too small to be considered meaningful. Consider, for example, a clinical psychologist interested in studying two competing therapies for depression—one a long-term therapy and one a short-term therapy. The researcher may be interested in demonstrating that the treatment outcomes for the short-term therapy are equivalent to that for the long-term therapy, in addition to being less time consuming, less expensive, and so on. In this example, the researcher may not need to show that the therapies are "exactly equivalent" (as with the traditional null hypothesis, H_o: $\mu_1 = \mu_2$) but only that differences in treatment outcomes are not large enough to warrant adoption of the more time-consuming and expensive therapy (i.e., $[\mu_1 - \mu_2] < D$, where D represents an a priori critical difference for determining equivalence). Second, it is well known that as sample size increases, the probability of finding even minute (and potentially meaningless) mean differences statistically significant approaches unity with traditional equivalence null hypothesis tests, such as Student's two-independent samples t test. This is especially important given the increased number of requests from, for example, textbooks, journal editors, and statisticians for researchers to justify the meaningfulness of their results by including measures of effect size such as d, η^2, and so on (see Baugh & Thompson, 2001; Thompson, 2002a, 2002b). Therefore, with large sample sizes researchers may wish to specify a critical difference between treatments that would be considered clinically meaningful.

Rogers et al. (1993) demonstrated how the results of studies investigating the equivalence of two experimental groups using nonequivalence null hypothesis testing methodologies often contradict the results obtained with traditional equivalence null hypothesis tests. For example, Rogers et al. compared subjects addicted to alcohol and subjects addicted to drugs on the correction scale of the MMPI (based on a study by Cannon, Bell, Fowler, Penk, & Finkelstein, 1990), using both the test of equivalence proposed by Schuirmann and Student's t test. The authors found that the two group means were declared statistically different with Student's t test, yet statistically equivalent using Schuirmann's

test of equivalence. The authors also found that on the schizophrenia scale of the MMPI, the two groups were found not statistically different with Student's *t* test, yet not equivalent with Schuirmann's test of equivalence.

Given recent recommendations in support of equivalence testing and the increased availability of equivalence tests in psychological research, it is important that researchers have clear guidelines for applying tests of equivalence as well as when tests of equivalence may not be appropriate. For example, in the illustration presented earlier, there is no way of knowing whether MMPI scale scores of subjects addicted to alcohol are equivalent to the subjects addicted to drugs, and thus no way of knowing whether the results of Schuirmann's test of equivalence are correct or if the results of the traditional Student *t* test are correct. Therefore, the purpose of this study is to compare the currently recommended method for assessing the equivalence of two group means proposed by Schuirmann (1987) with the traditional Student *t*-test method. The following discussion will (a) review Schuirmann's test of equivalence, (b) discuss the application of Schuirmann's test of equivalence within an example provided by Seaman and Serlin (1998, p. 405), and (c) utilize a Monte Carlo study to compare Schuirmann's method with the Student *t*-test method for assessing population mean equivalence.

Schuirmann's Test of Equivalence

The first step in conducting Schuirmann's test of equivalence is to establish a critical mean difference for declaring two population means equivalent (D). Any mean difference smaller than D would be considered meaningless within the framework of the experiment. The selection of an equivalency interval (D) is an important aspect of equivalence testing that is primarily dependent on a subjective "level of confidence" with which to declare two (or more) populations equivalent. This level of confidence can take on many different forms including a raw value (e.g., mean test scores different than ten points), a percentage difference (e.g., $\pm 10\%$), a percentage of the pooled standard deviation difference, and so on. Tryon (2001) described this level of confidence as "an amount that is considered inconsequential" (p. 379), and Rogers et al. (1993) stated that "any difference small enough to fall within that equivalence interval would be considered clinically and/or practically unimportant" (p. 553). We recommend that researchers debating an appropriate value of D consider the nature of the research. For example, if the long-term therapy discussed earlier took three times as long and was three times as costly as the short-term therapy, then a more significant difference in outcomes (e.g., $+20\%$) might be needed to conclude that the therapies are equivalent than if the long-term therapy took one-and-a-half times longer and was roughly equivalent in cost to the short-term therapy (where a 5% increase in outcomes may be appropriate for concluding that the therapies are equivalent).

It is assumed that two samples are randomly and independently selected from normally distributed populations with equal variance. Two one-sided hypothesis tests can be used to establish equivalence, where the null hypothesis relates to the nonequivalence of the population means and can be expressed as two separate composite hypotheses:

$$H_{o1}\colon \mu_1 - \mu_2 \geq D;\ H_{o2}\colon \mu_1 - \mu_2 \geq -D.$$

Rejection of H_{o1} implies that $\mu_1 - \mu_2 < D$, and rejection of H_{o2} implies that $\mu_1 - \mu_2 > -D$. Further, rejection of both hypotheses implies that $\mu_1 - \mu_2$ falls within the bounds of $(-D, D)$ and the means are deemed equivalent.

H_{o1} is rejected if $t_1 \leq -t_\nu^\alpha$ where:

$$t_1 = \frac{(\bar{X}_1 - \bar{X}_2) - D}{\sqrt{\frac{(n_1 + n_2)[(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2]}{n_1 n_2 (n_1 + n_2 - 2)}}},$$

and H_{o2} is rejected if $t_2 \geq t_\nu^\alpha$ where:

$$t_2 = \frac{(\bar{X}_1 - \bar{X}_2) - (-D)}{\sqrt{\frac{(n_1 + n_2)[(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2]}{n_1 n_2 (n_1 + n_2 - 2)}}}$$

$\bar{X}_1$ and $\bar{X}_2$ are the group means, n_1 and n_2 are the group sample sizes, s_1 and s_2 are the group standard deviations, and t_ν^α is the upper-tailed α-level t critical value with $\nu = n_1 + n_2 - 2$ dedrees of freedom.

Seaman and Serlin (1998) Example

Schuirmann's test of equivalence was demonstrated by Seaman and Serlin (1998, p. 405) within the following example. The critical difference (D) for declaring the population means equivalent was set at 5, the nominal significance level was set at $\alpha = .05$, and:

$$\bar{X}_1 = 65.7, n_1 = 25, s_1 = 9$$
$$\bar{X}_S = 65.0, n_S = 25, s_2 = 8.$$

Substituting the sample statistics generates the following test statistics:

$$t_1 = \frac{(65.7 - 65.0) - 5}{2.4} = -1.8$$
$$t_2 = \frac{(65.7 - 65.0) - (-5)}{2.4} = 2.4.$$

Since $t_1\,(-1.8) < -t_\nu^\alpha\,(-1.68)$ and $t_2\,(2.4) > t_\nu^\alpha\,(1.68)$ the population means are declared equivalent (i.e., the difference between the means does not exceed the critical difference).

Seaman and Serlin also explain that for this example the same decision would have been reached if a traditional Student t test had been applied with the equivalence null hypothesis. Specifically, the null hypothesis (H_o: $\mu_1 = \mu_2$) would not have been rejected in this example, and the means would again have been declared statistically equivalent.

An interesting question that surfaces from the previous example is how often the two methods (Schuirmann's test of equivalence and Student's t test) generate the same conclusions (or more importantly, how often the methods generate different conclusions). This question is of utmost importance given that researchers in psychology routinely adopt Student's t test even when the research hypothesis relates to the equivalence of the means. A simulation study was used to determine the probability of declaring the population means equivalent with Schuirmann's test of equivalence (i.e., H_o is rejected) and Student's t test (i.e., H_o is not rejected), where the sample sizes and statistics from the previous example were utilized as population parameters in the simulations. It is expected that Schuirmann's test of equivalence will declare the population means equivalent more often than Student's t test given that the alternate hypothesis for Schuirmann's test (H_a: $\mu_L - \mu_S < 5$) encompasses a larger region than the null hypothesis for Student's

Table 1
Simulation Study of the Probability of Detecting Population Equivalence Using the Parameters ($\mu_1 = 65$, $\mu_2 = 65.7$, $\sigma_1 = 8$, $\sigma_2 = 9$, $D = 5$) From Seaman and Serlin (1997, p. 405)

	Sample Sizes			
Statistical Test	$n_1 = n_2 = 25$[a]	$n_1 = n_2 = 50$	$n_1 = n_2 = 75$	$n_1 = n_2 = 100$
Schuirmann's test of equivalence	.311	.768	.913	.974
Student's *t* test	.939	.934	.914	.912

[a]The actual sample sizes used in the Seaman and Serlin example.

t test (H_o: $\mu_1 - \mu_2 = 0$), and because the difference between the means ($65.7 - 65.0 = 0.7$) clearly falls within the critical difference of $D = 5$. Five thousand simulations were conducted using a nominal significance level of $\alpha = .05$. The results of the simulation study are presented in Table 1. Contrary to the underlying logic of the two methods, the results indicate that with the means, variances, and sample sizes from Seaman and Serlin's example, the probability of declaring the means equivalent with Student's *t* (.939) was significantly greater than with Schuirmann's test of equivalence (.311). Further, even when the sample sizes were increased to $n = 50$ and $n = 75$, Student's *t* test was at least as likely as Schuirmann's test of equivalence to declare the means equivalent. It is not until the sample sizes were increased to $n = 100$ that Schuirmann's test of equivalence became more likely than Student's *t* test to declare the means equivalent. However, these results are based on specific population parameters and may not be representative of the general effectiveness of the procedures. Therefore, a more extensive comparison of Schuirmann's test of equivalence and Student's *t* test is required.

Monte Carlo Study

A simulation study was used to compare Schuirmann's test of equivalence with the traditional Student *t* test for detecting population equivalence under conditions commonly experienced by researchers in psychology. Several variables were manipulated in this study including (a) sample size, (b) population mean configuration, and (c) population variances. The critical mean difference for establishing population equivalence with Schuirmann's test of equivalence was maintained at 1 throughout all conditions.

One of the primary motivations for utilizing tests of equivalence is that as sample size increases, the probability of finding even trivial mean differences statistically significant becomes large. Therefore, treatment group sample sizes were manipulated in this study. Specifically, group sample sizes were set at $n_1 = n_2 = 10$, $n_1 = n_2 = 25$, $n_1 = n_2 = 50$, and $n_1 = n_2 = 100$.

The effectiveness of a test of equivalence is directly affected by the heterogeneity of the group means. Five mean configurations were evaluated in this study, including equivalent population means ($\mu_1 = \mu_2$) and four nonequivalent population means ($\mu_1 = \mu_2 + .4$, $\mu 1 = \mu 2 + .8$, $\mu_1 = \mu_2 + 1.2$, and $\mu_1 = \mu_2 + 1.6$). Given that the critical difference for population equivalence is set at 1, the equivalent means configuration and nonequivalent configurations with $\mu_2 - \mu_1 < 1$ fall under the alternate hypothesis of Schuirmann's test of equivalence (i.e., the population mean difference does not exceed the critical mean difference, and thus the means are expected to be declared equivalent with

Schuirmann's test), and nonnull configurations with $\mu_2 - \mu_1 > 1$ fall under the null hypothesis of Schuirmann' test of equivalence (i.e., the population mean difference exceeds the critical mean difference, and thus the means are expected to be declared nonequivalent with Schuirmann's test). For Student's *t* test, any population mean difference greater than zero falls under the alternate hypothesis ($\mu_1 \neq \mu_2$) and the means are expected to be declared nonequivalent.

Another important question is what effect increasing the population variability will have on Schuirmann's test of equivalence. Also in relation to population variances, in a review of published research in education and psychology, Keselman et al. (1998) found that unequal variances were the norm, rather than the exception. Specifically, researchers often report largest to the smallest variance ratios as large as four, and largest to smallest variance ratios as large as eight were not uncommon. Therefore, in addition to investigating the case where the variance of both groups was set at one, the effects of population variance inflation was also investigated in this study by setting the variance of one of the groups to four or eight.

Five thousand simulations were conducted for each condition using a nominal significance level of .05.

Results

A Priori Equivalence

The probabilities of detecting population equivalence with Schuirmann's test of equivalence and Student's *t* test for the simulated conditions are presented in Table 2. When a priori population mean differences were less than the critical mean difference, sample size was a major factor in comparing Schuirmann's test of equivalence and Student's *t* test (It should be noted here that for Student's *t* test any population mean difference

Table 2
Probability of Detecting Population Equivalence for Schuirmann's Test of Equivalence (Critical Difference for Equivalence = 1) and Student's t Test

		$\sigma_1^2 = \sigma_2^2 = 1$		$\sigma_1^2/\sigma_2^2 = 4$		$\sigma_1^2/\sigma_2^2 = 8$	
n	$\mu_1 - \mu_2$	S	*t*	S	*t*	S	*t*
10	0	.3930	.9488	.0248	.9466	.0026	.9412
	0.4	.2764	.8576	.0220	.9076	.0016	.9234
	0.8	.0966	.5962	.0114	.7940	.0008	.8582
25	0	.9384	.9534	.4426	.9506	.0698	.9488
	0.4	.6778	.7174	.3016	.8314	.0526	.8922
	0.8	.1698	.2036	.1056	.5832	.0308	.7528
50	0	.9996	.9510	.8660	.9494	.5230	.9514
	0.4	.9100	.4962	.5922	.7676	.3512	.8530
	0.8	.2562	.0202	.1544	.2974	.1174	.5320
100	0	1.000	.9496	.9946	.9510	.9083	.9498
	0.4	.9956	.1922	.8444	.5688	.6324	.7330
	0.8	.3998	.0000	.2388	.0544	.1714	.2526

S = Schuirmann's test of equivalence; *t* = Student's *t* test.

renders the null hypothesis false, and therefore the goal of the test under all nonnull conditions is to identify the nonequivalence, rather than the equivalence, of the means. The proportion of nonrejections of the null hypothesis for Student's *t* with mean differences greater than zero [i.e., Type II errors] are presented for comparison with Schuirmann's procedure only.) With 10 or 25 observations per group, Student's *t* test was more likely to declare two group means equivalent than Schuirmann's test of equivalence, regardless of the size of the population mean difference. In fact, with 10 subjects per group Schuirmann's test of equivalence never detected equivalence in *more* than 50% of the cases whereas Student's *t* never detected equivalence in *less* than 50% of the cases. On the other hand, with 50 or 100 subjects per group, Schuirmann's test of equivalence was more likely than Student's *t* to detect equivalence.

When population variances were inflated in one group, Schuirmann's test was very poor at detecting population mean equivalence. Specifically, when the variance of one group was set at eight, Schuirmann's test of equivalence was never more effective at detecting equivalence than Student's *t*-test, regardless of sample size or population mean configuration. When the variance of one group was set at four, Schuirmann's test of equivalence was only more effective than Student's *t*-test when there were at least 100 subjects per group. When the population group means were equal, Student's *t*-test, as expected, found the group means equivalent approximately 95% of the time, regardless of sample size. On the other hand, Schuirmann's test of equivalence often found the group means to be nonequivalent, even when the population means were identical. For example, with 10 subjects per group, equal population means and the variance of one group set at four, Schuirmann's test of equivalence found the means to be equivalent in only 2.48% of the cases, whereas Student's *t* found the means to be equivalent in 94.66% of the cases. Increasing the sample size improved the performance of Schuirmann's test, although the test still performed poorly relative to Student's *t*. For example, with 50 subjects per group, equal population means and the variance of one group set at four, Schuirmann's test of equivalence found the means to be equivalent in 86.60% of the cases, whereas Student's *t* found the means to be equivalent in 94.84% of the cases.

A Priori Nonequivalence

The probabilities of detecting population nonequivalence with Schuirmann's test of equivalence and Student's *t* test for the simulated conditions are presented in Table 3. When a priori population mean differences were greater than the critical difference, Schuirmann's test of equivalence was very accurate ($>$95% in all cases) in detecting the differences. However, it is clear that the superior ability of Schuirmann's test to detect mean differences larger than the critical difference is a function of the test's bias for declaring the populations nonequivalent even when the differences were smaller than the critical difference. The power of Student's *t* for detecting mean differences was, as expected, affected by sample sizes and variances, with power maximized with larger sample sizes and small variances.

Discussion

The present article investigated an alternative to the traditional Student *t* test for detecting the equivalence of two treatment group means. There are many examples (provided here and elsewhere) of clinical psychology research paradigms in which the question of interest is whether one treatment mean is practically equivalent to a second mean, or in other

Table 3
Probability of Detecting Population Nonequivalence for Schuirmann's Test of Equivalence (Critical Difference for Equivalence = 1) and Student's t Test

		$\sigma_1^2 = \sigma_2^2 = 1$		$\sigma_1^2/\sigma_2^2 = 4$		$\sigma_1^2/\sigma_2^2 = 8$	
n	$\mu_1 - \mu_2$	S	*t*	S	*t*	S	*t*
10	1.2	.9832	.7254	.9928	.3598	.9982	.2290
	1.6	.9980	.9280	.9974	.5890	.9992	.3780
25	1.2	.9900	.9874	.9834	.7486	.9886	.5006
	1.6	.9998	1.000	.9988	.9414	.9980	.7446
50	1.2	.9942	.9998	.9856	.9584	.9798	.7928
	1.6	1.000	1.000	.9998	.9990	.9990	.9588
100	1.2	.9984	1.000	.9924	.9994	.9876	.9764
	1.6	1.000	1.000	1.000	1.000	1.000	.9998

S = Schuirmann's test of equivalence; *t* = Student's *t* test.

words, the difference between two treatment means is not large enough to be considered meaningful. Recent articles have presented tests of equivalence to researchers in psychology and recommended these procedures for answering questions relating to the equivalence of two group means. These articles have increased both the availability and the popularity of these procedures. However, there has been little research into the statistical properties of the procedures, and guidelines for applying tests of equivalence are practically nonexistent.

Two simulation studies were performed in this article: (a) an extension of the example used by Seaman and Serlin (1998) to demonstrate the application of Schuirmann's test of equivalence and (b) a comparison of Schuirmann's test of equivalence and Student's *t* test under many conditions commonly experienced in behavioral science experiments. The purpose of the simulation studies was to highlight the statistical properties of tests of equivalence and how they relate to traditional null hypothesis testing. It should be noted that many of the results reported in this article could be predicted based on the underlying sampling distributions of the test statistics, although it is important that these results be quantified to allow researchers to make informed decisions in the selection of an appropriate test statistic. Both studies demonstrated that sample size is a crucial factor in deciding between Schuirmann's test of equivalence and the traditional Student *t* test. If the number of subjects per condition is large (25 or more), Schuirmann's test of equivalence can be more appropriate for detecting population equivalence than Student's *t*, especially when population mean differences are present but less than the critical difference. As the sample size and the difference between the means increase, Student's *t* test, as expected, becomes more powerful at detecting the differences and is thus less likely to declare the differences meaningless. On the other hand, as sample size increases, Schuirmann's test of equivalence becomes more powerful at detecting that these differences are less than the critical difference, and is therefore recommended over Student's *t* with large sample sizes.

However, an important qualification to the previous recommendation regarding sample size is required; when the group variances are even moderately inflated, the ability of Schuirmann's procedure to detect equivalence is substantially reduced. Although this finding is partially a result of the decreased power (i.e., increased standard error) of

Student's *t* for detecting mean differences with inflated variances, the extreme effect of inflating the variances on the Schuirmann test of equivalence (especially with small sample sizes) should not be overlooked. More specifically, it is extremely problematic that when there are no differences between the means (and the probability of concluding that the groups are equivalent with Student's *t* is approximately .95), that Schuirmann's test of equivalence performs so poorly with small samples sizes and/or increased variances. Further, it is also important to recognize that, although it is not the focus of this paper, unequal sample sizes that are paired with unequal variances can significantly affect the Type I and Type II error rates of Student's *t*, beyond what was observed in this study. It is expected that the effect of unequal samples sizes and variances will also significantly impact on the Type I and Type II error rates of Schuirmann's test of equivalence, and therefore research into a robust test of equivalence for unequal sample sizes and variances is necessary.

Although the results of this study are important in terms of providing recommendations for researchers regarding the selection of an appropriate test statistic for evaluating the equivalence of two group means, there are two important limitations of the current research that should be considered. First, the Monte Carlo study was unable to investigate all potential conditions of sample size, variance inequality, and so on; therefore, although we expect the results of this study to generalize to many common testing environments, these results are specific to the conditions investigated in this study. Second, although we have focused on the hypothesis-testing framework of evaluating the equivalence of means in this article, there have been important advances in the application of confidence interval approaches to equivalency testing that were beyond the scope here but that may be of interest to readers (see, e.g., Seaman & Serlin, 1998; Tryon, 2001).

To summarize, tests of equivalence are extremely popular in biopharmaceutical studies for demonstrating that the effects of two drugs are practically equivalent. It is expected that as the number of studies outlining the methodologies of equivalence tests grows, the popularity of tests of equivalence will increase in the field of psychology, given that researchers will be more prepared to identify situations in which equivalency tests are appropriate. Therefore, it is important that clear recommendations exist for applying these tests. The findings of this study emphasize the need to acknowledge that Schuirmann's test of equivalence and Student's *t* are diametrically opposed in their approach to hypothesis testing, and thus the same factors that significantly affect the power of Student's *t* to detect differences between means (e.g., sample size, error variability) also significantly affect the power of Schuirmann's test to detect equivalence.

References

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

Baugh, F., & Thompson, B. (2001). Using effect sizes in social science research: New APA and journal mandates for improved methodology practices. Journal of Research in Education, 11, 120–129.

Cannon, D.S., Bell, W.E., Fowler, D.R., Penk, W.E., & Finkelstein, A.S. (1990). MMPI differences between alcoholics and drug abusers: Effects of age and race. Psychological Assessment: A Journal of Clinical and Consulting Psychology, 2, 51–55.

Keselman, H.J., Huberty, C.J., Lix, L.M., Olejnik, S., Cribbie, R.A., Donahue, B., Kowalchuk, R.K., Lowman, L.L., Petoskey, M.D., Keselman, J.C., & Levin, J.R. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. Review of Educational Research, 68, 350–386.

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.

Rouanet, H. (1996). Bayesian methods for assessing importance of effects. Psychological Bulletin, 119, 149–158.

Schuirmann, D.J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing 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.

Selwyn, M.R., & Hall, N.R. (1984). On Bayesian methods for bioequivalence. Biometrics, 40, 1103–1108.

Thompson, B. (2002a). What future quantitative social science research could look like: Confidence intervals for effect sizes. Educational Researcher, 31, 24–31.

Thompson, B. (2002b). "Statistical," "practical," and "clinical": How many kinds of significance do counselors need to consider? Journal of Counseling and Development, 80, 64–71.

Tryon, W.W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis statistical tests. Psychological Methods, 6, 371–386.

Westlake, W.J. (1976). Symmetrical confidence intervals for bioequivalence trials. Biometrics, 37, 589–594.