Journal of Modern Applied Statistical Methods

Volume 6 | Issue 1 Article 13

5-1-2007

The Effects of Heteroscedasticity on Tests of Equivalence

Jamie A. Gruman
University of Guelph

Robert A. Cribbie
York University, cribbie@yorku.ca

Chantal A. Arpin-Cribbie
York University

Journal of Modern Applied Statistical Methods
May, 2007, Vol. 6, No. 1, 133-140

1538 – 9472/07/$95.00

The Effects of Heteroscedasticity on Tests of Equivalence

Jamie A. Gruman
University of Guelph

Robert A. Cribbie Chantal A. Arpin-Cribbie
York University

Tests of equivalence, which are designed to assess the similarity of group means, are becoming more popular, yet very little is known about the statistical properties of these tests. Monte Carlo methods are used to compare the test of equivalence proposed by Schuirmann with modified tests of equivalence that incorporate a heteroscedastic error term. It was found that the latter were more accurate than the Schuirmann test in detecting equivalence when sample sizes and variances were unequal.

Key words: Null hypothesis testing, heteroscedasticity, tests of equivalence.

Introduction

Over a half century ago, Hotelling, Bartky, Deming, Friedman & Hoel (1948) wrote that "Unfortunately, too many people like to do their statistical work as they say their prayers – merely substitute in a formula found in a highly respected book written a long time ago" (p. 103). This quote, which can be found cited in *The Task Force on Statistical Inference in Psychology*'s report outlining recommendations for the effective use of statistics (Wilkinson, 1999), underscores the fact that many researchers apply statistical methods thoughtlessly, without considering the methods' appropriateness to the research questions under consideration.

Many empirical questions in behavioral research involve testing the null hypothesis of no difference between groups on a specific dependent variable. In fact, formulating research questions involving two groups as tests of this null hypothesis is almost a conditioned reflex among scholars, even though such an hypothesis is frequently irrelevant to the research question (Westlake, 1976). Testing the null hypothesis of no difference is inappropriate for studies in which the primary objective is to demonstrate that two groups are equivalent, rather than different, on a particular measure. More specifically, when the research question deals with the equivalence of groups on a dependent measure, an equivalence test is the appropriate (and necessary) statistical method to be used. The present article will highlight the importance of equivalence tests in behavioral research and use a Monte Carlo study to compare tests of equivalence when the variances of the groups are not equal.

Researchers frequently conduct studies in which assessing the equivalence of two groups is the main purpose. For example, consider an investigation of two therapies for dealing with perfectionism. One therapy is lengthy and expensive; the other short and inexpensive. The pertinent research question may be to determine whether the therapies are equivalent in terms of their effectiveness. If they are equivalent, then the shorter, less expensive method can be implemented with considerable cost and time savings. Traditional statistical procedures such as t-tests and ANOVAs are ill-suited to answering these questions because they focus, conceptually and statistically, on assessing group differences. For research

Jamie A. Gruman is Assistant Professor of Organizational Behavior. E-mail: jgruman@uoguelph.ca Robert A. Cribbie is an Associate Professor of Psychology at York University in Toronto, Ontario. He specializes in robust test statistics and multiplicity control. Chantal A. Arpin-Cribbie is a doctoral student in the Department of Psychology at York University in Toronto, Ontario. She specializes in clinical health psychology.

questions pertaining to the equivalence of conditions, researchers require a statistical technique designed specifically to test the degree to which different conditions produce similar results. Tests of equivalence serve this purpose.

When employing tests of equivalence the goal is not to show that treatment conditions are perfectly identical, but only that the differences between the treatments are too small to be considered meaningful. Consider, for example, an investigation in which an attempt is made to demonstrate that scores on a computer-based test are equivalent to those from a paper and pencil based test (e.g., Epstein, Klinkenberg, Wiley & McKinley, 2001). In this example, the researchers may not need to show that the test scores are exactly equivalent (as with the traditional null hypothesis H_o: $\mu_1 = \mu_2$, but only that differences in test scores are inconsequential (i.e., $|\mu_1 - \mu_2| < D$, where D represents an a priori critical difference for determining equivalence).

A specific example may elucidate this issue more clearly. Alkhadher, Clarke & Anderson (1998) conducted an investigation designed to assess the equivalence of the paper-and-pencil version and a computer adaptive version of three subtests from the Differential Aptitude Tests (DAT), namely numerical ability (NA), abstract reasoning (AR) and mechanical reasoning (MR). It is noteworthy that the title of their article specifically underscores the equivalence of these subtests and that in their introduction they highlight that "their equivalence must be established empirically" (p.206). However, as a means of demonstrating the equivalence of the measures, Alkhader et al. proceeded to conduct ANOVAs, which are expressly designed to detect statistically significant group differences. Based on their analyses they claimed to have demonstrated the equivalence of two of the three subtests (AR and MR). However, what Alkhader et al. in fact demonstrated was merely that scores on the NA subtest on the computer adapted version of the DAT were statistically significantly different from the paper and pencil method as traditionally defined.

The question of the equivalence of the different administration methods on subtest scores remains a mystery. As Cribbie, Gruman & Arpin-Cribbie (2004) and Rogers, Howard & Vessey (1993) note, the rejection or nonrejection of the null hypothesis of traditional tests tells us very little about the potential equivalence of the groups in question. Effectively establishing whether the computer adapted version of the DAT produced subtest scores that were equivalent to the paper and pencil version would have required the use of a statistical technique that could assess the degree to which these measures produced similar results. This can be accomplished through the use of equivalence testing, the purpose of which is to demonstrate that two (or more) conditions are functionally the same (Stegner, Bostrom & Greenfield, 1996).

This approach to statistical analysis has been popular for many years in biology, where researchers interested in the interchangeability of genetically equivalent drugs have used the technique to determine drugs' comparative bioavailability, or bioequivalence (Westlake, 1976). However, researchers outside of biology have been slow to recognize the utility of this procedure and continue to use inappropriate statistics when conducting studies that consider the similarity of alternative conditions, tests, treatments, or procedures.

One of the more commonly discussed tests of equivalence was developed by Schuirmann (1987). Schuirmann's test of equivalence has been introduced to the behavioral sciences through influential articles by Rogers et al. (1993), Seaman & Serlin (1998) and others. The first step in applying 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 by 10 points), a percentage difference (e.g., +/- 10%), a percentage of the pooled standard deviation difference, etc.

Researchers debating an appropriate value of D should consider the nature of the

research. For example, if the paper-and-pencil test discussed above was ten times more expensive to administer than the computer-based test, even a very significant difference in outcomes (e.g., +20%) might be acceptable for concluding that the tests are equivalent; Whereas if the paper-and-pencil test was only twice as expensive to administer as the computer based test a difference in outcomes of no more than 5% might be required for concluding that the tests are equivalent. For a further discussion on establishing D readers are referred to Greene, Concato & Feinstein (2000).

When using this procedure it is assumed that the two samples are randomly and independently selected from normally distributed populations with equal variances. 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:

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

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 $t_1 \leq -t_v^{\alpha}$ where:

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

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

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

$\bar{x}_1$ 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 $t_{\alpha,df}$ is the upper-tailed α-level t critical value with df = $n_1 + n_2 - 2$ degrees of freedom.

One concern with the adoption of Schuirmann's test of equivalence is the potential effects of variance heterogeneity on the standard error of the statistic. This is an important consideration given that unequal variances (heteroscedasticity) appear to be the norm, rather than the exception in behavioral research (Keselman et al., 1998; Grissom, 2000). Keselman et al. have noted that researchers often report largest to smallest variance ratios as large as four to one, and largest to smallest variance ratios as large as eight to one are not uncommon. The standard error used with the Schuirmann test is identical to that used in the two independent samples t-test, and problems with this error term have a long history, termed the Behrens-Fisher problem (see, e.g., Scheffe, 1970).

One potential option is to use the heteroscedastic solution developed by Welch (1938) and Satterthwaite (1946). This idea was originally presented by Dannenberg, Dette & Munk (1994), although the procedure has received little attention given that in biopharmaceutical equivalence trials independent groups designs (where these methods would be appropriate) are rare relative to crossover designs (see Hauschke, Steinijans & Hothorn, 1996). However, independent groups designs are extremely common in behavioural research areas such as education, psychology, and management. Combining the numerator of Schuirmann's test with the error term of Welch's (1938) heteroscedastic test may provide an equivalence test that is robust to sample size and variance heterogeneity. For the Schuirmann-Welch test of equivalence H_{01} is rejected if $t_{W1} \leq -t_{\alpha,dfw}$ and H_{02} is rejected if $t_{w2} \geq t_{\alpha,dfw}$ where :

$$t_{W1} = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - D}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}},$$

$$t_{W2} = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - (-D)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}},$$

and

$$df_w = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)}}.$$

Recently, Tryon (2001) proposed a novel approach to equivalence testing that uses inferential confidence intervals to make decisions regarding the equivalence of two groups. Specifically, with Tyron's equivalence test two groups are declared equivalent if Rg. ≤ D, where:

$$Rg = \left| \begin{array}{l} \left[\overline{X}_1 - (t_{1-\alpha})(E)\left(s_{\overline{x}_1}\right)\right] \\ -\left[\overline{X}_2 + (t_{1-\alpha})(E)\left(s_{\overline{x}_2}\right)\right] \end{array} \right|$$

and

$$E = \frac{\sqrt{s_{\overline{x}_1}^2 + s_{\overline{x}_2}^2}}{s_{\overline{x}_1} + s_{\overline{x}_2}}.$$

$s_{\bar{x}}$ represents the usual standard error of the mean (i.e., $s_x / \sqrt{n}$) and $t_{1-\alpha}$ represents the (positive) two-tailed critical t value with df = n-1. A heteroscedastic version of the Tryon test is available by substituting the original degrees of freedom (df = n-1) by the Welch-Satterthwaite df divided by two (i.e., $df_w / 2$).

Methodology

Monte Carlo Study

A simulation study was used to compare the probability of detecting equivalence by: 1) Student t; 2) Welch t; 3) Schuirmann's equivalence test; 4) Schuirmann-Welch equivalence test; 5) Tryon equivalence test; and 6) Tryon-Welch equivalence test. Several variables were manipulated in this study including: a) sample size; b) population variances; and c) population mean configuration. Total sample sizes were set at N = 20 and N = 60. Sample sizes for N = 20 were: 1) n_1=10, n_2=10; 2) n_1=8, n_2=12; and 3) n_1=5, n_2=15. Sample sizes for N = 60 were: 1) n_1=30, n_2=30; 2) n_1=25, n_2=35; and 3) n_1=20, n_2=40. Population variances were set at: 1) 1, 1; 2) .5, 1.5; 3) 1.5, .5; 4) .2, 1.8; and 5) 1.8, .2. These conditions were crossed resulting in: 1) equal n or σ^2; 2) positively paired n and σ^2 (largest n with largest σ^2, smallest n with smallest σ^2); and 3) negatively paired n and σ^2 (largest n with smallest σ^2, smallest n with largest σ^2).

Six mean configurations were evaluated in this study, including equivalent population means ($\mu_1 = \mu_2$) and five nonequivalent population means ($\mu_2 = \mu_1 + .4$, $\mu_2 = \mu_1 + .8$, $\mu_2 = \mu_1 + 1$, $\mu_2 = \mu_1 + 1.2$ and $\mu_2 = \mu_1 + 1.6$). The critical mean difference for establishing population equivalence (D) was maintained at 1 throughout all conditions. Given that D is set at 1, the equivalent mean configuration and nonequivalent configurations with $\mu_2 - \mu_1 < 1$ fall under the alternate hypothesis of the Schuirmann and Tryon tests 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), and nonnull configurations with $\mu_2 - \mu_1 > 1$ fall under the null hypothesis of the Schuirmann and Tryon tests of equivalence (i.e., the population mean difference exceeds the critical mean difference and thus the means are expected to be declared nonequivalent). For the case where $\mu_2 - \mu_1 = 1 = D$, the expected probability of declaring the two populations equivalent is α.

Five thousand simulations were conducted for each condition using a nominal significance level of $\alpha = 0.05$.

Results

The probability of declaring the two independent populations equivalent for N = 20 and N = 60 are presented in Tables 1 and 2, respectively.

$\mu 2 - \mu 1 = 1 = D$

The Schuirmann-Welch maintained rejection (i.e., rejecting H_{o1} and H_{o2}) rates at approximately α (.039-.048) for N = 20 and exactly at α for N = 60 when $\mu 2 - \mu 1 = 1$ [recall that D=1 so $E(t_{w1}) = 0$], regardless of the pattern of sample sizes and variances. However, the Schuirmann test had rejection rates ranging from .019 to .092 under positively and negatively

Table 1. Probability of declaring the two populations equivalent for N = 20 under each of the testing conditions.

Pairing	$\mu_2 - \mu_1$	t	t-w	sch	sch-w	try	try-w
Equal n or σ	0	.948	.949	.352	.340	.289	.261
	.4	.867	.873	.251	.243	.217	.195
	.8	.621	.641	.094	.091	.091	.081
	1	.462	.487	.045	.044	.048	.043
	1.2	.309	.338	.018	.018	.022	.020
	1.6	.099	.123	.002	.002	.003	.003
Positive	0	.980	.951	.212	.454	.318	.299
	.4	.932	.858	.145	.318	.228	.213
	.8	.749	.582	.047	.107	.085	.079
	1	.600	.406	.019	.048	.040	.037
	1.2	.437	.252	.007	.018	.016	.015
	1.6	.166	.065	.001	.001	.002	.001
Negative	0	.865	.947	.403	.189	.218	.161
	.4	.773	.894	.317	.146	.175	.130
	.8	.536	.738	.156	.067	.092	.067
	1	.399	.632	.092	.039	.057	.041
	1.2	.273	.516	.051	.019	.033	.023
	1.6	.096	.300	.011	.004	.008	.006

Note. t = independent samples t; t-w = Welch t; sch = Schuirmann test of equivalence; sch-w = Schuirmann-Welch test of equivalence; try = Tyron test of equivalence; try-w = Tyron-Welch test of equivlance.

paired sample sizes and variances respectively, for N = 20, and rates ranging from .028 to .084 under positively and negatively paired sample sizes and variances respectively, for N = 60. Both the Tryon and Tryon-Welch equivalence tests had reasonably accurate rejection rates for $\mu2 - \mu1 = 1$ when N = 20, although rates were consistently mildly deflated under the unequal sample size and variance conditions when N = 60 (.032 - .036).

Rejection rates for the two independent samples t and Welch t for $\mu2 - \mu1 = 1$ reflect the power of these tests for detecting a true difference in means (see Table 1).

Table 2. Probability of declaring the two populations equivalent for N = 60 under each of the testing conditions.

Pairing	$\mu_2 - \mu_1$	t	t-w	sch	sch-w	try	try-w
Equal n or σ	0	.949	.950	.965	.964	.924	.918
	.4	.676	.680	.732	.729	.657	.650
	.8	.149	.153	.186	.185	.161	.157
	1	.037	.038	.050	.050	.044	.043
	1.2	.006	.006	.009	.009	.008	.008
	1.6	.000	.000	.000	.000	.000	.000
Positive 0		.975	.949	.961	.983	.944	.943
	.4	.757	.639	.682	.781	.671	.668
	.8	.189	.107	.131	.200	.144	.143
	1	.048	.021	.028	.050	.035	.035
	1.2	.007	.002	.003	.007	.005	.005
	1.6	.000	.000	.000	.000	.000	.000
Negative	0	.900	.950	.958	.913	.758	.725
	.4	.612	.732	.747	.647	.495	.466
	.8	.141	.239	.241	.165	.116	.106
	1	.041	.087	.084	.050	.036	.032
	1.2	.008	.023	.020	.010	.007	.007
	1.6	.000	.001	.000	.000	.000	.000

Note. t = independent samples t; wel-t = Welch t; sch = Schuirmann test of equivalence; sch-w = Schuirmann-Welch test of equivalence; try = Tyron test of equivalence; try-w = Tyron-Welch test of equivlance.

A Priori Equivalence ($\mu 2 - \mu 1 < D$)

When a priori population mean differences were less than the critical mean difference (D = 1), and either the sample sizes or variances were equal, the probability of declaring the two populations equivalent was almost identical for the Schuirmann, Schuirmann-Welch, Tyron and Tyron-Welch test statistics. The rates for the equivalence tests were significantly less than the rates for the Student t and Welch t when the total sample size was small (N = 20), although the rates were larger than those for the Student t and Welch t when the total sample size was large (N = 60).

The probability of declaring the two populations equivalent was greater for the Schuirmann-Welch test than the Schuirmann test when the sample sizes and variances were positively paired, whereas the probability of declaring the two populations equivalent was greater for the Schuirmann test than the Schuirmann-Welch test when the sample sizes and variances were negatively paired. This is due to the known bias in the non-heteroscedastic standard error, which becomes inflated when sample sizes and variances are positively paired and deflated when sample sizes and variances are negatively paired.

This bias can also be seen in the results for the traditional tests as the probability of declaring the two populations equivalent (i.e., a statistically non significant effect) was greater for the Student t than the Welch when the sample sizes and variances were positively paired, and the probability of declaring the two populations equivalent was greater for the Welch than the Student t when sample sizes and variances were negatively paired. The rates for the Tryon and Tryon-Welch tests were very similar across all conditions (primarily because the original Tryon test does not use the pooled standard error like the Schuirmann test) but consistently less than those of the Schuirmann-Welch test.

A Priori Nonequivalence ($\mu 2 - \mu 1 > D$)

When a priori population mean differences were greater than the critical difference (D = 1), and either the sample sizes or variances were equal, the probability of declaring the two populations equivalent was identical (and very low) for the Schuirmann and Schuirmann-Welch test statistics under all conditions and demonstrates an excellent ability to detect differences greater than D. This is due to the fact that the numerators of t_1 and t_{w1} have an expected positive value, whereas a rejection would only occur if t_1 and t_{w1} are LESS THAN $-t_{\alpha,df}$.

One way to think of this effect would be to relate it to traditional null hypothesis testing when testing a one-tailed alternative hypothesis (i.e., H_1: $\mu_1 - \mu_2 > 0$). We expect the Type I error rates to be approximately α when $\mu_1 - \mu_2 = 0$, but when $\mu_1 - \mu_2 < 0$ (i.e., an effect in the wrong direction) the Type I error rates will approach zero. The rates for the Schuirmann and Schuirmann-Welch tests were significantly less than the rates for the Student t and Welch t when the total sample size was small (N = 20), reflecting the fact that the Student t and Welch t have less power when N = 20, although the rates were very similar for all tests when the total sample size was large (N = 60). Similar to the results for a priori equivalence, the probability of declaring the two populations equivalent was greater for the Schuirmann-Welch test than the Schuirmann test when the sample sizes and variances were positively paired, whereas the probability of declaring the two populations equivalent was greater for the Schuirmann test than the Schuirmann-Welch test when the sample sizes and variances were negatively paired. The rates for the Tryon and Tryon-Welch tests were very similar across all conditions, and were also very similar to rates for the Schuirmann-Welch procedure.

Conclusion

Behavioral researchers reliably use traditional statistical procedures such as Student's t-test when comparing groups even when the primary objective is to demonstrate that groups are equivalent, rather than different, on a particular measure. The present article highlights the need for tests of equivalence and compared alternatives to the original Schuirmann (1987) and Tryon (2001) tests of equivalence for situations in which treatment group variances are unequal. The Schuirmann-Welch test incorporated a heteroscedastic error term and error degrees of freedom, while the Tryon-Welch test incorporated heteroscedastic degrees of freedom. It was expected that these modifications would improve the performance of the test statistics when sample sizes and variances were unequal. The results of this study support the hypothesis that equivalence rates for the Schuirmann-Welch were more accurate than for the Schuirmann test, correcting for a bias in the standard error of the Schuirmann test that dates back to Fisher and Behrens in the 1930s. Equivalence rates were also more accurate (and more powerful) for the Schuirmann test than for either of the Tryon or Tryon-Welch statistics.

The results also highlight the fact that traditional test statistics such as the Student t and Welch t are not appropriate for testing research hypotheses that relate to the equivalence of two populations. The traditional null hypothesis testing procedures have an extreme bias towards declaring equivalence when sample sizes are small (i.e., a lack of power for detecting small treatment group differences), and are less likely to be able to detect equivalence relative to the Schuirmann or Schuirmann-Welch tests when sample sizes become large.

Tests of equivalence are 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 grow, the popularity of tests of equivalence will increase in behavioral fields such as education, psychology, and management. Thus, methodologists should provide recommendations for applying these tests. The findings of this study emphasize the need for robust tests of equivalence (such as the Schuirmann-Welch test investigated in this paper) for situations in which data conditions are not optimal. Empirical data rarely meet all of the underlying assumptions of test statistics (Keselman et al., 1998; Micceri, 1989; Welch; 1988), and instead one should be cognizant of assumption violations and apply appropriate test statistics that minimize the likelihood that incorrect inferences are drawn regarding the results.

References

Alkhadher, O., Clarke, D. D., & Anderson, N. (1998). Equivalance and predictive validity of paper-and-pencil and computerized adaptive formats of the differential aptitude tests. *Journal of Occupational and Organizational Psychology, 71*, 205-217.

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

Epstein, J., Klinkenberg, W. D., Wiley, D., & McKinley, L. (2001). Insuring sample equivalence across internet and paper-and-pencil assessments. *Computers in Human Behavior, 17*, 339-346.

Greene, W. L., Concato, J., & Feinstein, A. R. (2000). Claims of equivalence in medical research: Are they supported by the evidence? *Annals of Internal Medicine, 132*, 715-722.

Grissom, R. J. (2000). Heterogeneity of variance in clinical data. *Journal of Consulting and Clinical Psychology, 68*, 155-165.

Hauschke, D., Steinijans, V. W., & Hothorn, L. A. (1996). A note on Welch's approximate 't'-solution to bioequivalence assessment. *Biometrika, 83*, 236-237.

Hotelling, H., Bartky, W., Deming, W. E., Friedman, M., & Hoel, P. (1948). The teaching of statistics. *Annals of Mathematical Statistics, 19*, 95-115.

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.

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

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.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin, 2*, 110–114.

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.

Scheffe, H. (1970). Practical solutions of the Behrens-Fisher problem. *Journal of the American Statistical Association, 65*, 1501-1508.

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*(3), 193-198.

Tyron, 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.

Welch, B. L. (1938). The significance of the difference between two means when population variances are unequal. *Biometrika, 29*, 350-362.

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

Wilkinson, L., and the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*(8), 594-604.