British Journal of Mathematical and Statistical Psychology (2014), 67, 213–230

The British Psychological Society

www.wileyonlinelibrary.com

Evaluating clinical significance: Incorporating robust statistics with normative comparison tests

Katrina van Wieringen and Robert A. Cribbie*

Department of Psychology, York University, Toronto, Canada

The purpose of this study was to evaluate a modified test of equivalence for conducting normative comparisons when distribution shapes are non-normal and variances are unequal. A Monte Carlo study was used to compare the empirical Type I error rates and power of the proposed Schuirmann–Yuen test of equivalence, which utilizes trimmed means, with that of the previously recommended Schuirmann and Schuirmann–Welch tests of equivalence when the assumptions of normality and variance homogeneity are satisfied, as well as when they are not satisfied. The empirical Type I error rates of the Schuirmann–Yuen were much closer to the nominal α level than those of the Schuirmann or Schuirmann–Welch tests, and the power of the Schuirmann–Yuen was substantially greater than that of the Schuirmann or Schuirmann–Welch tests when distributions were skewed or outliers were present. The Schuirmann–Yuen test is recommended for assessing clinical significance with normative comparisons.

1. Introduction

In the field of clinical psychology, much research is dedicated to evaluating the effectiveness of various interventions. An important aspect of the effectiveness of an intervention is clinical significance, which can be operationally defined as the practical or applied importance of the intervention. This importance is normally described in terms of whether the intervention makes a real difference in the everyday life of the client or individuals who interact with the client, or whether the intervention is able to bring the client back to a state of normal functioning. Following numerous recommendations for the reporting of clinical significance (e.g., Kendall, 1997), clinical researchers are beginning to make measures of clinical significance an important part of their clinical intervention studies (e.g., Sánchez-Ortuño & Edinger, 2010; Wallach, Safir & Bar-Zvi, 2009).

Numerous methods have been put forward to evaluate the effectiveness of an intervention; however, there are a few issues with these methods. For example, one issue is that most methods do not analyse whether the treated group returns to a state of normal functioning. In other words, traditional methods of evaluating interventions may assess whether the treated group has experienced significant change, but may fail to specify how much the group has changed, what this change means, or whether the change has returned the clients in the group to a level of normal functioning. A second issue is that many methods assess clinical significance at the individual rather than the group level,

*Correspondence should be addressed to Robert A. Cribbie, Department of Psychology, York University, Toronto, Ontario M3J 1P3, Canada (e-mail: cribbie@yorku.ca).

DOI:10.1111/bmsp.12015

which can make global assessments of intervention efficacy difficult. For example, what proportion of improved clients would indicate that the intervention was effective? Although individual-level assessments are important for the clinicians to examine in order to be able to identify clients with extreme responses to the treatment, researchers reviewing intervention studies are often interested in global assessments of the efficacy of the intervention. Another important issue is that most methods ignore the fact that the distributions of scores in treated, control, and normal comparison groups are often non-normal, and that the variances are often very different across these groups.

One of the more promising recent methods for assessing clinical significance involves evaluating the equivalence of a treated group of clients and a normal comparison group on an outcome (e.g., symptoms of depression) of interest (Kendall, Marrs-Garcia, Nath & Sheldrick, 1999). The advantages of this approach are that the assessments are at the group level, and that the method directly addresses the question of whether the clients in the group have returned to a state of normal functioning. The purpose of this research was to investigate improvements to existing test statistics for evaluating the equivalence of a group of treated clients and a normal comparison group for situations in which the distributions of the groups are non-normal and/or the variances of the groups are unequal.

2. Traditional methods for determining the effectiveness of an intervention

Traditionally, statistical analyses of intervention studies involved only comparisons of pre- and post-treatment data to determine whether the treatment was responsible for the change observed. These comparisons were typically made relative to a control group. Traditional statistical significance tests – for example, Student's two-related-samples *t*-test or an ANOVA *F* test – are used to compare the pre- and post-treatment data of the experimental group or the experimental and control groups. Using traditional statistical significance tests to evaluate treatment efficacy is limited in that it does not provide information about the strength of the relationship or whether it has clinical meaning or significance – for example, whether the clients return to a state of normal functioning (Jacobson & Truax, 1991; Kraemer, Morgan, Leech, Gliner, Vaske & Harmon, 2003).

Effect size measures (e.g., Cohen's *d*) can be used to assist in interpreting practical significance, as they provide a measure of the strength of the relationship. However, Kraemer *et al.* (2003) discuss an important limitation of effect sizes that relates to their ability to act as a measure of clinical significance. The issue is that effect sizes are not interpretable in terms of how much individuals are affected by the treatment because they were not originally designed to be indices of clinical significance. However, more recent indices of effect size for clinical interventions – for example, the success rate difference (SRD) – provide more appropriate indices of an effect for randomized studies. The SRD represents the difference between the probability that a client in the treatment group has a treatment outcome preferable to a client in the control group and the probability that a client in the control group has a treatment outcome preferable to a client in the treatment group (Kraemer & Kupfer, 2006). More specifically, although effect sizes assess the strength of the association (e.g., how much more the experimental group improved than the control group), they do not tell us if the clients have been brought back to a state of normal functioning.

Popular methods of assessing clinical significance, such as Jacobson and Truax's (1991) method, measure change at the individual level by determining whether the

treated clients move outside the range of the dysfunctional population or within the range of the functional population. The first step with the Jacobson and Truax approach is to calculate a cut-off point for clinically significant change, which represents the point that the client must cross at the time of the post-treatment assessment in order to be classified as changed to a clinically significant degree. Jacobson and Truax proposed three ways to calculate the cut-off score: (1) the level of functioning following an intervention should fall outside the range of the dysfunctional population, where the range is defined as extending to two standard deviations beyond the mean of the dysfunctional population; (2) the level of functioning following an intervention should fall within the range of the functional population, where the range is defined as within two standard deviations of the mean of that population; or (3) the level of functioning following an intervention places that client closer to the mean of the functional population than it does to the mean of the dysfunctional population. The second step is to calculate a reliable change index, (RCI), to determine whether a client's change from pre-test to post-test is reliable and not due to measurement error since it is possible for post-test scores to cross the cut-off point yet not be statistically reliable.

$$\mathrm{RCI} = \frac{x_2 - x_1}{S_{\mathrm{diff}}},$$

Here x_1 represents a client's pre-test score, x_2 represents that same client's post-test score, and S_{diff} is the standard error of difference between the two test scores. S_{diff} is traditionally calculated as $s_{\mathrm{pre}}(1 - r_{\mathrm{pre}})^{1/2}$, where s_{pre} is the standard deviation of the pre-test score and r_{pre} is the internal consistency or test–retest reliability of the pre-test score (see Martinovich, Saunders & Howard, 1996, for the advantages and disadvantages of each).

There are a couple of important criticisms related to using Jacobson and Truax's method. First, it is unclear how robust the method is when the assumption that the dysfunctional and functional distributions are normal is violated, since the method assumes normal distributions. This is especially important since several previous reviews have found that distributions in Psychology are rarely normal (e.g., Micceri, 1989). Secondly, this method is designed to study clients at the individual level (i.e., each individual is investigated separately), making it difficult to make general statements about the effectiveness of an intervention. In many cases researchers will compute the proportion of individuals who recovered, improved, were unchanged or deteriorated, and others may even compare the proportions in each category; however, these tests do not directly answer the question of whether the previously clinical group is now performing equivalent to a normal comparison group.

3. Normative comparisons

Normative comparisons are a procedure for evaluating the clinical significance of therapeutic interventions. For example, imagine that an intervention has been shown to produce significantly greater improvement than a placebo, another treatment, etc., but that it is unclear how well the treated clients are functioning relative to a normal comparison group. Normative comparisons involve comparing the mean of a treated group (i.e., the clinical group following the intervention) with that of a normal comparison group in order to address the clinical significance of the research. Specifically, the goal of the research is to determine if the two groups score equivalently on the

measure or behaviour of interest. Kendall *et al.* (1999) raise two important questions related to the clinical significance of an intervention. First, is the amount of change that has occurred because of the treatment large enough to be considered meaningful? Second, are the treated individuals distinguishable from normal individuals? Normative comparison tests address these questions at the group level. This method allows one to assess the effectiveness of an intervention against a standard independent of the initially disordered individuals. However, since the statistical analyses involve demonstrating the equivalence of, rather than the difference between, the treated and normal comparison groups, specialized equivalence testing methods are required.

However, there are a couple of important cautions to raise regarding normative comparisons before discussing the methods in detail. First, the methods described in this paper are intended for interventions in which a realistic goal is to return the clients to a state of functioning that falls within the normal range. Although this is not the case for all interventions (e.g., treating behavioural issues in autistic populations), we believe that it is the case for many interventions. As discussed by an anonymous reviewer of this paper, the use of psychiatric medications (e.g., antidepressants) could also be an aspect of the intervention that helps to return the individuals' behaviour to a state of normal functioning; normative comparisons would still be useful in this setting unless the treated group was only taking the medication on a short-term basis. A second caution, also raised by an anonymous reviewer, is that because normative comparisons only utilize a measure of behaviour at a single time point, they cannot take into account the possibility of relapses in the future. Finally, it must also be emphasized that the comparison population (typically referred to as the normal comparison group) must be selected to match as closely as possible the relevant characteristics of the treated group. For example, imagine that a community sample of individuals diagnosed with major depressive disorder was treated with cognitive behavioural therapy. If the researcher wanted to conduct normative comparisons it would not be appropriate to compare the treated community sample to, for example, a normative population comprised of university students because the latter is well known to have higher than normal rates of depression (and thus the treated sample may seem equivalent to the university sample on depression, even though they might not be equivalent to a community sample on depression). Without a comparable normative group the comparisons made can be very misleading.

4. Equivalence testing

Equivalence testing is a statistical method often used in biopharmaceutical studies to determine the equivalence of two experimental drugs. Rogers, Howard and Vessey (1993) explain that equivalence testing methods can be used to evaluate many important hypotheses in psychology and related fields. Recent literature (e.g., Cribbie & Arpin-Cribbie, 2009; Gruman, Cribbie & Arpin-Cribbie, 2007; Rogers *et al.*, 1993; Seaman & Serlin, 1998) has highlighted that traditional difference-based tests are not appropriate when the goal is to determine if two groups are equivalent on an outcome variable, and encourages the use of equivalence tests for evaluating these types of hypotheses. When an investigator is trying to demonstrate the equivalence of groups with a traditional test of significance, he or she is often in the position of trying to confirm rather than reject the null hypothesis. It is widely known that with a large enough sample size even minute differences will be statistically significant (Rogers *et al.*, 1993). Further, when the sample

sizes are small, the means will almost always be declared equivalent (i.e., the null hypothesis of no difference will rarely be rejected).

The null hypothesis of equivalence tests states that the difference between the groups falls outside of the equivalence interval specified by the investigator, and the alternative hypothesis states that the difference between the groups falls inside of the specified interval (Rogers *et al.*, 1993). Schuirmann (1987) proposed an equivalence testing method in which the investigator must first define an equivalence interval and then perform two simultaneous one-sided hypothesis tests. When defining the equivalence interval, one must consider what the smallest meaningful difference would be given the nature of the research. The range must be bounded by two values: the lower value is the negative delta ($-\delta$), and the upper value is the positive delta (δ). Choosing a large δ will increase the probability of declaring the groups equivalent, but at the same time it reduces the likelihood that the differences between the groups can be considered meaningless. A smaller δ makes it harder to establish the equivalence of the two groups, but at the same time there is more confidence regarding statements of equivalence (Cribbie & Arpin-Cribbie, 2009). Let μ_1 and μ_2 represent the two population means being compared and let δ represent the smallest difference between the means that would be considered important. The statistical hypotheses are defined as:

$$H_{01} : \mu_1 - \mu_2 \geq \delta, H_{02} : \mu_1 - \mu_2 \leq -\delta;$$
$$H_{a1} : \mu_1 - \mu_2 < \delta, H_{a2} : \mu_1 - \mu_2 > -\delta.$$

In order to establish the equivalence of the means, two simultaneous one-sided tests are used to test H_{01} and H_{02}. In test 1, the goal is to reject the null hypothesis asserting that the difference between the means is greater than or equal to δ. In test 2, the goal is to reject the null hypothesis asserting that the difference is less than or equal to $-\delta$. H_{01}: $\mu_1 - \mu_2 \geq \delta$ is rejected if $t_1 \leq -t_{\alpha,df}$, where

$$t_1 = \frac{(M_1 - M_2) - \delta}{\sqrt{\frac{(n_1+n_2)[(n_1-1)S_1^2+(n_2-1)S_2^2]}{n_1 n_2 (n_1+n_2-2)}}},$$

and H_{02}: $\mu_1 - \mu_2 \leq -\delta$ is rejected if $t_2 \geq t_{\alpha,df}$, where

$$t_2 = \frac{(M_1 - M_2) - (-\delta)}{\sqrt{\frac{(n_1+n_2)[(n_1-1)S_1^2+(n_2-1)S_2^2]}{n_1 n_2 (n_1+n_2-2)}}},$$

M_1 and M_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 $n_1 + n_2 - 2$ degrees of freedom. It is important to point out that an alternative method for conducting these analyses is to determine if the $1 - 2\alpha$ confidence interval is completely contained within the equivalence interval. In other words, reject H_{01} and H_{02} if:

$$M_1 - M_2 \pm t_{2\alpha,df} \sqrt{\frac{(n_1 + n_2)[(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2]}{n_1 n_2 (n_1 + n_2 - 2)}},$$

is completely contained within ($-\delta$, δ).

For the purpose of this paper, the focus of equivalency tests is on whether an intervention is effective by comparing the post-test scores of a treated group to a normal comparison group. The use of equivalence tests to compare treated and normative populations is entitled 'normative comparisons', and these comparisons are described in the next section.

5. Kendall's normative comparison procedure

The first step in the Kendall *et al.* (1999) normative comparisons approach is to specify a range of closeness within which two groups will be considered clinically equivalent (i.e., the equivalence interval). Kendall *et al.* (1999) mention that although one standard deviation may be a guideline for published norms, the defined range can be tailored differently depending on the particular comparison. For example, the selection of δ can be guided by cut-off scores from published measures, percentages of the normal or treated group mean, or effect size estimates of differences that are not clinically significant (Kendall *et al.*, 1999). Cribbie and Arpin-Cribbie (2009) discuss how, in instances where the researcher does not have a clearly defined equivalence interval, a single δ value does not allow him or her to quantify the level of closeness established by the therapy. Cribbie and Arpin-Cribbie (2009) suggest, in cases where researchers evaluating the equivalence of treated and normative groups have little information on which to select an appropriate single interval, that multiple equivalence intervals be utilized. The following levels of δ are provided as suggestions: (1) $\delta = 0.5s_{normal}$; (2) $\delta = s_{normal}$; and (3) $\delta = 1.5s_{normal}$, where s_{normal} is the standard deviation of the normal comparison group. Again, the δ levels should be selected based on the nature of the specific study, since the difficulty associated with returning behaviour to normal functioning varies greatly from behaviour to behaviour and study to study (Cribbie & Arpin-Cribbie, 2009). Cribbie and Arpin-Cribbie also discuss the importance of conducting a preliminary test that assesses whether the clinical group at pre-test is different from the normal comparison group (using a traditional difference-based test), since, if the groups do not differ before the intervention, the value of demonstrating that they are equivalent following the intervention is reduced substantially.

The next step is to conduct the equivalence test to determine if the treated and normal comparison groups can be considered equivalent. In equivalence testing two popular approaches are the original Schuirmann (1987) two one-sided test (TOST) procedure (presented above) and the Schuirmann–Welch test (Dannenberg, Dette & Munk, 1994; Gruman *et al.*, 2007). The original Schuirmann procedure assumes that the treated and normal comparison population variances are equal, whereas the Schuirmann–Welch does not require the population variances to be equal. Although Kendall *et al.* (1999), and others (e.g., Golinski & Cribbie, 2009; Keselman *et al.*, 1998) have indicated that population variances are rarely equal in psychological studies, Kendall *et al.* present the homoscedastic TOST procedure proposed by Schuirmann (1987) as the method for conducting normative comparisons.

Dannenberg *et al.* (1994) proposed a modification to the original Schuirmann test of equivalence that incorporated the heteroscedastic standard error and degrees of freedom due to Welch (1938) and Satterthwaite (1946). This modification was done because the original Schuirmann test of equivalence utilizes the same standard error and degrees of freedom as the independent-samples *t*-test, and thus the sample size and variance inequality issues that affect the independent-samples *t*-test also affect Schuirmann's

equivalence test (Cribbie & Arpin-Cribbie, 2009). Empirical Type I error rates for Schuirmann's test of equivalence have been found to deviate substantially from the nominal α level when sample sizes and variances are unequal (Gruman *et al.*, 2007). For the Schuirmann–Welch test of equivalence, H_{01}: $\mu_1 - \mu_2 \geq \delta$ is rejected if $t_{W1} \leq -t_{\alpha,df_W}$ and H_{02}: $\mu_1 - \mu_2 \leq -\delta$ is rejected if $t_{W2} \geq t_{\alpha,df_W}$, where

$$t_{W1} = \frac{(M_1 - M_2) - \delta}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}},$$

$$t_{W2} = \frac{(M_1 - M_2) - (-\delta)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}},$$

and

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

6. Example of normative comparisons

Manzoni, Cribbie, Villa, Arpin-Cribbie, Gondoni and Castelnuovo (2010) used an observational pre–post study to examine the effectiveness of a 4-week cardiac rehabilitation programme for improving obese patients' psychological well-being. The patients completed the Psychological General Well-Being Inventory (PGWBI) at admission and at discharge. The researchers used the normative comparisons procedure to determine the clinical significance of the 4-week residential programme. Manzoni *et al.* hypothesized that obese patients with heart disease would have lower scores than normal on a measure of quality of life at baseline and that a cardiac rehabilitation intervention would be effective for improving mental health to normal levels. Equivalence testing with normative comparisons was used to evaluate whether the 4-week rehabilitation programme was effective at improving impaired PGWBI dimensions to normal levels. Results showed that patients' mean scores on the PGWBI scales that were impaired at baseline in comparison to norms (global score, self-control and vitality) were equivalent to normal means at discharge. Even at the subgroup level, the PGWBI scales that were impaired at baseline significantly improved to normal levels at the end of the intervention.

7. Improving normative comparisons

An important issue when evaluating the equivalence of treated and normal comparison groups is that the distributions of the groups often are not normal and/or contain outlying cases. For example, with treated clinical populations often a large proportion of the group improves but a small subset of the group either does not improve or even deteriorates. This can result in a negatively skewed distribution with outlying cases. Further, normal comparison group scores on popular psychological scales measuring depression, anxiety, etc. often produce highly positively skewed distributions with many outlying cases in the upper tail. These are only a couple of examples of distributions that are often skewed in clinical research.

Even though the Schuirmann–Welch test is robust to violations of the variance homogeneity assumption, it is unclear how robust this test will be when the distributions are skewed or contain outliers. For example, both Student's t (which assumes equal variances) and Welch's t (which does not assume equal variances) have poor control over Type I error rates when distributions are non-normal and variances are unequal (Keselman, Othman, Wilcox & Fradette, 2004). A popular approach for dealing with non-normal distributions and/or outliers has been to remove the atypical values by trimming the data. Wilcox (1997), Keselman *et al.* (2004), and others indicate that rates of Type I error and power to detect effects are much less affected when trimmed means are substituted for the usual group means. Past studies have recommended setting the amount of trimming at 20% (e.g., Keselman, Wilcox, Lix, Algina & Fradette, 2007; Wilcox & Keselman, 2001; Wilcox, 1997), so that the analyses are computed after removing the most extreme 20% of the cases from each tail (see details below). Yuen (1974) proposed a trimmed two-sample t-test, based on the standard error and degrees of freedom of the Welch (1938) test, that was found to be more powerful and had more accurate empirical Type I error rates than the original Student's t when the distributions were non-normal.

In order to improve the normative comparisons approach for situations in which the distributions are non-normal or contain outlying cases, the Schuirmann–Welch test of equivalence was modified by replacing the original means and variances with the trimmed means and Winsorized variances (entitled the Schuirmann–Yuen procedure). H_{01} is rejected if $t_{Y1} \leq -t_{\alpha,df_Y}$ and H_{02} is rejected if $t_{Y2} \geq t_{\alpha,df_Y}$, where

$$t_{Y1} = \frac{M_{t1} - M_{t2} - \delta}{\sqrt{d_1 + d_2}},$$

$$t_{Y2} = \frac{M_{t1} - M_{t2} - (-\delta)}{\sqrt{d_1 + d_2}},$$

and

$$df_y = \frac{(d_1 + d_2)^2}{\frac{d_1^2}{(h_1 - 1)} + \frac{d_2^2}{(h_2 - 1)}}.$$

h_1 and h_2 represent the sample sizes after trimming and M_{t1} and M_{t2} are the population trimmed means. In order to obtain the trimmed means, let $Y_{(1)j} \leq Y_{(2)j} \leq \ldots \leq Y_{(nj)j}$ and let $g_j = [\gamma n_j]$ indicate that γn_j is rounded down to the nearest integer; γ represents the proportion of observations that are to be trimmed in each tail of the distribution. The effective sample size for the jth group becomes $h_j = n_j - 2g_j$. Therefore the sample trimmed mean is

$$M_{tj} = \frac{1}{h_j} \sum_{i=g_j+1}^{n_j - g_j} Y_{ij}.$$

Further, d is defined as

$$d_j = \frac{(n_j - 1)\hat{\sigma}_{wj}^2}{h_j(h_j - 1)},$$

where $\hat{\sigma}_{wj}^2$ is the Winsorized variance. The sample Winsorized variance, which is required to get a theoretically valid estimate of the standard error of trimmed means, is then given by

$$\hat{\sigma}^2_{wj} = \frac{1}{n_j - 1}\sum_{i=1}^{n_j}\left(X_{ij} - \hat{\mu}_{wj}\right)^2.$$

The sample Winsorized mean, necessary to compute the Winsorized variance, is computed as

$$\hat{\mu}_{wj} = \frac{1}{n_j}\sum_{i=1}^{n_j} X_{ij},$$

where

$$\begin{aligned} X_{ij} &= Y_{(g_j+1)j} \text{ if } Y_{ij} \le Y_{(g_j+1)j} \\ &= Y_{ij} \text{ if } Y_{(g_j+1)j} < Y_{ij} < Y_{(n_j-g_j)j} \\ &= Y_{(n_j-g_j)j} \text{ if } Y_{ij} \ge Y_{(n_j-g_j)j} \end{aligned}.$$

In other words, the Winsorized mean is computed by replacing the trimmed observations with the most extreme untrimmed value from the respective tail.

Therefore, the purpose of this project is to determine if the Schuirmann–Yuen test, proposed in this paper, is more accurate, under realistic data conditions, than the previously proposed Schuirmann and Schuirmann–Welch tests for identifying when the treated and normal comparison groups are equivalent. With regard to the data conditions investigated, in addition to unequal variances and distributions that are non-normal, but equal in form, this study will extend the robustness literature by investigating the properties of the tests when the underlying population distributions differ (e.g., one skewed, one normal). This is an important aspect of this research because the distribution shapes of clinical and normal comparison groups often differ substantially.

8. Method

A Monte Carlo simulation study was used to compare the probability of detecting equivalence with Schuirmann's equivalence test, the Schuirmann–Welch equivalence test, and the Schuirmann–Yuen equivalence test. Several variables were manipulated in this study, including sample size, population standard deviations, and distribution shape. The conditions utilized in this study are summarized in Table 1. The sample sizes chosen were n = 40, 100, and 400, and both equal and unequal sample sizes were used. There were five standard deviation conditions: in addition to equal standard deviations across groups, we investigated two levels of unequal standard deviations that were either positively or negatively paired with the unequal sample sizes. Positively paired sample sizes and standard deviations means that the larger sample size is paired with the larger standard deviation, and the smaller sample size is paired with the smaller standard deviation. Negatively paired sample sizes and standard deviations means that the larger sample size is paired with the smaller standard deviation, and the smaller sample size is paired with the larger standard deviation.

A total of nine combinations of distribution shapes were used, based on a normal distribution, a skewed distribution, and distributions containing outliers in one or both tails. Positively and negatively skewed distributions were generated using the *g*- and *h*-distribution (Hoaglin, 1985), where g represents the skewness parameter and h

Table 1. Conditions for the Monte Carlo simulation study

n_1, n_2	σ_1, σ_2	γ_1, γ_2
20, 20	1, 1	Normal, normal
15, 25	0.7, 1.3	Normal, +skew
25, 15	0.5, 1.5	Normal, outlier (+/−)
50, 50	1.3, .7	Normal, outlier (+)
25, 75	0.7, 1.3	+skew, +skew
75, 25		+skew, outlier (+/−)
200, 200		−skew, +skew
150, 250		Outlier (+/−), outlier (+/−)
250, 150		Outlier (+), outlier (+)

Note. γ = distribution shape; +skew = positively skewed; −skew = negatively skewed; outlier (+) = outliers in the upper tail only; outlier (+/−) = outliers in the upper and lower tails.

represents the kurtosis parameter. More specifically, a highly skewed distribution was generated using $g = 1$ and $h = 0$. To generate data from a *g*- and *h*-distribution, standard unit normal variables (Z_{ij}) were converted to the random variable

$$X_{ij} = \frac{e^{gZ_{ij}} - 1}{g} e^{\frac{hZ_{ij}^2}{2}},$$

according to the values of g and h selected for investigation. A negatively skewed distribution is created by reflecting the positively skewed distribution before modifying the mean of the distribution. To obtain a distribution with standard deviation σ_j, each X_{ij} was multiplied by a value of σ_j. It is important to note that this does not affect the value of the null hypothesis when $g = 0$ (see Wilcox, 1994). However, when $g > 0$, the population mean for a *g*- and *h*-variable is

$$\mu_{gh} = \frac{1}{g\sqrt{(1-h)}} \left(e^{\frac{g^2}{2(h-1)}} - 1 \right).$$

Thus, for those conditions where $g > 0$, μ_{gh} was first subtracted from X_{ij} before multiplying by σ_j. When working with trimmed means, the population trimmed mean for the *j*th group was subtracted from the variate before multiplying by σ_j.

A distribution containing outliers was generated by adding outliers to one or both of the tails of a normal distribution. Following the method described by Zimmerman (1994), the outliers were drawn from a normal distribution with a standard deviation five times as large as that of the original distribution. For both the single-tail and double-tail outlier conditions, 10% of the total cases were drawn from the outlier distribution. As with the *g*- and *h*-distributed data, the distributions that contained outliers in only one tail were adjusted so that the population means (or trimmed means) were equal to 0. The different distribution shape conditions were crossed with the sample size and standard deviation conditions, resulting in 405 unique conditions that were each evaluated when the null hypothesis was true (Type I error) and when the null hypothesis was false (power).

Ten thousand simulations were conducted for each condition using a nominal significance level of $\alpha = 0.05$ and an equivalence interval of $(-1, 1)$. For the Type I error

conditions the means were set to 0 and 1 (i.e., the difference between the means is at the bounds of the equivalence interval). For the power conditions the means were set at 0 and 0.66 (i.e., the difference between the means falls within the equivalence interval).

9. Results

Empirical Type I error rates within the interval of 0.025–0.075 (i.e., $\alpha \pm 0.5\alpha$) are considered to be robust. Figures 1 and 2 contain empirical Type I error rates for the conditions involving identically shaped distributions and non-identically shaped distributions, respectively. Tables 2 and 3 contain the empirical power values for the

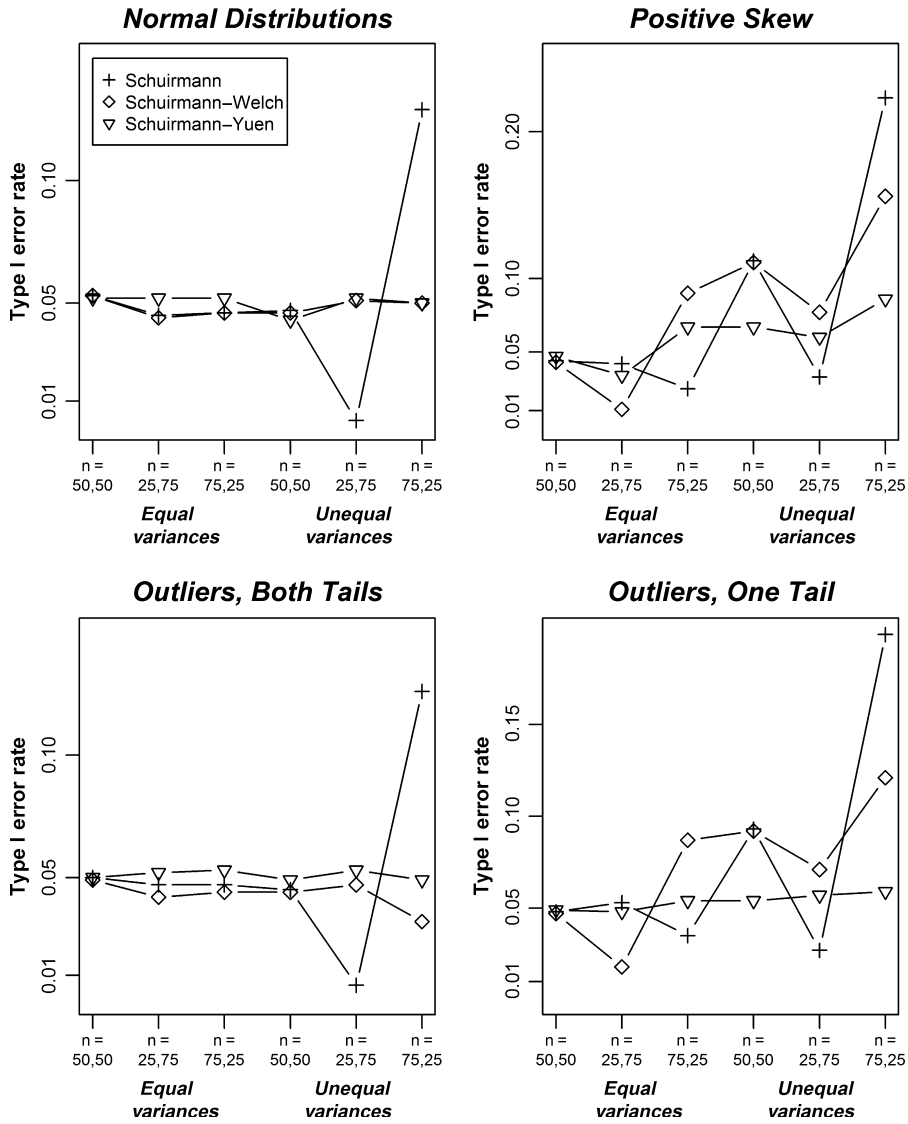

Figure 1. Type I error rates with identical distribution shapes

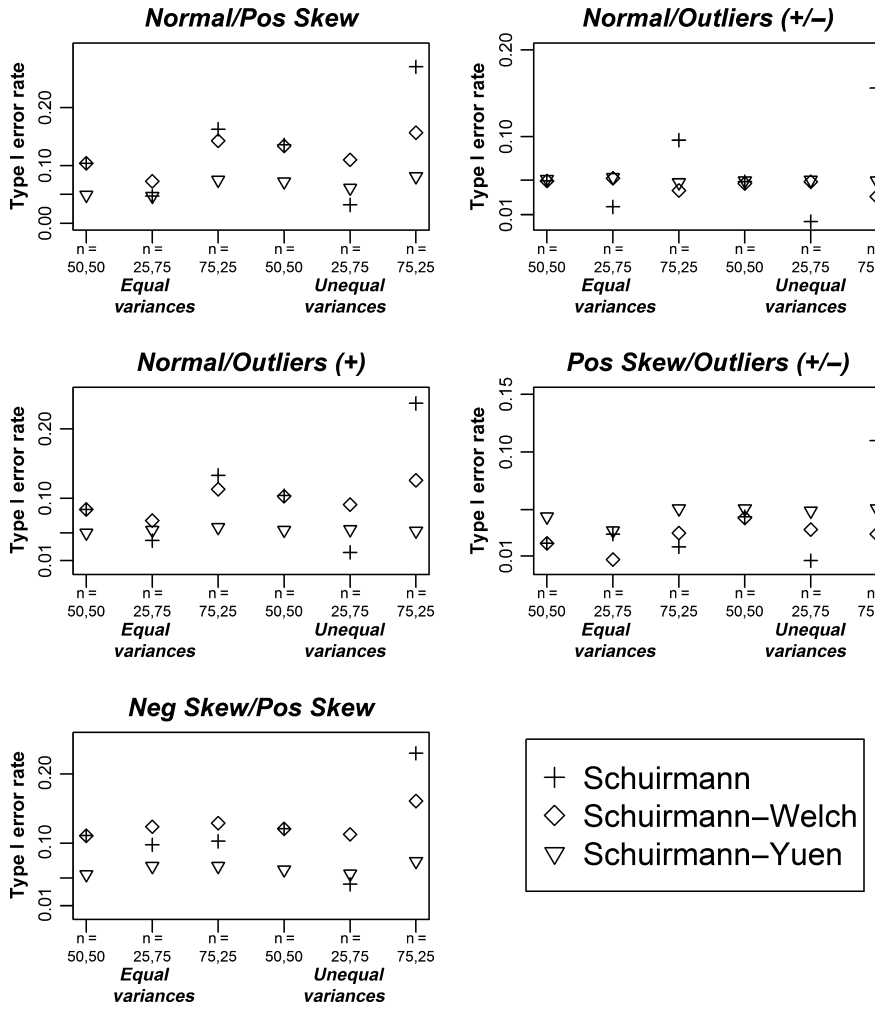

Figure 2. Type I error rates with different distribution shapes.
Note: Pos Skew = positively skewed; Neg Skew = negatively skewed; Outliers (+/−) = outliers in the upper and lower tails; Outliers (+) = outliers in the upper tail only

conditions involving identically shaped distributions and non-identically shaped distributions, respectively. Note that when the distribution shapes are the same, it is not necessary to reverse the order of the population standard deviations because that would just replicate previous conditions, but when the distribution shapes are different, reversing the order of the population standard deviations results in unique conditions.

Since the pattern of results was very similar across the moderately and extremely unequal standard deviation conditions, we report only the results for the extremely unequal standard deviation conditions, since this is the condition that has the greatest effect on the test statistics. Further, since the pattern of results for $N = 40$, $N = 100$, and $N = 400$ was very similar, only the Type I error rates for $N = 100$ and power rates for $N = 40$ and 100 are displayed. However, we do discuss the results below and full tables of results are available by contacting the authors.

Table 2. Power rates with identical distribution shapes

n_1, n_2	Sch	Sch-W	Sch-Y
	$\sigma_1 = 1, \sigma_1 = 0.5,$ $\sigma_2 = 1\ \sigma_2 = 1.5$	$\sigma_1 = 1, \sigma_1 = 0.5,$ $\sigma_2 = 1\ \sigma_2 = 1.5$	$\sigma_1 = 1, \sigma_1 = 0.5,$ $\sigma_2 = 1\ \sigma_2 = 1.5$
$\gamma_1 = \gamma_2$ = normal			
20, 20	0.282 0.243	0.281 0.238	0.254 0.217
15, 25	0.274 0.170	0.270 0.277	0.247 0.248
25, 15	0.265 0.298	0.261 0.195	0.255 0.172
50, 50	0.511 0.238	0.510 0.445	0.467 0.387
25, 75	0.438 0.277	0.436 0.513	0.409 0.446
75, 25	0.414 0.195	0.409 0.291	0.362 0.274
$\gamma_1 = \gamma_2$ = +skew			
20, 20	0.089 0.212	0.085 0.208	0.207 0.250
15, 25	0.077 0.141	0.042 0.182	0.123 0.247
25, 15	0.098 0.222	0.133 0.170	0.225 0.225
50, 50	0.203 0.302	0.203 0.298	0.394 0.357
25, 75	0.140 0.110	0.069 0.262	0.275 0.399
75, 25	0.157 0.422	0.283 0.314	0.355 0.294
$\gamma_1 = \gamma_2$ = outlier (+/−)			
20, 20	0.089 0.096	0.087 0.093	0.223 0.183
15, 25	0.087 0.043	0.093 0.103	0.197 0.218
25, 15	0.087 0.109	0.087 0.058	0.194 0.117
50, 50	0.251 0.221	0.250 0.218	0.402 0.319
25, 75	0.174 0.048	0.203 0.260	0.332 0.385
75, 25	0.183 0.297	0.194 0.120	0.329 0.229
$\gamma_1 = \gamma_2$ = outlier (+)			
20, 20	0.100 0.178	0.099 0.176	0.219 0.184
15, 25	0.091 0.121	0.067 0.192	0.194 0.221
25, 15	0.079 0.238	0.121 0.159	0.196 0.149
50, 50	0.243 0.291	0.243 0.290	0.398 0.355
25, 75	0.188 0.129	0.129 0.264	0.309 0.403
75, 25	0.215 0.394	0.295 0.246	0.337 0.235

Note. Sch = Schuirmann; Sch-W = Schuirmann–Welch; Sch-Y = Schuirmann–Yuen; γ = distribution shape; +skew = positively skewed; outlier (+/−) = outliers in the upper and lower tails outlier; (+) = outliers in the upper tail only; grey colour = Type I error rate not controlled within $\alpha \pm 0.5\alpha$ (0.025–0.075).

9.1. Type I error

9.1.1. Identical distribution shapes

The results showed that when both the treated and the normative comparison groups had a normal distribution, Type I error rates were maintained within the interval of 0.025–0.075 for all three equivalence tests. The exception was when standard deviations and sample sizes were unequal, in which case the Type I error rates for the original Schuirmann procedure were inflated in the negative pairing situation (e.g., 0.129 for $N = 100$) and deflated in the positive pairing situation (e.g., 0.002 for $N = 100$).

When the distribution shapes were both skewed or contained outliers, Type I error rates were maintained within the conservative bounds most effectively with the Schuirmann–Yuen test. Although the Schuirmann–Yuen had some Type I error rates that fell outside the acceptable bounds when both distributions were extremely skewed

Table 3. Power rates with different distribution shapes

n_1, n_2	Sch	Sch-W	Sch-Y
	$\sigma_1 = 1, \sigma_1 = 0.5,$ $\sigma_2 = 1\ \sigma_2 = 1.5$	$\sigma_1 = 1, \sigma_1 = 0.5,$ $\sigma_2 = 1\ \sigma_2 = 1.5$	$\sigma_1 = 1, \sigma_1 = 0.5,$ $\sigma_2 = 1\ \sigma_2 = 1.5$
γ_1 = normal; γ_2 = +skew			
20, 20	0.296 0.271	0.296 0.265	0.249 0.240
15, 25	0.238 0.188	0.263 0.283	0.210 0.238
25, 15	0.318 0.362	0.294 0.245	0.244 0.224
50, 50	0.376 0.322	0.376 0.321	0.418 0.355
25, 75	0.263 0.138	0.339 0.352	0.365 0.422
75, 25	0.443 0.450	0.352 0.293	0.349 0.259
γ_1 = normal; γ_2 = outlier (+/−)			
20, 20	0.192 0.116	0.191 0.109	0.243 0.166
15, 25	0.124 0.054	0.172 0.119	0.223 0.210
25, 15	0.191 0.134	0.157 0.060	0.192 0.140
50, 50	0.307 0.239	0.307 0.233	0.395 0.341
25, 75	0.182 0.046	0.300 0.273	0.363 0.428
75, 25	0.369 0.358	0.233 0.134	0.323 0.222
γ_1 = normal; γ_2 = outlier (+)			
20, 20	0.287 0.262	0.286 0.256	0.243 0.193
15, 25	0.223 0.148	0.242 0.244	0.216 0.174
25, 15	0.306 0.322	0.285 0.211	0.225 0.180
50, 50	0.373 0.306	0.371 0.302	0.444 0.331
25, 75	0.232 0.123	0.328 0.311	0.351 0.403
75, 25	0.422 0.450	0.336 0.283	0.340 0.218
γ_1 = +skew; γ_2 = outlier (+/−)			
20, 20	0.027 0.061	0.026 0.060	0.185 0.123
15, 25	0.033 0.032	0.024 0.056	0.206 0.115
25, 15	0.028 0.086	0.042 0.048	0.102 0.121
50, 50	0.145 0.201	0.145 0.198	0.333 0.368
25, 75	0.128 0.039	0.078 0.170	0.387 0.314
75, 25	0.103 0.288	0.170 0.120	0.211 0.298
γ_1 = −skew; γ_2 = +skew			
20, 20	0.245 0.230	0.244 0.225	0.224 0.224
15, 25	0.242 0.153	0.249 0.240	0.228 0.232
25, 15	0.218 0.271	0.237 0.187	0.221 0.234
50, 50	0.310 0.303	0.310 0.302	0.375 0.345
25, 75	0.299 0.145	0.299 0.304	0.321 0.373
75, 25	0.299 0.407	0.299 0.282	0.325 0.266

Note. Sch = Schuirmann; Sch-W = Schuirmann–Welch; Sch-Y = Schuirmann–Yuen; γ = distribution shape; +skew = positively skewed; outlier (+/−) = outliers in the upper and lower tails outlier; (+) = outliers in the upper tail only; grey colour = Type I error rate not controlled within $\alpha \pm 0.5\alpha$ (0.025–0.075).

and variances were unequal, the original Schuirmann and Schuirmann–Welch procedures failed in many more conditions than the Schuirmann–Yuen, and when they did fail the rates were often very disparate from the nominal level. For example, with $n_1 = 25$ and $n_2 = 15$, and unequal variances, the Type I error rate for the Schuirmann test was 0.177 and for the Schuirmann–Welch was 0.145. Further, in the large sample size condition

($N = 400$), rates for the Schuirmann–Yuen were always maintained within the acceptable Type I error bounds, whereas the pattern for the Schuirmann and Schuirmann–Welch procedures was similar to that for the small sample sizes, with rates regularly falling below or above the acceptable levels.

9.1.2. Different distribution shapes

When the two groups had different distribution shapes, the empirical Type I error control of the Schuirmann–Yuen procedure was substantially better than that of the original Schuirmann or the Schuirmann–Welch. More specifically, although the Schuirmann–Yuen procedure had Type I error rates that fell outside the acceptable bounds in one of the 30 conditions reported in Table 3 (normal and skewed distribution, and unequal variances), the Schuirmann and Schuirmann–Welch procedures each had Type I error rates that fell outside the acceptable bounds in more than half of the conditions. Further, although the rates across all conditions ranged from 0.010 to 0.091 for the Schuirmann–Yuen, the rates for the Schuirmann ranged from 0.000 to 0.271, and the rates for the Schuirmann–Welch ranged from 0.000 to 0.180. When sample sizes were large, as with identical distribution shapes, the empirical Type I error rates were well controlled by the Schuirmann–Yuen in all conditions, but commonly fell outside the acceptable bounds with the Schuirmann and Schuirmann–Welch.

9.2. Power

The power values shown in grey in Tables 2 and 3 represent conditions in which the Type I error rates were not within the interval of robustness, and thus the power rates are biased (and should not be interpreted).

9.2.1. Identical distribution shapes

When distribution shapes were normal, standard deviations were equal, and sample sizes were equal, the power rates were slightly higher for the Schuirmann and Schuirmann–Welch tests than the Schuirmann–Yuen. This result is due to the fact that the Schuirmann–Yuen employs trimmed means and thus the test statistics were based on a smaller sample size. When the distribution shapes were non-normal but identical, the power rates for the Schuirmann–Yuen were generally much higher than those for the Schuirmann or Schuirmann–Welch (and it also important to notice that many of the power rates for the Schuirmann and Schuirmann–Welch are not interpretable because the Type I error rates fell outside the acceptable range). For example, with equal sample sizes and variances and $N = 100$, power for both the Schuirmann and Schuirmann–Welch procedures was 0.203, whereas the power of the Schuirmann–Yuen was 0.394.

When sample sizes were large ($N = 400$), the differences in power with non-normal distributions are even more exaggerated, with power rates for the Schuirmann and Schuirmann–Welch ranging from about 0.3 to 0.6 and rates for Schuirmann–Yuen ranging from about 0.7 to 0.9. For example, when both distribution were extremely positively skewed, sample sizes were 150 and 250, and variances were unequal, the power rates for the Schuirmann and Schuirmann–Welch were 0.385 and 0.481, respectively, and the power for the Schuirmann–Yuen was 0.812.

9.2.2. Different distribution shapes

When the distributions had different shapes, the same pattern of results emerged; however, because of the poor Type I error control of the Schuirmann and Schuirmann–Welch procedures under these conditions, there were very few opportunities for unbiased comparisons with the Schuirmann–Yuen. However, for the conditions in which comparisons were possible, the Schuirmann–Yuen was always more powerful. For example, when one distribution was skewed and one distribution contained outliers in both tails, $n_1 = 50$ and $n_2 = 50$, and variances were unequal, power for the Schuirmann was 0.201, for the Schuirmann–Welch was 0.198 and for the Schuirmann–Yuen was 0.368.

When sample sizes were large ($N = 400$), the same pattern emerged as for the smaller sample sizes, with the power rates for the Schuirmann–Yuen often substantially higher than the rates for the other procedures. For example, when one distribution was skewed and one distribution contained outliers in both tails, $n_1 = 150$ and $n_2 = 250$, and variances were equal, power for the Schuirmann was 0.475, for the Schuirmann–Welch was 0.455 and for the Schuirmann–Yuen was 0.876.

10. Discussion

When a client enters therapy for a given issue, such as depression, the end goal in many cases is for the client's behaviour on that issue to return to a state of normal functioning. Normative comparison tests allow researchers to determine if the treated group has gone through a change back to normal functioning by comparing the treated group to a normative group on the measure of interest. Kendall *et al.* (1999) employed Schuirmann's (1987) two one-sided tests of equivalence to determine if a treated and a normative comparison group are equivalent on a particular measure. Equivalence tests have been shown to be more effective than traditional hypothesis tests (e.g., Student's *t*-test) for determining equivalence, but traditional equivalence tests are not robust when the assumptions of normality and/or homogeneity of variance are violated. Dannenberg *et al.* (1994) incorporated a heteroscedastic test statistic (Schuirmann–Welch test) that was found to be robust when the variance homogeneity assumption was violated. However, there remained concern over the robustness of this test to distribution non-normality or the presence of outliers.

The purpose of this paper was to compare the robustness of the original Schuirmann (1987) and Schuirmann–Welch (Dannenberg *et al.*, 1994) tests of equivalence when the standard deviations were unequal and the population distributions were non-normal, to that of the trimmed means based Schuirmann–Yuen procedure described in this paper. It was expected that the Schuirmann–Yuen test, which incorporated a heteroscedastic test statistic with trimmed means, would provide better Type I error control and power under conditions of variance inequality and distribution non-normality.

The results of this study indicate that the Schuirmann–Yuen test provides much better control of empirical Type I error rates and higher power than the Schuirmann or Schuirmann–Welch tests. Although the Schuirmann–Yuen did not always provide acceptable Type I error control with small sample sizes, the rates with larger sample sizes were well controlled and the rates across all conditions were much better controlled than those of the original Schuirmann or Schuirmann–Welch procedures. Further, the power of the Schuirmann–Yuen procedure when at least one of the distributions was non-normal was regularly higher than that for the Schuirmann or Schuirmann–Welch procedures, an

advantage that is of utmost importance to researchers conducting normative comparisons. One caution raised by an anonymous reviewer is that, although the Schuirmann–Yuen compares trimmed means and is thus an effective method for comparing the typical individual in one group with the typical individual in another group, it possible that the underlying distribution shapes could be very different. Thus, researchers are encouraged to explore the distribution shapes of their variables and be cautious of comparing central tendencies when distribution shapes differ. In order to improve the accessibility of the methods described in this paper, an R (R Development Core Team, 2010) function for conducting the normative comparison tests described in this paper is available at http://www.psych.yorku.ca/cribbie/norm_comparisons_rprogram_web.txt. R is an open-source statistical software program that is available at http://www.r-project.org.

The need for robust statistical tests has increased over the years with the realization that traditional statistical tests are unreliable when the assumptions of normality and homogeneity of variance are violated. This is especially important for researchers in the field of clinical psychology, where these assumptions are rarely satisfied. To conclude, the proposed Schuirmann–Yuen equivalence test is recommended for conducting normative comparisons because it provides better Type I error control and greater power than the original Schuirmann or Schuirmann–Welch equivalence tests.

References

Cribbie, R. A., & Arpin-Cribbie, C. A. (2009). Evaluating clinical significance through equivalence testing: Extending the normative comparisons approach. *Psychotherapy Research*, *19*, 677–686.

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

Golinski, C., & Cribbie, R. A. (2009). The expanding role of quantitative methodologists in advancing psychology. *Canadian Psychology*, *50*, 83–90.

Gruman, J. A., Cribbie, R. A., & Arpin-Cribbie, C. A. (2007). The effects of heteroscedasticity on tests of equivalence. *Journal of Modern Applied Statistical Methods*, *6*, 132–140.

Hoaglin, D. C. (1985). Summarizing shape numerically: The *g*- and *h*-distributions. In D. Hoaglin, F. Mosteller & J. Tukey (Eds.), *Exploring data tables, trends, and shapes* (pp. 461–513). New York: Wiley.

Jacobson, N. S., & Truax, P. (1991). Clinical significance: a statistical approach to defining change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, *59*, 12–19.

Kendall, P. C. (1997). Editorial. *Journal of Consulting and Clinical Psychology*, *15*, 3–5.

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*, *67*, 285–299.

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R., Donahue, B., . . . 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.

Keselman, H. J., Othman, A. R., Wilcox, R. R., & Fradette, K. (2004). The new and improved two-sample *t* test. *Psychological Science*, *15*, 47–51.

Keselman, H. J., Wilcox, R. R., Lix, L. M., Algina, J., & Fradette, K. (2007). Adaptive robust estimation and testing. *British Journal of Mathematical and Statistical Psychology*, *60*, 267–293.

Kraemer, H. C., & Kupfer, D. J. (2006). Size of treatment effects and their importance to clinical research and practice. *Biological Psychiatry*, *59*, 990–996.

Kraemer, H. C., Morgan, G. A., Leech, N. L., Gliner, J. A., Vaske, J. J., & Harmon, R. J. (2003). Measures of clinical significance. *Journal of the American Academy of Child & Adolescent Psychiatry*, *42* (12), 1524–1529.

Manzoni, G. M., Cribbie, R. A., Villa, V., Arpin-Cribbie, C. H., Gondoni, L., & Castelnuovo, G. (2010). Psychological well-being in obese inpatients with ischemic heart disease at entry and at discharge from a four-week cardiac rehabilitation program. *Frontiers in Psychology*, *1*, 1–7.

Martinovich, Z., Saunders, S., & Howard, K. (1996). Some comments on 'assessing clinical significance'. *Psychotherapy Research*, *6*, 124–132.

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

R Development Core Team (2010). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

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.

Sánchez-Ortuño, M. M., & Edinger, J. D. (2010). A penny for your thoughts: Patterns of sleep-related beliefs, insomnia symptoms and treatment outcome. *Behaviour Research and Therapy*, *48*, 125–133.

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-sided tests procedure and the power approach for assigning 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.

Wallach, H. S., Safir, M., & Bar-Zvi, M. (2009). Virtual reality cognitive behavior therapy for public speaking anxiety: A randomized clinical trial. *Behavior Modification*, *33*, 314–333.

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

Wilcox, R. R. (1994). A one-way random effects model for trimmed means. *Psychometrika*, *59*, 289–306.

Wilcox, R. R. (1997). *Introduction to robust estimation and hypothesis testing*. San Diego, CA: Academic Press.

Wilcox, R. R., & Keselman, H. J. (2001). Using trimmed means to compare *K* measures corresponding to two independent groups. *Multivariate Behavioral Research*, *36*, 421–444.

Yuen, K. K. (1974). The two-sample trimmed *t* for unequal population variances. *Biometrika*, *61*(1), 165–170.

Zimmerman, D. W. (1994). A note on the influence of outliers on parametric and nonparametric tests. *Journal of General Psychology*, *121*, 391–401.

Received 10 April 2012; revised version received 21 March 2013