Evaluating Clinical Significance Through Equivalence Testing:

Extending the Normative Comparisons Approach

Robert A. Cribbie & Chantal A. Arpin-Cribbie

York University

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

Abstract

The field of psychology, as with many other disciplines, has been increasingly interested in being able to measure the effectiveness of behavioral interventions. This trend has led to a number of different approaches for measuring clinical significance, each addressing a slightly different aspect of the clinical outcome. Recently, clinical psychologists (and clients) have supported the contention that one of the most important therapeutic questions is whether clients are functioning equivalently to normal controls following an intervention. To address this question, Kendall, Marrs-Garcia, Nath and Sheldrick (1999) presented an approach to measuring clinical significance that utilizes tests of equivalence. The present paper clarifies the nature of the hypotheses being conducted in measuring clinical significance with tests of equivalence, and extends the approach by incorporating recent advances in equivalence testing. A revised approach for evaluating clinical significance via equivalence testing is proposed, and an empirical example demonstrating this approach is provided.

Evaluating Clinical Significance Through Equivalence Testing:
Extending the Normative Comparisons Approach

Kendall and Grove stated that "convincing demonstrations of therapeutic efficacy must provide evidence, where possible, that once troubled and disordered clients are now, after treatment, not distinguishable from a meaningful and representative nondisturbed reference group" (1988, p. 148). Further, Jacobsen and Revenstorf (1988) claim that clients “expect to be as normal as their functioning counterparts by the time therapy has ended” (p. 134). Kendall, and many others (e.g., Jacobsen, Follette & Revenstorf, 1984; Jacobson & Truax, 1991), have also highlighted the inability of traditional statistical methods, which compare changes in response to the intervention across treatment conditions, to address the question of whether the treated individuals are equivalent to a normal comparison group following the intervention. This is also important in light of the fact that for many clinical issues, the level attained by the end of therapy is considerably more predictive of long-term functioning than the magnitude of change (e.g., Baucom & Mehlman, 1984). In the 1980s, Kendall and his colleagues (e.g., Kendall & Grove, 1988; Kendall & Norton-Ford, 1982) discussed several methods that attempted to assess the important question of whether the treated and normal comparison populations are equivalent, however Kendall faced several statistical issues that limited the ability of the procedures to directly answer this question. However, advances in the field of equivalence testing led Kendall to develop the highly regarded normative comparisons approach (Kendall, Marrs-Garcia, Nath, & Sheldrick, 1999), which provided a method for evaluating the equivalence of treated and normal comparison groups. Not only did Kendall et al.’s equivalence based approach provide clinical researchers with a fresh approach to the issue of clinical significance, it finally directly addressed the question of whether the treated population was equivalent to a normal comparison group.

An important distinction to make when discussing clinical significance is between methods for evaluating group level and individual level clinical significance. Group level methods for evaluating clinical significance address the question of whether the intervention was effective across the entire treatment group, whereas individual level methods address the effectiveness of an intervention separately for each individual. In this paper we specifically deal with methods for evaluating group level clinical significance, which is not to say that methods for evaluating individual level clinical significance are not important. In fact, methods for evaluating individual level clinical significance due to Jacobsen and his colleagues (e.g., Jacobsen & Truax, 1991) are the most popular of all available methods for assessing clinical significance (Ogles, Lunnen, & Bonesteel, 2001). Further, these methods have been recommended against alternatives in a review of several techniques for assessing clinical significance (Bauer, Lambert & Nielsen, 2004), and can be very effective at calculating the proportion of individuals who are ‘recovered’, ‘unchanged’, etc. However, we tend to favor group level methods for evaluating clinical significance because clinicians, and clinical researchers, are often interested in knowing globally whether an intervention is effective, and this question, in our opinion, is better addressed at the group level.

The purpose of the current paper is to review the equivalence based method for assessing group level clinical significance proposed by Kendall et al. (1999), as well as extend the method by addressing some of the issues that were raised by Kendall and his colleagues in the original paper. Specifically, the goals of this paper are to: 1) Clarify the logic behind conducting only one of the two one-sided *t* tests when conducting the test of equivalence, and provide a simple solution to this issue; 2) Continue the discussion by Kendall et al. on selecting an appropriate equivalence interval and offer a recommendation that is based on utilizing multiple equivalence intervals; 3) Address the important issues (raised by Kendall et al.) of sample size and variance heterogeneity across the treated and normal comparison samples by recommending a heteroscedastic test of equivalence; and 4) Discuss whether the third and fourth steps of Kendall's method (i.e., implementing a traditional test of the difference between the means of the treated and normal comparison groups, and comparing those results to the findings of the equivalence test) are necessary. We end by presenting an applied example that demonstrates the incorporation of the suggestions offered in this paper. The goal is to be able to provide clinical researchers with a meaningful, logical, and easy to implement approach to evaluating the clinical significance of an intervention.

Kendall's Equivalence Based Approach to Clinical Significance

Kendall et al. (1999) raise two important questions that are at the heart of evaluating clinical significance: 1) Are the treated individuals no longer affected by their initial condition?; and 2) Are the treated individuals distinguishable from a normative sample of individuals on relevant measures of the condition? The second question directly addresses the issue of whether the group of treated individuals is equivalent to the group of normal control individuals. It is important to point out that in some cases this question is not a realistic goal of the intervention. For example, Kazdin (2001) states that autism is an example of a disorder with behaviors that are extremely difficult to change and therefore equivalence based methods of demonstrating clinical significance are inappropriate. Wise (2004) also describes dual diagnosis disorders (especially those including medical problems) as an example of a case where improvement to normal may be unrealistic. However, for clinical issues where full (or close to full) recovery is attainable, Kendall et al. suggest that clinical researchers evaluate the second question above directly by determining whether the treated and normal comparison populations are equivalent using the two independent samples test of equivalence proposed by Schuirmann (1987). The normal comparison population would be selected to be as representative of the clinical population as possible, except without having any clinical diagnoses. Population samples may be appropriate in many situations, although in some situations (i.e., when the clinical sample is distinctly different from the population sample on certain characteristics) it is recommended that the researcher collect normative data from a more representative sample. A researcher would declare the treated and normal comparison groups (μ_t and μ_n, respectively) equivalent if H_{o1}: $\mu_t - \mu_n > \delta$ and H_{o2}: $\mu_t - \mu_n < -\delta$ are both rejected. δ represents the critical mean difference for declaring the two population means equivalent; in other words, any mean difference smaller than δ would be considered meaningless within the framework of the experiment. H_{o1} is rejected if $t_1 \leq -t_{\alpha,\mathrm{df}}$ where:

$$t_1 = \frac{\left(M_1 - M_2\right) - \delta}{\sqrt{\frac{\left(n_1 + n_2\right)\left[\left(n_1 - 1\right)s_1^2 + \left(n_2 - 1\right)s_2^2\right]}{n_1 n_2 \left(n_1 + n_2 - 2\right)}}}$$

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

$$t_2 = \frac{\left(M_1 - M_2\right) - (-\delta)}{\sqrt{\dfrac{\left(n_1 + n_2\right)\left[\left(n_1 - 1\right)s_1^2 + \left(n_2 - 1\right)s_2^2\right]}{n_1 n_2 (n_1 + n_2 - 2)}}} \quad .$$

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 $t_{\alpha,\mathrm{df}}$ is the upper-tailed α-level *t* critical value with $n_1 + n_2 - 2$ degrees of freedom (df).

As discussed in detail by Cribbie, Gruman and Arpin-Cribbie (2004), an important consideration with any test of equivalence is the power of the test statistics. For example, with a traditional independent samples *t* test, power for detecting differences between the means increases as sample sizes increase (assuming all other factors are held constant). However, when a test of equivalence is used to explore whether two groups are equivalent, increased sample sizes no longer increase power for detecting differences, but instead increase power for detecting equivalence. The overall power of the equivalence test is a function of the critical mean difference, sample size, the difference between the means of the groups, the variability within the groups, and the Type I error rate; power increases with a larger critical mean difference, larger sample sizes, smaller differences between the means, less variability within the groups and a larger Type I error rate. For a more thorough discussion of the power of equivalence tests under several different conditions see Cribbie et al. (2004).

Extending Kendall's Approach

One t-test or Two? In Kendall et al. (1999) it is stated that "if the range is symmetrical ($|\delta_1| = \delta_2$) around zero, then the two *t* tests are identical; therefore, only one test needs to be conducted" (p. 287). This point is emphasized in Sheldrick, Kendall and Heimberg (2001), where only one of the above *t* tests (t_1, t_2) is used in an empirical example because "the specified range of closeness in this case is symmetrical about the normative mean" (p. 427). It is important to clarify that with the two independent groups equivalence test due to Schuirmann (the procedure that Kendall et al. describe is originally due to Schuirmann, 1987, although they reference a more recent article, Rogers, Howard & Vessey, 1993, that outlines the test due to Schuirmann), that both *t*-tests must be statistically significant in order to declare the groups equivalent. In other words, rejection of H_{o1} implies that $\mu_1 - \mu_2 < \delta$, and rejection of H_{o2} implies that $\mu_1 - \mu_2 > -\delta$. Rejection of both hypotheses implies that $\mu_1 - \mu_2$ falls within the bounds of ($-\delta$, δ) and the means are deemed equivalent.

It is important to point out here that Kendall et al. use the terms δ_1 and δ_2 to represent δ and $-\delta$. From the above formulae it should be evident why both tests need to be conducted, as t_1 will only equal t_2 if $M_1 = M_2$ (which would have near zero probability). Kendall also notes that "if the range is asymmetrical ($|\delta_1| \neq \delta_2$), then only the more stringent t-test corresponding to the smaller delta value needs to be conducted. If this test is significant, then the other must be as well" (Kendall et al., 1999, p. 287). Again, in this situation, it is important to clarify that both tests would need to be conducted. For example, imagine that δ_1 is set at -20 and δ_2 is set at 10. Which test is most stringent will depend on $M_1 - M_2$. If $M_1 - M_2 = -18$ then the test associated with δ_1 will be most stringent, whereas if $M_1 - M_2 = 18$ then the test associated with δ_2 will be most stringent. The only exception to this rule, which Kendall et al. experienced in their applied examples, occurs when you set δ_2 equal to infinity (∞). In this situation, only one test can be conducted because t_1 is undefined with ∞ in the equation. Streiner (2003) describes the approach of one-tailed equivalence testing (i.e., setting one of the equivalence limits to ∞, which is often referred to as noninferiority testing), and explains how it is valuable when the goal is to demonstrate that a new therapy is no worse than the standard therapy. However, we do not recommend this strategy (i.e., setting one of the limits to ∞) when evaluating clinical significance as this negates the possibility of finding that the clinical group is not equivalent to the normative group at posttest because they are actually scoring *better* than the normative group. Although this situation will be rare, and it should be probed extensively to determine the cause, it could highlight situations in which the therapy specifically addresses issues that are probed on the measuring instruments (e.g., questionnaires) and the treated group may show artificially inflated improvements. To summarize, unless one of the equivalence bounds are set to ∞, which we do not recommend, both *t* tests should be conducted in order to establish the equivalence of the treatment and normative conditions.

Establishing an Equivalence Interval. The first step in conducting Schuirmann's test of equivalence is to establish a critical mean difference for declaring two population means equivalent (δ). 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). Within the framework of clinical significance testing, setting δ amounts to establishing what difference between the treated and normative groups at posttest would be clinically meaningless (Cribbie et al., 2004).

The selection of δ 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., ±10%), a percentage of the pooled standard deviation difference, and so on. As δ increases, the probability of declaring the groups equivalent increases, but greater (and potentially important) differences between the groups are considered meaningless. On the other hand, smaller values of δ make it harder to establish equivalence, although there is more confidence that differences between groups declared equivalent are clinically insignificant. In the applied examples of Kendall et al. (1999) and Sheldrick et al. (2001), they utilized an equivalence interval of one standard deviation unit (calculated using the normative group data). As in Kendall's examples, it is most common in equivalence testing to utilize a single value of δ, however, we find this strategy uninformative in equivalence based clinical significance testing because there are clearly different degrees of "closeness" between the treated and normative groups. In other words, using a single value of δ does not allow the researcher to quantify the level of closeness established by the therapy (unless equivalence was established with the smallest practical value of δ or nonequivalence was concluded with the largest practical value of δ).

We recommend that researchers assessing the equivalence of treated and normative groups, assuming that returning the clinical population to normal functioning is a realistic goal of the intervention, utilize the following levels of δ: 1) Definitive equivalence, $\delta = .5\ (s_{normal})$; 2) Probable equivalence, $\delta = s_{normal}$; and 3) Potential equivalence, $\delta = 1.5\ (s_{normal})$, where s_{normal} is the standard deviation of the normal comparison group scores. It is important to highlight that although these values provide a general framework for qualifying equivalence, that researchers are encouraged to consider alternative quantifications/qualifications of δ that may be more appropriate for their specific studies. For example, establishing equivalence with δ = 1.5 may have a completely different meaning in a population that is difficult to return to normal functioning than in a population where returning to normal functioning is a realistic goal of the intervention. Applications of the above values of δ are presented in the examples below.

The Problem of Sample Size and Variance Heterogeneity. Kendall et al. (1999) identified a serious issue with evaluating the equivalence of treated and normal comparison groups with Schuirmann's (1987) approach, namely that the sample sizes and variances of the groups are regularly different. Boneau (1960), Kohr and Games (1974), and many others since, have identified that the independent samples *t* test is not accurate when sample sizes and variances are unequal. More specifically, Boneau found that empirical Type I error rates (when α = .05) could be as large as .16 or as small as .01 when sample sizes and variances are unequal, but that rates for the independent samples Welch *t* test were maintained at approximately α even when sample sizes and variances were extremely disparate. The direction of the bias affecting the independent samples *t* test depends on the pattern of unequal sample sizes and variances. If the larger sample size is paired with the larger variance (and hence the smaller sample size is paired with the smaller variance), then the test will be conservative and it will be difficult to reject H_{o1} and H_{o2} (i.e., power is deflated). If the larger sample size is paired with the smaller variance (and hence the smaller sample size is paired with the larger variance), then the test will be liberal and the probability of committing a Type I error will exceed α. Boneau, Kohr and Games, and others, have also shown that there is only a very slight advantage for the original two independent samples *t* test over the Welch test when sample sizes and variances are equal.

Because Schuirmann's test of equivalence is based on the independent samples *t* test, the sample size and variance inequality issues that affect the independent samples *t* test also affect Schuirmann's equivalence test. Gruman, Cribbie and Arpin-Cribbie (2007) recently demonstrated that empirical Type I error rates for Schuirmann's test of equivalence deviate substantially from the nominal α level when sample sizes and variances are unequal. Gruman et al. also presented a heteroscedastic procedure for testing the equivalence of two independent groups that draws on the heteroscedastic standard error and degrees of freedom due to Welch (1938) and Satterthwaite (1946). For the Schuirmann-Welch test of equivalence, H_{01} is rejected if $t_{W1} \le -t_{\alpha,dfw}$ and H_{02} is rejected if $t_{W2} \ge t_{\alpha,dfw}$ where :

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

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

and

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

Type I error rates for the Schuirmann-Welch are maintained at approximately α even when sample sizes and variances are extremely unequal (Gruman et al.). Further, there is very little power lost by using the Schuirmann-Welch procedure, instead of the original Schuirmann equivalence testing procedure, when sample sizes and variances are equal. Therefore, given that sample sizes and variances are often unequal across treated and normal comparison groups, and that there is very little power lost by the Schuirmann procedure when sample sizes and variances are equal, we recommend that researchers evaluating clinical significance via equivalence testing routinely utilize the Schuirmann-Welch procedure described above.

A related issue is what effect nonnormal distributions will have on the Schuirmann and Schuirmann-Welch tests of equivalence. Although a full treatment of this topic is beyond the scope of this article, previous evidence has indicated that the modified Welch statistics have reasonable Type I error rates when distributions are slightly to moderately skewed and sample sizes and variances are unequal (e.g., Algina, Oshima, & Lin, 1994). However, when distributions become very asymmetric, Welch statistics no longer produce accurate Type I error rates when sample sizes and variances are unequal. Instead, researchers should look to trimmed means or rank-based solutions (e.g., Yuen, 1974; Zimmerman & Zumbo, 1989).

Should a Test of Differences be Done on the Posttest Means? The final two steps of Kendall et al.'s (1999) procedure for assessing clinical significance with equivalence testing require that researchers conduct a traditional two independent-samples *t* test to determine if posttest mean differences exist between the treated and normal comparison groups, and further to compare the results of this test to the results of the equivalence test. In our opinion, this step is inconsistent with the research question being addressed, namely whether or not the groups are equivalent. In other words, since the null hypothesis being evaluated by a two independent samples *t* test is that the population means are exactly equal (i.e., H_o: $\mu_1 = \mu_2$), with a large enough sample size (and recall from the discussion by Achenbach, 2001, that the normative samples can often be very large) there will always be significant differences between the treated and normal comparison groups, regardless of how small the mean differences are. But what does this tell us? Our interest is in whether the differences in the posttest means of the treated and normal comparison groups fall within the established equivalence interval (i.e., $-\delta$ to δ), not whether there are any, potentially trivial, differences between the group means. We should point out here that with a large enough sample size the power of the equivalence test will also approach one, however the fact that equivalence is being evaluated within an interval makes the hypothesis more meaningful. To summarize, we see no reason to conduct a traditional test of mean differences on the posttest means.

However, we do see a lot of value in conducting a traditional test of mean differences between the pretest mean of the treated group and the normal comparison group mean. In other words, if the clinical (i.e., group to be treated) and the normal comparison groups are not different at baseline (assuming ample statistical power), then testing to see if the groups are equivalent at posttest is likely unnecessary. [It is also possible to compare the pretest and/or posttest scores of the control group to the normal comparison group with a traditional test of mean difference, which would contribute information about the status of the control group, although this is not central to the approach discussed in this paper which focuses on the status of the treated population.] It is important to make it clear that the reason for conducting a traditional test of mean difference (as opposed to an equivalence test) is that the research question is whether the groups differ, not whether the groups are equivalent. It should also be clear that it would not be recommended that researchers utilize a standard two independent samples *t* test given that sample sizes and variances, as for the equivalence procedure, will likely be unequal. Therefore, we recommend the Welch (1938) heteroscedastic procedure that is routinely reported in most software packages, and discussed earlier in the paper.

Recommended Procedure for Evaluating Clinical Significance via Equivalence Testing

From the previous discussion, we recommend the following steps in assessing whether the means of a treated and normal comparison group are equivalent:

Step 1: Compare the means of the pretest clinical group (i.e., group to receive the intervention) and the normal comparison group with a two independent samples Welch *t* test. If this test is statistically significant, continue to Step 2. If this test is not statistically significant, then there is no difference between the pretest clinical and normal comparison groups and thus evaluating the equivalence of these groups at posttest is not meaningful.

Step 2a: Determine if the posttest treated group mean is equivalent to the normal comparison group mean using an equivalence interval of $\delta = .5\ (s_{\text{normal}})$, where again s_{normal} is the standard deviation of the normal comparison group scores. If this test is statistically significant, ‘definitive equivalence’ has been established. If this test is not statistically significant, continue to Step 2b.

Step 2b: Determine if the posttest treated group mean is equivalent to the normal comparison group mean using an equivalence interval of $\delta = s_{\text{normal}}$. If this test is statistically significant, ‘probable equivalence’ has been established. If this test is not statistically significant, continue to Step 2c.

Step 2c: Determine if the posttest treated group mean is equivalent to the normal comparison group mean using an equivalence interval of $\delta = 1.5$ (s_{normal}). If this test is statistically significant, ‘potential equivalence’ has been established. If this test is not statistically significant, equivalence of the treated and normal comparison groups cannot be established.

Note that it is possible that equivalence may not have been established at Step 2a, Step 2b or Step 2c because the treated group is actually performing better than the normal comparison group at posttest. Although this may seem like a best case scenario, this outcome should also be cause for investigating whether some aspect of the intervention resulted in the clinical group responding in a biased manner on the posttest measures. For example, if the intervention focused specifically on material covered in the outcome measures, then the treated group, although demonstrating significant therapeutic change on the specific outcome measures utilized, may not demonstrate such extreme improvement on other measures of posttest performance. An anonymous reviewer also highlights that “teaching to the test” is one of several potential threats to the validity of any intervention study that should always be considered when interpreting the results of psychotherapy studies.

Empirical Example

Arpin-Cribbie, Irvine and Ritvo (2009) conducted a randomized clinical trial to evaluate the effectiveness of a 10 week online cognitive behavioral therapy (CBT) for perfectionism. The CBT for perfectionism included topics related to accepting reality, examining and reevaluating expectations, recognizing how certain ways of thinking cause distress, dealing with negative moods, keeping perspective on desires, and dealing with academic and performance anxiety. Using a sample of undergraduate students demonstrating extreme levels of perfectionism, Arpin-Cribbie randomly assigned subjects to receive either the perfectionism based CBT, or no intervention (control). Arpin-Cribbie et al. found that the group receiving the perfectionism based CBT improved significantly more than the control group that received no intervention on several measures of perfectionism. Specifically, the group receiving the CBT improved significantly more than the control group on the: 1) Perfectionism Cognitions Inventory (PCI, Flett, Hewitt, Blankstein, & Gray, 1998); 2) Concern for Mistakes subscale of the Frost Multidimensional Perfectionism Scale (MPSF_CM, Frost, Marten, Lahart, & Rosenblate, 1990); 3) Self Oriented Perfectionism subscale of the Hewitt and Flett Multidimensional Perfectionism Scale (HF_SOP; Hewitt & Flett, 1991); and 4) Socially Prescribed Perfectionism subscale of the Hewitt and Flett Multidimensional Perfectionism Scale (HF_SPP; Hewitt & Flett, 1991). The means and standard deviations for the CBT group, control group, and normal comparison group are presented in Table 1. The normal comparison group data (N = 107) was collected from a sample that was expected to be very similar to the clinical group (i.e., the group that demonstrated elevated perfectionism levels, N = 77); specifically the normal comparison group was comprised of undergraduate students that were at the same academic level as the clinical group, and the data were collected at same time of year as the posttest clinical group data.

An important consideration in evaluating the effectiveness of this therapy is whether the results are ‘clinically significant’. In other words, within the framework of Kendall et al.’s (1999) approach for evaluating clinical significance through equivalence testing, an important question is whether the CBT group is ‘equivalent’ to the normal comparison group at posttest. To evaluate this question we utilized the equivalence testing based approach to assessing group clinical significance described above, and the results are presented in Table 2. The results indicate that the normal comparison group was statistically different from the CBT group at pretest on all measures of perfectionism. As indicated above, this is an important step because if the groups are not different at pretest then the need for an intervention (or evaluating the equivalence of the groups at posttest) is suspect. The results also indicate that the posttest CBT mean was found to be equivalent to the normal comparison group mean on all perfectionism measures, with the groups being declared 'definitively equivalent' on the MPSF_CM and HF_SPP, and 'probable equivalence' was declared for the PCI and HF_SOP. Appendix A provides detailed information on steps for assessing equivalence for the PCI measure.

Discussion

It is now widely recognized that statistical tests for demonstrating that an experimental group has improved significantly more than a control group in a randomized clinical trial fall short of addressing the issue of 'clinical significance'. Further, as psychology (and other disciplines) increasingly value evidence-based therapeutic methods (e.g., Kendall, 1997), it will be very important that valid methods for evaluating clinical significance are available to clinical researchers. Advances in statistical methods for assessing group equivalence (e.g., Schuirmann, 1987) provided the groundwork for Kendall et al.'s (1999) normative comparison based method for assessing clinical significance, which has become the premier method for assessing group level clinical significance. In this paper, we extend the method proposed by Kendall et al. by clarifying the nature of the null hypotheses being conducted in each step of the process, and incorporating recent advances in statistical methods for assessing equivalence (e.g., Gruman et al., 2007).

It is important to highlight that although this paper has addressed many of the issues surrounding the application of equivalence based normal comparison tests in clinical interventions, there are other important issues that require attention. For example, in the paper we briefly introduce the idea that more advanced methods may be required when distributions are extremely nonnormal, or when distribution shapes differ across groups. Solutions to these problems, including trimmed means and rank-based methods may be useful, but more research is necessary before definitive recommendations can be made.

Another important issue, raised by an anonymous reviewer of this paper, is that of non-independence. More specifically, in many clinical studies the clients receiving treatment are nested within the different participating therapists. It is expected that the data analytic strategy that is used to assess whether there is significant improvement in the individuals following the intervention (usually in relation to a control group) would control for any nesting that occurs when multiple therapists are employed (e.g., a hierarchical linear modeling program), and further that ample statistical power is available (which becomes increasingly important in hierarchical designs where it is necessary to ensure that there are enough subjects within each cluster, e.g., therapist). Moreover, an important question that arises is whether normative comparison based tests, such as those discussed in this paper, should be adjusted for the nested nature of the design. There are two ways to address this question. One is to recognize that the normative comparisons are being conducted post hoc and therefore any potential nesting during the intervention is irrelevant to tests being conducted following the intervention. In this manner, the question being asked is whether treated subjects (regardless of which therapist they were assigned to during the intervention) are equivalent to a group of normal comparison subjects following the intervention. The second way to address the problem, that would be most appropriate if significant therapist level effects were identified, would be to take into account these effects when conducting the normative comparisons. A simple method for controlling for the different effects of the therapists would be to investigate normative comparison tests separately within each therapist; in other words, compare treated subjects from each therapist separately to the normal comparison group. The disadvantage of this approach would be that the sample sizes within each therapist may be small and would limit the power of the normative comparison tests. A second approach would be to use posttest means (and standard errors) that are adjusted for therapist level effects. The adjusted means could be obtained from a hierarchical modeling program that allowed for the nesting of subjects within therapists.

The revised approach that we recommend in this paper is intended to provide clinical researchers with a method for specifically addressing the question of whether clients are functioning equivalently to normal controls at the end of the therapeutic process. This does not mean that we are recommending that other potential statistical approaches (e.g., comparing pre-post changes between treated and control groups, evaluating clinical significance at the individual level) are abandoned, but only that these tests provide a very important and unique method for addressing clinical significance that can be used in conjunction with these other statistical methods. In other words, normative comparison based tests of clinical significance should be used in conjunction with statistical tests of the change in outcomes from pretest to posttest (that are preferably relative to a control group, and that also incorporate any nested structures to the data), and individual-level tests of clinical significance (e.g., Jacobsen & Truax, 1991). As was discussed in the introduction, equivalence based normal comparison methods provide the most direct attempt to answer the question of whether the treated group is equivalent to a normally functioning control group. It is important to point out again that while it is appropriate to expect clients with many clinical issues/disorders to have a full (or near full) recovery (i.e., return to normal functioning) during therapy, for many issues or disorders (e.g., autism) it is not realistic to expect clients to return to normal functioning during therapy. Further, as an anonymous reviewer of this manuscript pointed out, even the best designed clinical trials experience non-trivial numbers of participants that fail to respond to the intervention. These non-responsive individuals, in addition to increasing the variability of posttest scores (and therefore reducing the power of normative comparison tests), more importantly highlight the importance of looking at individual level measures of clinical significance as a way of identifying which (and possibly why) specific individuals did not respond to the intervention.

We hope that the revised method for conducting normal comparison based assessments of clinical significance is logical, easy to conduct, and clinically meaningful. In order to make the procedure more widely available, anyone interested in receiving an R program (R is a free statistical software program available at http://www.r-project.org) for conducting the approach outlined in this paper can contact the authors of this paper.

References

Achenbach, T. M. (2001). What are norms and why do we need valid ones? *Clinical Psychology: Science and Practice, 8,* 446-450.

Arpin-Cribbie, C. A., Irvine, J. & Ritvo, P. (2009). Perfectionism related cognitions and psychological distress: A randomized trial evaluating the relative effectiveness of a web-based cognitive behavioural intervention protocol. Unpublished manuscript.

Algina, J., Oshima, T. C., & Lin, W.-Y. (1994) Type I error rates for Welch's test and James's second order test under nonnormality and inequality of variance when there are two groups. Journal of Educational and Behavioral Statistics, 19, 275-292.

Baucom, D. H., & Mehlman, S. K. (1984). Predicting marital status following behavioral marital therapy: A comparison of models of marital relationships. In K. Hahlweg & N. S. Jacobson (Eds.), Marital interactions: Analysis and modification (pp. 89–104). New York: Guilford.

Bauer, S., Lambert, M. J., & Nielsen, S. L. (2004). Clinical significance methods: A comparison of statistical techniques. Journal of Personality Assessment, 82, 60-70.

Boneau, C. A. (1960). The effects of violations of assumptions underlying the t test. *Psychological Bulletin, 57*, 49-64.

Cribbie, R. A., Gruman, J. & Arpin-Cribbie, C. (2004). Recommendations for applying tests of equivalence. *Journal of Clinical Psychology, 60*, 1-10.

Flett, G. L., Hewitt, P. L., Blankstein, K., & Gray, L. (1998). Psychological distress and the frequency of perfectionistic thinking. *Journal of Personality and Social Psychology, 75*, 1363-1381.

Frost, R. O., Marten, P., Lahart, C., & Rosenblate, R. (1990). The dimensions of perfectionism. *Cognitive Therapy and Research, 14,* 449-468.

Gruman, J., Cribbie, R. A., & Arpin-Cribbie, C. A. (2007). The effects of heteroscedasticity on tests of equivalence. *Journal of Modern Applied Statistical Methods, 6,* 133-140.

Hewitt, P. L., & Flett, G. L. (1991). Perfectionism in the self and social contexts: Conceptualization, assessment, and association with psychopathology. *Journal of Personality and Social Psychology, 60,* 456-470.

Jacobsen, N. S., Follette, W. C. & Revenstorf, D. (1984). Psychotherapy outcome research: Methods for reporting variability and evaluating clinical significance. *Behavior Therapy, 15,* 336–552.

Jacobsen, N. S., & Revenstorf, D. (1988). Statistics for assessing the clinical significance of psychotherapy techniques: Issues, problems, and new developments. *Behavioral Assessment, 10*, 133–145.

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

Kazdin, A. E. (2001). *Behavior modification in applied settings (6th ed.)*. New York: Wadsworth.

Kendall, P. C. (1997). Editorial. *Journal of Consulting and Clinical Psychology, 65,* 3-5.

Kendall, P. C., & Grove, W. M. (1988). Normative comparisons in therapy outcome. *Behavioral Assessment, 10,* 147-158

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.

Kendall, P. C. & Norton-Ford, J. D. (1982). *Clinical psychology: Scientific and professional dimensions*. NY: Wiley.

Kohr, R. L., & Games, P. A. (1974). Robustness of the analysis of variance, the Welch procedure and a Box procedure to heterogeneous variances. *Journal of Experimental Education, 43,* 61-69.

Ogles, B. M., Lunnen, K. M., & Bonesteel, K. (2001). Clinical significance: History, application, and current practice. *Clinical Psychology Review, 21,* 421-446.

Rogers, J. L., Howard, K. I. & Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin, 113*, 553-565.

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics, 15*, 657-680.

Sheldrick, R., Kendall, P. C., & Heimberg, R. (2001). The clinical significance of treatments: A comparison of three treatments for conduct disordered children. *Clinical Psychology: Science and Practice, 8,* 418–430.

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

Streiner, D. L. (2003). Unicorns do exist: A tutorial on “proving” the null hypothesis. *Canadian Journal of Psychiatry, 48*, 756-761.

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

Wise, E. A. (2004). Methods for analyzing psychotherapy outcomes: A review of clinical significance, reliable change, and recommendations for future directions. *Journal of Personality Assessment, 82,* 50-59.

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

Zimmerman, D.W., & Zumbo, B.D. (1989). A note on rank transformations and comparative power of the Student *t* test and Wilcoxon-Mann-Whitney test. *Perceptual and Motor Skills, 68,* 1139-1146.

Table 1

Means and Standard Deviations for each Group on each of the Perfectionism Measures.

Group	PCI	MPSF_CM	HF_SOP	HF_SPP
Normal Control	47.37 (16.76)	24.87 (7.04)	66.80 (14.59)	55.56 (11.57)
Treatment_CBT				
Pre	66.14 (15.55)	29.43 (6.94)	85.49 (9.62)	64.83 (13.87)
Post	50.24 (15.72)	23.34 (5.02)	73.20 (10.98)	55.52 (10.84)
Treatment_Control				
Pre	69.75 (12.50)	30.21 (7.87)	84.37 (12.15)	65.92 (13.52)
Post	70.36 (12.35)	30.23 (8.59)	85.17 (14.53)	67.76 (13.25)

Note: PCI = Perfectionism Cognitions Inventory; MPSF_CM = Frost Multidimensional Perfectionism Scale, Concern for Mistakes Subscale; HF_SOP = Hewitt and Flett Multidimensional Perfectionism Scale, Self-Oriented Perfectionism Subscale; HF_SPP = Hewitt and Flett Multidimensional Perfectionism Scale, Socially-Prescribed Perfectionism Subscale; Treatment_CBT = Cognitive Behavioral Therapy Treatment Group; Treatment_Control = Treatment Group that Received No Intervention.

Table 2

Normative Comparisons for each of the Perfectionism Measures in the Arpin et al. Study.

Stage of Testing	PCI	MPSF_CM	HF_SOP	HF_SPP
Step 1: Normal Control Group & Pretest Treatment Group Different?	Yes ($p<.001$)	Yes ($p<.001$)	Yes ($p<.001$)	Yes ($p<.001$)
Decision:	Go to Step 2	Go to Step 2	Go to Step 2	Go to Step 2
Step 2: Normal Control Group & Posttest Treatment_CBT Group Equivalent?				
a) EI = .5 (s_{normal})	No ($p_1 = .053$; $p_2 <.001$)	Yes ($p_1 = .045$; $p_2 <.001$)	No ($p_1 = .359$; ($p_2 < .001$)	Yes ($p_1 = .008$; ($p_1 = .007$)
b) EI = s_{normal}	Yes ($p_1 <.001$; $p_2 <.001$)	NA	Yes ($p_1 < .001$; ($p_2 < .001$)	NA
c) EI = 1.5 (s_{normal})	NA	NA	NA	NA

Note: PCI = Perfectionism Cognitions Inventory; MPSF_CM = Frost Multidimensional Perfectionism Scale, Concern for Mistakes Subscale; HF_SOP = Hewitt and Flett Multidimensional Perfectionism Scale, Self-Oriented Perfectionism Subscale; HF_SPP = Hewitt and Flett Multidimensional Perfectionism Scale, Socially-Prescribed Perfectionism Subscale; Treatment_CBT = Cognitive Behavioral Therapy Treatment Group; EI = Equivalence Interval; s_{normal} = standard deviation of the normal control group on the variable of interest; NA = not applicable because equivalence was declared at a smaller equivalence interval.

Appendix A

$$t_{W1} = \frac{\left(\overline{X}_{CBT} - \overline{X}_{NC}\right) - \delta}{\sqrt{\frac{s_{CBT}^2}{n_{CBT}} + \frac{s_{NC}^2}{n_{NC}}}} = \frac{(50.24 - 47.37) - 8.38}{\sqrt{\frac{15.72^2}{29} + \frac{16.76^2}{107}}} = -1.65$$

Calculations for determining if the CBT group is equivalent to a normal comparison group on PCI scores, using data from Arpin-Cribbie et al. (2009).

Step 1: Determining if the mean of the CBT group at pretest is statistically different from the normal comparison group.

H$_o$: $\mu_{CBT} = \mu_{NC}$

$$t_w = \frac{\overline{X}_{CBT} - \overline{X}_{NC}}{\sqrt{\frac{s_{CBT}^2}{n_{CBT}} + \frac{s_{NC}^2}{n_{NC}}}} = \frac{66.14 - 47.37}{\sqrt{\frac{15.55^2}{29} + \frac{16.76^2}{107}}} = 6.02$$

$$df_w = \frac{\left(\frac{s_{CBT}^2}{n_{CBT}} + \frac{s_{NC}^2}{n_{NC}}\right)^2}{\frac{s_{CBT}^4}{n_{CBT}^2(n_{CBT}-1)} + \frac{s_{NC}^4}{n_{NC}^2(n_{NC}-1)}}$$

$$df_w = \frac{\left(\frac{s_{CBT}^2}{n_{CBT}} + \frac{s_{NC}^2}{n_{NC}}\right)^2}{\frac{s_{CBT}^4}{n_{CBT}^2(n_{CBT}-1)} + \frac{s_{NC}^4}{n_{NC}^2(n_{NC}-1)}}$$

$$= \frac{\left(\frac{15.72^2}{29} + \frac{16.76^2}{107}\right)^2}{\frac{15.72^4}{29^2(29-1)} + \frac{16.76^4}{107^2(107-1)}} = 46.73.$$

$$= \frac{\left(\frac{15.55^2}{29} + \frac{16.76^2}{107}\right)^2}{\frac{15.55^4}{29^2(29-1)} + \frac{16.76^4}{107^2(107-1)}} = 47.27.$$

Thus, since t$_w$ (6.02) > $t_{w,\ \alpha=.05,\ df=47.27}$ (1.67), we reject H$_o$: $\mu_{CBT} = \mu_{NC}$ and continue to Step 2.

Step 2a: Determine if the postttest mean of the CBT group is equivalent to the normal comparison group with δ = .5 (s$_{normal}$).

δ = .5 (s$_{normal}$) = .5 (16.76) = 8.38

H$_{o1}$: μ_{CBT} - μ_{NC} > 8.38; H$_{o2}$: μ_{CBT} - μ_{NC} < - 8.38

$$t_{W2} = \frac{\left(\overline{X}_{CBT} - \overline{X}_{NC}\right) - (-\delta)}{\sqrt{\frac{s^2_{CBT}}{n_{CBT}} + \frac{s^2_{NC}}{n_{NC}}}} = \frac{(50.24 - 47.37) - (-8.38)}{\sqrt{\frac{15.72^2}{29} + \frac{16.76^2}{107}}} = 3.37$$

Therefore, since t_{W2} (3.37) > $t_{w,\ \alpha=.05,\ df=46.73}$ (1.68), but t_{W1} (-1.65) > $-t_{w,\ \alpha=.05,\ df=46.73}$ (-1.68), we do not reject H_{o1}: $\mu_{CBT} - \mu_{NC} > 8.38$ and therefore we conclude that the groups are not equivalent (and continue to Step 2b).

Step 2b: Determine if the postttest mean of the CBT group is equivalent to the normal comparisons group with δ = s_{normal}.

$\delta = s_{normal} = 16.76$

H_{o1}: $\mu_{CBT} - \mu_{NC} > 16.76$; H_{o2}: $\mu_{CBT} - \mu_{NC} < -16.76$

$$t_{W1} = \frac{\left(\overline{X}_{CBT} - \overline{X}_{NC}\right) - \delta}{\sqrt{\frac{s^2_{CBT}}{n_{CBT}} + \frac{s^2_{NC}}{n_{NC}}}} = \frac{(50.24 - 47.37) - 16.76}{\sqrt{\frac{15.72^2}{29} + \frac{16.76^2}{107}}} = -4.16$$

$$t_{W2} = \frac{\left(\overline{X}_{CBT} - \overline{X}_{NC}\right) - (-\delta)}{\sqrt{\frac{s^2_{CBT}}{n_{CBT}} + \frac{s^2_{NC}}{n_{NC}}}} = \frac{(50.24 - 47.37) - (-16.76)}{\sqrt{\frac{15.72^2}{29} + \frac{16.76^2}{107}}} = 5.88$$

$$df_w = \frac{\left(\frac{s^2_{CBT}}{n_{CBT}} + \frac{s^2_{NC}}{n_{NC}}\right)^2}{\frac{s^4_{CBT}}{n^2_{CBT}(n_{CBT} - 1)} + \frac{s^4_{NC}}{n^2_{NC}(n_{NC} - 1)}}$$

$$= \frac{\left(\frac{15.72^2}{29} + \frac{16.76^2}{107}\right)^2}{\frac{15.72^4}{29^2(29 - 1)} + \frac{16.76^4}{107^2(107 - 1)}} = 46.73.$$

Therefore, since t_{W1} (-4.16) < - $t_{w,\ \alpha=.05,\ df=46.73}$ (-1.68) and t_{W2} (5.88) > $t_{w,\ \alpha=.05,\ df=46.73}$ (1.68), we reject H_{o1}: μ_{CBT} - μ_{NC} > 16.76 and H_{o2}: μ_{CBT} - μ_{NC} < - 16.76 and conclude that the groups are equivalent at $\delta = s_{normal}$ = 16.76 (which is labeled ‘probable equivalence). At this point Step 2c is unnecessary because equivalence has been established within a smaller interval than would be evaluated at Step 2c [i.e., 1.5 (s_{normal})].