Equivalence-based Measures of Clinical Significance:

Assessing Treatments for Depression

George Nasiakos, & Robert A. Cribbie

Department of Psychology

York University

Chantal Arpin-Cribbie

Department of Psychology

Laurentian University

Nasiakos, G., Cribbie, R. A., & Arpin-Cribbie, C. A. (2010). Equivalence based tests of clinical significance: Assessing treatments for depression. Psychotherapy Research, 20, 647-656.

Abstract

Treatment efficacy is largely determined by statistical significance testing, and clinical significance testing is often used to quantify or qualify the efficacy of a treatment at the individual or group level. This study applies the equivalence based clinical significance model proposed by Kendall, Marrs-Garcia, Nath and Sheldrick (1999), and a revised model proposed by Cribbie and Arpin-Cribbie (2009), to the assessment of treatments for depression. Using several studies that investigated treatments for depression, we tested whether the post-treatment means were equivalent to the means for a similar normal comparison group. All of the studies had significant improvement from pretest to posttest, although for many of the studies the treated group was not equivalent to a normal comparison group at posttest. Further, there are important differences between the conclusions drawn from the Kendall et al. and Cribbie and Arpin-Cribbie methods for assessing equivalence based clinical significance.

Equivalence-based Measures of Clinical Significance: Assessing Treatments for Depression

Psychotherapy research has uncovered many treatments to mitigate the symptoms of, or help individuals cope with, various ailments. Traditionally, tests of statistical significance act as the hallmark for efficacy, although it is important to highlight that qualitative outcomes also play an important role in understanding the efficacy of an intervention. For example, the patient's subjective evaluation of the intervention effects, or more specifically the therapeutic alliance (Warren, 2001) and personal significance (Sweeney, MacAuley & Pereira Gray, 1998) of the intervention are important factors in evaluating an intervention. From a quantitative standpoint, a "statistically significant" difference between the treatment and control groups in response to an intervention leads readers to infer that the intervention does indeed ameliorate clients' psychological well-being. However, is this sort of assessment sufficient in branding a treatment as "effective"? The treatment may have yielded a considerable improvement in the condition of the client, but overall, was the treatment "clinically significant"?

Jacobson and Truax (1991) argue that tests of "statistical significance" neither indicate the variability of response to the treatment in question within the sample, nor translate to the clinical efficacy of the treatment. Thompson (2002) argues that "statistical significance" alone is not enough to assess the efficacy of a particular treatment, and instead favors using effect sizes, a concept he deems "practical significance". Thompson also discusses "clinical significance", a concept pursued by researchers for several years. Researchers interested in assessing the efficacy of treatments have used the term "clinical significance" to indicate whether or not these

treatments are capable of helping patients return to a state of normalcy. For instance, Ogles, Lunnen and Bonesteel (2001) dichotomize evaluation into "subjective" and "social" components, the former pertaining to the qualitative change in the individual as observed by others, the latter to comparisons to nondeviant peers. An intervention is often described as being clinically significant when the behavior of the clients is indistinguishable from a normal reference group, though the researcher must determine what constitutes a normative level and when the use of such a reference group is appropriate. Finding an appropriate normal reference group can often be difficult, especially when trying to match the clinical and normal comparison groups on important individual characteristics (e.g., age, cultural background).

Assessing Clinical Significance

Several models have been proposed to quantitatively assess "clinical significance" at the individual level. Jacobson and Truax (1991) define clinically significant change as that which brings the client's level of functioning closer to the "functional" population, and cite three potential cut-off points for clinically significant change: either the post-treatment score lies within two standard deviations of the "functional" mean, at least two standard deviations away from the "dysfunctional" mean, or beyond the halfway point between these two values (Jacobson & Truax, 1991; Ogles et al., 2001). Jacobson and Truax also present a "reliable change index" (RCI) to account for overlap in functional and dysfunctional distributions (Jacobson & Truax, 1991; Jacobson, Follette & Revenstorf, 1984); in other words, the RCI is designed to ensure that a post-test score that crosses the "functional" cut-off point is indeed statistically reliable (Jacobson & Truax, 1991). Bauer, Lambert & Nielsen (2004) compared five measures of

individual level clinical significance and recommended the Jacobson and Truax method for its ease of calculation and compatibility with multiple measures. An important characteristic of these methods is that they assess clinical significance at the individual level. However, the methods that will be discussed in this paper take a different approach to quantifying clinical significance, specifically addressing clinical significance at the group level. The assessment of group level clinical significance is performed through the use of equivalence testing methods, which are introduced in the following sections.

Introduction to Equivalence Testing

The goal of mean equivalence testing is to determine if two (or more) group means are comparable (equivalent), within an appropriate interval. Schuirmann (1987) developed a two one-sided testing approach for determining if two group means are equivalent, the hypotheses for which are as follows:

H_{01}: $\mu_1 - \mu_2 \leq \theta_1$; H_{02}: $\mu_1 - \mu_2 \geq \theta_2$

H_{11}: $\mu_1 - \mu_2 > \theta_1$; H_{02}: $\mu_1 - \mu_2 < \theta_2$

where μ_1 and μ_2 refer to the group means and $[\theta_1, \theta_2]$ the lower and upper limits of the "equivalence interval" (which are usually symmetrical, i.e., $\theta_1 = -\theta_2$). Contrary to conventional statistical tests, the composite null hypothesis denotes a difference, while the composite alternate hypothesis relates to equivalence. H_{o1} is rejected if $t_1 \geq t_{\alpha,df}$ where:

$$t_1 = \frac{\left(M_1 - M_2\right) - \theta_1}{\sqrt{\frac{\left(n_1 + n_2\right)\left[\left(n_1 - 1\right)s_1^2 + \left(n_2 - 1\right)s_2^2\right]}{n_1 n_2 \left(n_1 + n_2 - 2\right)}}}$$

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

$$t_2 = \frac{\left(M_1 - M_2\right) - \theta_2}{\sqrt{\frac{\left(n_1 + n_2\right)\left[\left(n_1 - 1\right)s_1^2 + \left(n_2 - 1\right)s_2^2\right]}{n_1 n_2 \left(n_1 + n_2 - 2\right)}}} \quad .$$

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 $t_{\alpha,df}$ is the upper-tailed α-level *t* critical value with $n_1 + n_2 - 2$ degrees of freedom (df). It is important to note that in order to declare the means equivalent both null hypotheses must be rejected; in other words, rejection of H_{O1} implies that the difference in the means is greater than θ_1 and rejection of H_{O2} implies that the difference in the means is less than θ_2. Rogers, Howard and Vessey (1993) define an equivalence interval as a difference in the means that is "important enough to make the groups non-equivalent" (p. 554). From the opposite perspective, it would be the largest difference that is meaningless within the framework of the research.

Clinical Significance through Equivalence Testing

Kendall, Marrs-Garcia, Nath and Sheldrick (1999) outlined a 'normative comparisons' approach which involves evaluating the equivalence of a treated population and a representative normative population. These normative comparisons can provide insight into the efficacy of a clinical intervention. In other words, if the goal of a treatment is to return the participants to a state of normal functioning, then a logical way to evaluate the treatment is to determine if the

mean of the treated population is equivalent to the mean of the normal population. Kendall et al. are in accord with Rogers, Howard and Vessey (1993), as they believe that determining an equivalence interval must be tailored to the application in question, though they suggest, as a heuristic, a range of ±1.0 standard deviations (SDs) from the normative mean (i.e., $\theta_1 = (-1)$ SD_{normal} and $\theta_2 = (1)$ SD_{normal}. Kendall et al. propose testing for both the equivalence of the means and for differences in the means.

There are two approaches by which normative data can be obtained, the researcher can utilize existing published data or the researcher can collect her or his own normative sample (Kendall et al., 1999). The latter is often a preferred option, as the treated group will often not be comparable to the normative group on important dimensions. Kendall and Sheldrick (2000) cite several advantages to collecting representative normative data: (1) the process can be tailored to the target research question; (2) normative data can be obtained for all required measures, rather than using published data in a piecemeal manner; (3) one can control for test-sensitization and reactivity; and (4) the normative data be matched to the clinical data on important demographic characteristics. However, despite its numerous advantages, collecting data is quite expensive, requiring much time and money, thereby prompting some investigators to favor using published normative data as a basis for normative comparisons.

Advancing Clinical Significance through Equivalence Testing.

Recently, Cribbie and Arpin-Cribbie (2009) expanded upon the normative comparisons approach proposed by Kendall et al. (1999). The Cribbie and Arpin-Cribbie method involves two (hierarchical) steps: First, the pre-test mean is compared to the normative mean using a

difference-based two independent samples *t* test with a non-pooled standard error (i.e., a test that does not assume that the variances of the groups are equal). If the pre-test and normative means differ, then the second step is to evaluate whether the post-test and normative group means are equivalent (if the pre-test and normative means do not differ, then no further testing is conducted because the treatment group and normative group do not differ before the intervention).

With regard to a test statistic for evaluating equivalence, it is important that the test statistic be robust to situations in which the sample sizes and variances of the groups differ because often the normative group is larger and less variable than the clinical group (which can substantially inflate the Type I error rate of the Schuirmann test, see Gruman, Cribbie & Arpin-Cribbie, 2007). Kendall et al. (1999) recognized the problem of variance heterogeneity in equivalence testing, but failed to account for it in their model (i.e., their method utilizes the original Schuirmann procedure). To more formally illustrate this issue, imagine a group of individuals treated for major depression. Many will likely respond satisfactorily to an established intervention, but some will experience little or no change (or even deterioration) and some will improve substantially. This will in most cases lead to a post-test distribution with greater variability than the distribution of depression scores in the normal population. In other words, though the distribution of scores in the general population will undoubtedly include outliers, typically the distribution will be narrower than that of the treated population.

To address this problem, Cribbie and Arpin-Cribbie integrate Welch's (1938) heteroscedastic standard error and degrees of freedom into their model (Schuirmann-Welch test), thus accounting for the sometimes vast differences between sample sizes and variances of the treated and normal samples. Cribbie and Arpin-Cribbie assess equivalence using the

Schuirmann-Welch method at intervals of 0.5, 1.0 and 1.5 SDs (of the normative group). Assessing equivalence at multiple levels affords the researcher the opportunity to determine the degree to which a declaration of equivalence is made. Declarations of equivalence at the .5, 1.0 and 1.5 SD levels were labeled 'definitive equivalence', 'probable equivalence', and 'potential equivalence', respectively, by Cribbie and Arpin-Cribbie. Cribbie and Arpin-Cribbie also recommend against using infinity as a limit (which was used by Kendall & Sheldrick, 2000, in conducting normative comparisons) for the researcher then disregards instances in which the treatment group outperforms the normal comparison group, which might be an interesting event in and of itself. Oddly, Bauer et al. (2004) make no mention of Kendall et al. (1999) in their review of methods for evaluating clinical significance testing, although it is quite conceivable to presume that these models were simply overlooked by the researchers when assessing methodologies. However, it is also quite common for one to be reluctant to utilize new advancements in assessment, much like with any new discovery in any field, especially if the established model is perceived as valid and reliable.

Application of Equivalence-Based Measures of Clinical Significance

Sheldrick, Kendall & Heimberg (2001) employed the Kendall et al. (1999) methodology when testing treatments for conduct disorder. Their study examined three treatments across 14 studies, comparing treatment scores to age-appropriate normative data on the Child Behavior Checklist (Achenbach & Edelbrock, 1983) and the Eyberg Child Behavior Inventory (Eyberg & Ross, 1978). A total of 50 sets of pre- and post-treatment scores were examined across 23 treatment conditions, and out of these, all but three exhibited clinically significant change as

indicated by the Jacobson and Truax (1991) "reliable change index" (adapted for group comparisons). Sixteen out of the 50 post-treatment scores were deemed "equivalent" (significant equivalence test and nonsignificant difference test), 32 "different" (non-significant equivalence test and significant difference test) and two "equivocal" (non-significant equivalence test and non-significant difference test)

The Current Study

The purpose of this study is to evaluate the utility of the Kendall et al. (1999) and Cribbie and Arpin-Cribbie (2009) equivalence-based methods of clinical significance by applying these methods to a sample of studies employing interventions for depression. Depression was selected because it is a common ailment affecting a vast number of people, and a large number of studies have been published assessing treatments; for example, cognitive behavioural therapy (Bodenmann et al., 2008; MacKinnon, Griffiths & Christensen, 2008; Horowitz, Garber, Ciesla, Young & Mufson, 2007; Kingston, Dooley, Bates, Lawlor & Malone, 2007), problem-solving therapy (Eskin, Ertekin & Demir, 2008) and emotion-focused therapy (Ellison, Greenberg, Goldman & Angus, 2009). This study compares the post-treatment scores of clinical trials to published normative data on corresponding measures. The aim of this study is three-fold: (1) evaluate the clinical significance of interventions using equivalence testing approaches, (2) compare the conclusions derived from the equivalence based tests of clinical significance to the traditional tests of significance, and (3) compare the results for the Kendall et al. and Cribbie and Arpin-Cribbie equivalence-based measures of clinical significance.

Method

Selection of Studies

We applied various search criteria in the PsycINFO and Google Scholar databases, including combinations of "depression", "treatment", "intervention" and "randomized control trial", and retrieved studies accordingly. The studies utilized numerous outcome variables, including two versions of the Beck Depression Inventory, BDI (Beck, Ward, Mendelson, Mock & Erbaugh, 1961) and BDI-II (Beck, Steer & Brown, 1996), the Center for Epidemiological Studies Depression scale (CES-D; Radloff, 1977), and the Hamilton Rating Scale for Depression (HRSD; Hamilton, 1960) and its derivatives.

Studies must have reported post-test means, standard deviations and sample sizes to be utilized in this study. As equivalence testing compares treated to normal groups, waiting list and other control groups were omitted. All but one study included pre-test means and standard deviations, which were used, with the post-test data, to calculate Cohen's standardized effect size (*d*) and whether there was a statistically significant change from pre-test to post-test. Although a few studies included measures of effect size, in many cases it was difficult to determine how it was computed (e.g., just for the pre-post change or for the pre-post change relative to a control group) and many studies did not include an effect size. Therefore, Cohen's *d* was calculated for each study assuming a correlation between pre-test and post-test of $r = .5$. Follow-up data provide insight into persistence and relapse rates, but were not considered necessary for this investigation. Furthermore, with respect to symptoms, there must exist the potential for clients to return to a state of normalcy (see Cribbie & Arpin-Cribbie, 2009). For example, one might not anticipate a group of individuals to achieve a state of normal functioning on, say, a depression

scale, if the group also suffers from a comorbid disorder, such as avoidant personality disorder, that limits the effectiveness of an intervention (e.g., Papakostas et al., 2003).

Normative Comparisons

In order to determine clinical significance, treatment groups must be compared to a sample representative of the "normal" population. For a most appropriate comparison, it is advised that representative normative data be collected by the researchers (Kendall & Sheldrick, 2000), or that normative data that matches important characteristics of the sample (e.g., age, gender, socioeconomic status) be identified in the literature. It would be impractical for us to attempt to collect representative data to match each study we identified, and thus we used existing published normative data for our comparisons (in each case trying to find normative data that closely matched the characteristics of the populations used in the intervention studies).

Statistical Analysis

We modeled both the Kendall et al. (1999) and the Cribbie and Arpin-Cribbie (2009) procedures with the *R* statistical software application (R Foundation for Statistical Computing, 2008). For Kendall et al. (1999), the equivalence interval was set at plus or minus 1.0 SD of the normative sample. For Cribbie and Arpin-Cribbie (2009), three equivalence intervals were utilized, specifically 0.5, 1.0 and 1.5 times the SD of the normative sample. For each of the Kendall et al. and Cribbie and Arpin-Cribbie methods, the treated and normative samples are deemed equivalent if the null hypothesis for both t_1 and t_2 can be rejected.

Results

The references for published normative data used in this investigation are presented in Table 1. Appropriate normative data could only be found for the BDI, BDI-II, CES-D, and 17-item version of the HRSD, and therefore other versions of the HRSD were excluded. An important finding of this study was that it is very difficult to find representative normative data for many of the depression scales. Specifically, although we believe we have found acceptable normative data for each of the depression studies we collected, the lack of normative data for many of these scales limited how precise we could be in matching the characteristics of the treated and normative samples. For example, MacKinnon et al. (2008) conducted their study of Internet cognitive-behavioural therapy for depression using the CES-D with adults from Great Britain, yet their results were compared with a normative sample of CES-D scores consisting of adults in the Netherlands (Bouma, Ranchor, Sanderman, & Van Sonderen, 1995) because more representative data was not available. Future studies will hopefully improve on the availability of representative normative data.

Fifteen studies were selected that investigated 22 treatments for depression. Some studies included multiple samples, multiple interventions and/or multiple outcome variables. In order to limit the number of studies, and avoid non-independence issues, only one sample, intervention and/or outcome was randomly selected from each study.

Table 2 summarizes pre- and post-test data for each trial, sorted by clinical measure. Specifically, these tables highlight the treatments investigated in each trial, the means, standard deviations and sample sizes of each group, whether there was a statistically significant change from pre-test to post-test, and the effect size for the pre-test to post-test change (Cohen's *d*). All

of the interventions had significant change from pre-test to post-test and had Cohen's *d* values greater than .5.

Table 3 summarizes findings of "clinical significance" as ascertained from the Kendall and Cribbie and Arpin-Cribbie methods, sorted by clinical measure. The first step of the Cribbie and Arpin-Cribbie method, comparing the pre-test and normative means, had to be skipped for one study (Cuijpers, Smit, Woordouw, & Kramer, 2005) because pre-test data were not provided. For the remaining studies the pre-test means all differed significantly from the normative means.

Of the 15 treated groups we investigated, 11 exhibited statistically significant differences and the treated samples were found to be equivalent to the normal comparison group using the Kendall method. On the other hand, only five exhibited the same result using the Cribbie and Arpin-Cribbie method at the 1.0 SD interval, and 10 at the 1.5 SD interval. Therefore, four groups exhibited statistically significant differences, but the treated group was not equivalent to the normal comparison group according to the Kendall method, whereas 10 exhibited the same result using the Cribbie and Arpin-Cribbie method at the 1.0-SD interval, and 5 at the 1.5 SD interval.

With respect to the Cribbie and Arpin-Cribbie method, it is clear that declaring the treated group equivalent to the normal comparison group with an equivalence interval of half the SD of the normal comparison group (0.5 SD) is very difficult, relative to declarations of equivalence at the 1.0 and 1.5-SD intervals. In fact, only one study (Ellison et al., 2009) featured a treatment (emotion-focused therapy) that was clinically significant at the 0.5 SD level.

Table 4 compares the frequencies of equivalence decisions for both the Kendall and Cribbie and Arpin-Cribbie methods at an equivalence interval of 1.0 SD. Most importantly, there

were six instances in which the Kendall method declared the treated participants equivalent to the normal comparison group and the Cribbie and Arpin-Cribbie method did not. The only difference between the methods is that the Kendall et al. procedure uses the traditional standard error and degrees of freedom from the two independent samples t-test, whereas the Cribbie and Arpin-Cribbie model uses a non-pooled standard error and adjusted degrees of freedom, as suggested by Welch (1938), that are robust to violations of the variance homogeneity assumption. Therefore, only the results due to the Welch-based statistics can be considered accurate.

Discussion

Clinical significance testing provides a necessary complement to traditional statistical significance testing. A statistically significant result tells the reader that the treatment was effective in reducing the potency of symptoms (often relative to a control group), but does not imply that the client has returned to a state of "normal" functioning, which is the primary goal of most intervention studies for depression. In the current study, we found that, regardless of the statistical significance of the change from pre- to post-test or the associated effect size, the equivalence-based measures of clinical significance added important information about the effectiveness of the intervention. In short, an intervention can have a statistically significant effect on clients, i.e., it can significantly mitigate one's symptoms, but not necessarily result in a return to a "normal" state, as compared with normative scores on a given clinical measure. This finding has a particularly important implication pertaining to the administration of an intervention. A given intervention may render a statistically significant difference from pre- to

post-test for the treatment group, but if the intervention does not return the group to a state of normal functioning, this may highlight potential issues with the treatment (e.g., not enough time for complete results to be observed). In the current study, we found that a variety of treatments were effective in returning clients to a state of normal functioning, including mindfulness-based CBT, emotion-focused therapy, and various online therapies.

Traditional methods that examine clinical significance at the individual level (e.g., Jacobson & Truax, 1991) are beneficial when assessing the efficacy of an intervention for a particular individual, or for knowing the proportion of individuals that improved, deteriorated, etc. However, testing for clinical significance at the group level (i.e., testing for equivalence of the treatment group and a normative sample) better assesses overall treatment efficacy. More specifically, one can assess clinical significance by examining the performance of each individual in a treatment group and noting for whom the particular intervention was effective, but when assessing the effectiveness of the intervention itself, a test of the equivalence of the treated group and an appropriate normal comparison group can be very effective. However, it is important to mention that this method is only appropriate when the goal of the intervention is to return clients to a state of normal functioning.

The results of this study illustrate the importance of two important factors in equivalence-based clinical significance testing: (1) accounting for variance heterogeneity; and (2) utilizing multiple equivalence intervals. First, the Kendall et al. model incorporates the traditional standard error and degrees of freedom from the sampling distribution of the difference between two means; i.e. it utilizes the original Schuirmann (1987) test of equivalence. When comparing a group treated with an intervention to a normal comparison group, often the normative sample

will both be much larger in size and have smaller variability than the treated sample. This combination of a large sample size with a small amount of variability for the normative sample (and hence a small sample size with a large amount of variability for the treated sample) means that the standard error for the *t* tests used in the Kendall et al. (1999) method will be extremely underestimated. In other words, the smaller variability of the normal comparison group gets weighted much higher than the larger variability of the treated group when computing the pooled variance term (and hence the pooled standard error). Thus, the Kendall et al. method will have an inflated Type I error rate; i.e. this method will often declare treated and normative comparison populations equivalent to a normal comparison group when, in fact, these populations are not equivalent to this group (see Gruman et al., 2007). Another important consequence of the use of non-pooled standard error by the Cribbie and Arpin-Cribbie (2009) method is that this test has been shown to more robust to moderate violations of normality than tests that adopt a pooled standard error (Algina, Oshima, & Lin, 2004). However, if the non-normality is more extreme, robust estimators (e.g., trimmed mean) may be necessary (Wilcox, 1994). When comparing the Kendall et al. and Cribbie and Arpin-Cribbie (2009) methods at an equivalence interval of 1.0 standard deviation, in all cases in which the results did not match, the former returned an "equivalent" result while the latter returned a "not equivalent" result. It is expected that the majority of these equivalence decisions for the Kendall method are Type I errors (although obviously there is no way to verify this). Therefore, when using equivalence-based testing for clinical significance, if the homogeneity of variance assumption is violated, the validity of the results of the Kendall approach must be treated with caution.

Using multiple equivalence levels also adds a measure of confidence when determining

whether or not a treated group is indeed equivalent to the normal group. For example, if a treated group falls within 0.5 standard deviations of the normative mean, we can be confident that the two groups are equivalent; however, if the treated group falls within 1.5 standard deviations of the normative mean, readers should not have as much confidence in the result.

Another very important finding of this study is the extreme paucity of normative/referent data for the outcomes investigated. This was a very surprising finding given that we were investigating very popular depression scales (e.g., BDI). The lack of normative data for these, and likely many other, psychological scales is a significant hindrance to researchers conducting normative comparisons (if they have not collected separate normative data). More specifically, researchers will often be interested in matching their treated and normative samples on many different characteristics, and currently this will be very difficult to do given the lack of normative data available in the literature. It is hoped that future studies will collect normative data from a wide variety of populations.

There are a couple of limitations to the methodology of the current study. First, although we have tried to obtain a large sample of intervention studies for depression, this is obviously not an exhaustive list. Thus, it is unclear how the results (effect sizes, traditional statistical tests, normative comparisons) of the studies not selected for investigation would compare to those selected. Second, the effect sizes (Cohen's *d*) utilized in this study for the pretest to posttest change were calculated under the assumption of a .5 correlation between pretest and posttest, because in most cases the standard error of the difference was not provided, or it was unclear how the effect size that was provided was calculated. It is possible that the true effect sizes deviated slightly from the calculated values.

To summarize, this study examined interventions for depression and the results highlight that several therapeutic interventions are effective at not only improving the subject's well being, but also returning the subjects depression levels to those similar to that of a normal comparison group. In the future, we recommend that equivalence-based normative comparisons be utilized whenever an intervention for a psychological malady is undertaken and the goal is to return the members of the clinical group to a state of normal functioning. An R (The R Foundation for Statistical Computing, 2008) function for conducting the normative comparison tests discussed in this article is available at http://www.psych.yorku.ca/cribbie.

References

Achenbach, T.M. & Edelbrock, C.S. (1983). *Manual for the Child Behavior Checklist and Revised Child Behavior Profile.* Burlington, VT: University Associates in Psychiatry.

Achenbach, T., & Edelbrock, C. (1986). *Manual for the TRF and the Child Behavior Profile.* Burlington, VT: University of Vermont.

Algina, J., Oshima, T. C. & Lin, W. (1994). Type I error rates for Welch's test and James' second-order test under nonnormality and inequality of variance when there are two groups. *Journal of Educational and Behavioral Statistics, 19,* 275-291.

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

Beck, A.T., Steer, R.A. & Brown, G.K. (1996). *Beck Depression Inventory-II Manual.* San Antonio, TX: The Psychological Corporation.

Beck, A.T., Ward, C.H., Mendelson, M., Mock. J. & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry, 4,* 561-571.

Bodenmann, G., Plancherel, B., Beach, S.R.H., Widmer, K., Gabriel, B., Meuwly, N., Charvoz, L., Hautzinger, M. & Schramm, E. (2008). Effects of coping-oriented couples therapy on depression: A randomized clinical trial. *Journal of Consulting and Clinical Psychology, 76,* 944-954.

Bouma, J., Ranchor, A.V., Sanderman, R., & Van Sonderen, E. (1995). *Het meten van Symptomen van depressie met de CES-D; een handleiding [The measurement of depressive symptoms with the CES-D; a manual]*. Groningen: Noordelijk Centrum voor Gezondheidsvraagstuk ken.

Campbell, A. (2008). Clinical significance in real-world settings. *Australian and New Zealand Journal of Family Therapy, 29,* 107-110.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences.* Hillsdale, NJ: Lawrence Erlbaum.

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

Cuijpers, P., Smit, F., Woordouw, I. & Kramer, J. (2005). Outcome of cognitive behaviour therapy for minor depression in routine practice. *Psychology and Psychotherapy: Theory, Research and Practice, 78,* 179-188.

De Berardis, D., Carano, A., Gambi, F., Campanella, D., Giannetti, P., Ceci, A., Mancini, E., La Rovere, R., Cicconetti, A., Penna, L, Di Matteo, D., Scorrano, B., Cotellessa, C., Salerno, R.M., Serroni, N. & Ferro, F.M. (2007). Alexithymia and its relationships with body checking and body image in a non-clinical female sample. *Eating Behaviors, 8,* 296-304.

Dimidjian, S., Hollon, S.D., Dobson, K.S., Schmaling, K.B., Kohlenberg, R.J., Addis, M.E., Gallop, R., McGlinchey, J.B., Markley, D.K., Gollan, J.K., Atkins, D.C., Dunner, D.L. & Jacobson, N.S. (2006). Randomized trial of behavioral activation, cognitive therapy, and antidepressant medication in the acute treatment of adults with major depression. *Journal of Consulting and Clinical Psychology, 74,* 658-670.

Ellison, J.A., Greenberg, L.S., Goldman, R.N. & Angus, L. (2009). Maintenance of gains following experiential therapies for depression. *Journal of Consulting and Clinical Psychology, 77,* 103-112.

Eskin, M., Ertekin, M. & Demir, H. (2008). Efficacy of a problem-solving therapy for depression and suicide potential in adolescents and young adults. *Cognitive Therapy and Research, 32*, 227-245.

Eyberg, S.M. & Ross, A.W. (1978). Assessment of child behavior problems: The validation of a new inventory. *Journal of Clinical Psychology, 16,* 113–116.

Friedman, M.A., Cardemil, E.V., Uebelacker, L.A., Beevers, C.G., Chestnut, C. & Miller, I.W. (2005). The GIFT program for major depression: Integrating group, individual, and family treatment. *Journal of Psychotherapy Integration, 15*, 147-168.

Gerrits, R.S., van der Zanden, R.A.P., Visscher, R.F.M. & Conijn, B.P. (2007). Master your mood online: a preventive chat group intervention for adolescents. *Australian e-Journal for the Advancement of Mental Health, 6,* 1-11.

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.

Grundy, C.T., Lambert, M.J. & Grundy, E.M. (1996). Assessing clinical significance: The Hamilton Rating Scale for Depression. *Journal of Mental Health, 5,* 25-33.

Hageman,W.J. & Arrindell,W.A. (1999). Establishing clinically significant change: Increment of precision between individual and group level of analysis. *Behavior Research and Therapy, 37,* 1169–1193.

Haringsma, R., Engels, G.I., Cuijpers, P. & Spinhoven, P. (2006). Effectiveness of the Coping With Depression (CWD) course for older adults provided by the community-based mental health care system in the Netherlands: a randomized controlled field trial. *International Psychogeriatrics, 18,* 307-325.

Hellerstein, D.J, Batchelder, S., Hyler, S., Arnaout, B., Corpuz, V., Coram, L. & Weiss, G. (2008). Aripiprazole as an adjunctive treatment for refractory unipolar depression. *Progress in Neuro-Psychopharmacology & Biological Psychiatry, 32,* 744-750.

Horowitz, J.L., Garber, J., Ciesla, J.A., Young, J.F. & Mufson, L. (2007). Prevention of depressive symptoms in adolescents: A randomized trial of cognitive-behavioral and interpersonal prevention programs. *Journal of Consulting and Clinical Psychology, 75,* 693-706.

Hsu, L.M. (1989). Reliable changes in psychotherapy: Taking into account regression toward the mean. *Behavioral Assessment, 11,* 459–467.

Jacobson, N.S., Follette, W.C. & Revenstorf, D. (1984). Psychotherapy outcome research: Methods for reporting variability and evaluating clinical significance. *Behavior Therapy, 15,* 336-352.

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.

Joseph, S., Lewis, C.A. & Olsen, C. (1996). Convergent validity of the Depression-Happiness Scale with measures of depression. *Journal of Clinical Psychology, 52,* 551-554.

Karavidas, M.K., Lehrer, P.M., Vaschillo, E.G., Vaschillo, B., Marin, H., Buyske, S., Malinovsky, I., Radvanski, D. & Hassett, A. (2007). Preliminary results of an open label study of heart rate variability biofeedback for the treatment of major depression. *Applied Psychophysiology and Biofeedback, 32,* 19-30.

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. & Sheldrick, R.C. (2000). Normative data for normative comparisons. *Journal of Consulting and Clinical Psychology, 68,* 767-773.

Kingston, T., Dooley, B., Bates, A., Lawlor, E. & Malone, K. (2007). Mindfulness-based cognitive therapy for residual depressive symptoms. *Psychology and Psychotherapy: Theory, Research and Practice, 80,* 193-203.

Learmonth, D., Trosh, J., Rai, S., Sewell, J. & Cavanagh, K. (2008). The role of computer-aided psychotherapy within an NHS CBT specialist service. *Counselling and Psychotherapy Research, 8,* 117-123.

MacKinnon, A., Griffiths, K.M. & Christensen, H. (2008). Comparative randomised trial of online cognitive-behavioural therapy and an information website for depression: 12-month outcomes. *The British Journal of Psychiatry, 192,* 130-134.

Murrell, S.A., Himmelfarb, S. & Wright, K. (1983). Prevalence of depression and its correlates in older adults. *American Journal of Epidemiology, 117,* 173-185.

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

Papakostas, G. I., Petersen, T. J., Farabaugh, A. H., Murakami, J. L., Pava, J. A., Alpert, J. E., Fava, M., & Nierenberg A. A. (2003). Psychiatric comorbidity as a predictor of clinical response to nortriptyline in treatment-resistant major depressive disorder. *Journal of Clinical Psychiatry, 64,* 1357-61.

R Foundation for Statistical Computing (2008). *R version 2.7.2.* Vienna, Austria: Author.

Radloff, L.S. (1977). The CES-D Scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement, 1,* 385-401.

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 assigning equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics, 15,* 657-680.

Seggar, L.B., Lambert, M.J. & Hansen, N.B. (2002). Assessing clinical significance: Application to the Beck Depression Inventory. *Behavior Therapy, 33,* 253-269.

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

Speer, D.C. (1992). Clinically significant change: Jacobson & Truax (1991) revisited. *Journal of Consulting and Clinical Psychology, 60,* 402–408.

Speer, D.C. & Greenbaum, P.E. (1995). Five methods for computing significant individual client change and improvement rates: Support for an individual growth curve approach. *Journal of Consulting and Clinical Psychology, 63,* 1044–1048.

Spek, V., Nyklicek, I.,Smits, N., Cuijpers, P., Riper, H., Keyzer, J. & Pop, V. (2007). Internet-based cognitive behavioural therapy for subthreshold depression in people over 50 years old: A randomized controlled clinical trial. *Psychological Medicine, 37,* 1797-1806.

Spitzer, R.L. & Wakefield, J.C. (1999). DSM-IV Diagnostic criterion for clinical significance: does it help solve the false positives problem? *The American Journal of Psychiatry, 156,*

1856-1864.

Strauman, T.J., Vieth, A.Z., Merrill, K.A., Kolden, G.G., Woods, T.E., Klein, M.H., Papadakis, A.A., Schneider, K.L. & Kwapil, L. (2006). Self-system therapy as an intervention for self-regulatory dysfunction in depression: A randomized comparison with cognitive therapy. *Journal of Consulting and Clinical Psychology, 74,* 367-376.

Sweeney, K. G., MacAuley, D., & Pereira Gray, D. (1998). Personal significance: The third dimension. *The Lancet,* 351, 134-136.

Teri, L. (1982). The use of the Beck Depression Inventory with adolescents. *Journal of Abnormal Child Psychology, 10,* 277-284.

Thompson, B. (2002). 'Statistical', 'Practical', and 'Clinical': How many kinds of significance do counselors need to consider? *Journal of Counselling and Development, 80,* 64–71.

Uebelacker, L.A., Courtnage, E.S. & Whisman, M.A. (2003). Correlates of depression and marital dissatisfaction: Perceptions of Marital Communication Style. *Journal of Social and Personal Relationships, 20,* 757-769.

Warren, C. S. (2001). Negotiating the therapeutic alliance: A relational treatment guide. *Psychotherapy Research, 11,* 357 - 359.

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). Some results on the Tukey-McLaughlin and Yuen methods for trimmed means when distributions are skewed. *Biometrical Journal, 36,* 259–273.

Table 1
Published normative data used in the normative comparison analyses.

Scale	Reference	M	SD	N	Characteristics
BDI	De Berardis et al. (2007)	9.0	8.4	288	Undergraduate women in Italy
	Seggar et al. (2002)	7.22	6.33	28,905	Community sample
	Teri (1982)	8.47	8.03	568	Adolescents in high school
	Joseph et al. (1996)	7.33	6.54	194	Undergraduate students in N. Ireland
BDI-II	Uebelacker et al. (2003)	7.4*	6.6*	127	Married adults in U.S.A.
CES-D	Joseph et al. (1996)	14.23	10.87	194	Undergraduate students in N. Ireland
	Murrell et al. (1983)	9.36	9.2	936	Persons age 55 and over in Kentucky
	Bouma et al. (1995)	9.7	8.6	2768	Citizens of the Netherlands
HRSD-17	Grundy et al. (1996)	6.25	4.24	203	General population in U.S.A.

Note: * = average of female and male means

Table 2
Statistical data from the target studies.

Reference	Intervention	Pre-test			Post-test			Pre-Post Change p<.05	Cohen's d
		M	SD	N	M	SD	N		
BDI									
Kingston et al. (2007)	Mindfulness-based CBT	30.3	7.6	6	12.3	9.7	6	Yes	2.23
Eskin et al. (2008)	Problem-solving	26.7	9.4	27	10.7	10.4	27	Yes	1.64
Ellison et al. (2009)	Emotion-focused	26.3	6.9	27	6.7	5.8	27	Yes	3.11
Straumann et al. (2006)	Cognitive Therapy	24.6	6.2	18	10.7	7.1	18	Yes	2.13
BDI-II									
Friedman et al. (2005)	GIFT program	25.4	7.1	13	9.9	8.7	13	Yes	2.01
Learmonth et al. (2008)	Beat the Blues	24.2	11.1	244	15.8	11.0	244	Yes	0.76
Spek et al. (2007)	Internet-based CBT	19.1	7.2	102	11.9	8.0	102	Yes	0.95
Karavidas et al. (2007)	HRV biofeedback	26.1	3.4	11	15.8	2.4	8	Yes	3.64
CES-D									
MacKinnon et al. (2008)	Internet-based CBT	21.8	10.5	182	15.9	9.8	136	Yes	0.58
Gerrits et al. (2007)	Online interaction	32.6	9.3	50	18.7	9.4	50	Yes	1.5
Haringsma et al. (2006)	Coping class	31.9	8.2	21	21.5	9.6	21	Yes	1.19
Cuijpers et al. (2005)	Coping class	Not reported		187	17.0	9.9	128	NA	0.84*
HRSD-17									
Bodenmann et al. (2003)	Interpersonal psychotherapy	13.9	3.3	20	9.3	5.8	18	Yes	0.94
Dimidjian et al. (2006)	Cognitive therapy	22.7	2.6	25	10.3	7.6	18	Yes	1.91
Hellerstein et al. (2008)	Adjunctive aripiprazole	21.6	4.3	14	12.6	7.5	14	Yes	1.43

Note: CBT = Cognitive Behavioral Therapy; HRV = Heart Rate Variability; * = value reported in the paper.

Table 3
Results for the Kendall et al. and Cribbie & Arpin-Cribbie normative comparisons.

Reference	Intervention	Normative Reference	Kendall 1.0 SD	Cribbie & Arpin-Cribbie		
				0.5 SD	1.0 SD	1.5 SD
BDI						
Kingston et al. (2007)	Mindfulness-based CBT	De Berardis et al. (2007)	E	N/E	N/E	E
Eskin et al. (2008)	Problem-solving	Teri (1982)	E	N/E	E	E
Ellison et al. (2009)	Emotion-focused	Seggar et al (2002)	E	E	E	E
Straumann et al. (2006)	Cogntive Therapy	Seggar et al (2002)	E	N/E	N/E	E
BDI-II						
Friedman et al. (2005)	GIFT program	Uebelacker et al. (2003)	E	N/E	N/E	E
Learmonth et al. (2008)	Beat the Blues	Uebelacker et al. (2003)	N/E	N/E	N/E	N/E
Spek et al. (2007)	Internet-based CBT	Uebelacker et al. (2003)	E	N/E	E	E
Karavidas et al. (2007)	HRV biofeedback	Uebelacker et al. (2003)	N/E	N/E	N/E	N/E
CES-D						
MacKinnon et al. (2008)	Internet-based CBT	Bouma et al. (1995)	E	N/E	E	E
Gerrits et al. (2007)	Online interaction	Joseph et al. (1996)	E	N/E	E	E
Haringsma et al. (2006)	Coping class	Murrell et al. (1983)	N/E	N/E	N/E	N/E
Cuijpers et al. (2005)	Coping class	Bouma et al. (1995)	E	N/E	N/E	E
HRSD-17						
Bodenmann et al. (2003)	Interpersonal psychotherapy	Grundy et al. (1996)	E	N/E	N/E	E
Dimidjian et al. (2006)	Cognitive therapy	Uebelacker et al. (2003)	E	N/E	N/E	N/E
Hellerstein et al. (2008)	Adjunctive aripiprazole	Seggar et al. (2002)	N/E	N/E	N/E	N/E

Note: E = treated and normal comparison groups equivalent, N/E = treated and normal comparison groups not equivalent; CBT = Cognitive Behavioral Therapy; HRV = Heart Rate Variability.

Table 4
Normative comparison outcomes for the Kendall et al. and Cribbie & Arpin-Cribbie methods at an equivalence interval of 1 SD.

		Cribbie & Arpin-Cribbie	
		Equivalent	Not Equivalent
Kendall et al.	Equivalent	5	6
	Not Equivalent	0	4