Psychotherapy Research, 2015
Vol. 25, No. 2, 239–248, http://dx.doi.org/10.1080/10503307.2014.889329

EMPIRICAL PAPER

Assessing clinical significance using robust normative comparisons

HAYKAZ MANGARDICH & ROBERT A. CRIBBIE

Department of Psychology, York University, Toronto, Ontario, Canada

(Received 9 April 2013; revised 19 September 2013; accepted 24 January 2014)

Abstract
Objective: Clinical significance determines whether an intervention makes a real difference in the everyday life of a client. One of the most recommended approaches for conducting group-level analyses of clinical significance is to evaluate whether the treated clinical group is equivalent to a normal comparison group (normative comparisons). The purpose of this study was to demonstrate the analytical and practical power of assessing clinical significance using normative comparisons that are robust to violations of normality and homogeneity of variance assumptions. **Method:** Six datasets were gleaned from published intervention studies for depression. **Results:** We found that normative comparisons using a robust Schuirmann-Yuen test determined equivalency for 11% fewer clinical samples compared to original normative comparisons that use a Schuirmann test of equivalence. **Conclusions:** We recommend that researchers conducting normative comparisons utilize the Schuirmann-Yuen procedure as it provides the most reliable method available for determining if a treated clinical group is equivalent to a normative comparison group.

Keywords: clinical significance; intervention; depression; normative comparisons; normality; variance homogeneity; assumptions

When a clinician attempts to determine which therapy is appropriate for a particular client, he or she is likely to consider the clinical significance of that therapeutic intervention. Clinical significance, the measure of whether an intervention makes a real difference in the everyday life of the clients or others with whom the clients interact, represents an important advance in the treatment, prevention, and rehabilitation of mental health concerns (Kazdin, 1999). Its introduction by Jacobson, Follette, and Revenstorf (1984) was regarded as an important development in methodology (Lambert, Shapiro, & Bergin, 1986) and it has since become an expected statistic in published intervention studies by many journal editors (Bauer, Lambert, & Nielsen, 2004). There have been a number of different statistical approaches proposed to assess clinical significance, each with a different representation of what constitutes a "real difference" in a therapeutic outcome.

At the root of these different representations is the amount or degree of change clients experience from pre- to post-treatment. Prior to the introduction of clinical significance, this client change was assessed using only traditional analyses and findings that were statistically significant represented an efficacious intervention (Moleiro & Beutler, 2009). There has been, however, a growing recognition among clinical researchers that reliance on traditional inferential statistical analyses to evaluate treatment efficacy is problematic (Kazdin, 1999; Kendall & Sheldrick, 2000; Kraemer et al., 2003; Lunnen & Ogles, 1998) and that statistically significant differences between groups do not necessarily indicate practical, meaningful, or clinically significant differences between groups, nor for individuals within the groups (Ogles, Lunnen, & Bonesteel, 2001). Therefore, when investigators infer the efficacy of certain interventions by referencing statistically significant differences found between two group means following treatment, the ameliorative effect of the intervention is not established and may not be genuine. Thus, the purpose of the

Correspondence concerning this article should be addressed to Haykaz Mangardich, York University, Department of Psychology, Toronto, Ontario, Canada. Email: 12hm30@queensu.ca

current study is to provide a review of normative comparisons, i.e., group-level measures of clinical significance that focus on the equivalence of treated and normal comparison populations, and specifically to highlight an improved normative comparison procedure that is less susceptible to problems due to non-normality and unequal group variances. We begin with a brief introduction to clinical significance before introducing normative comparison procedures and demonstrating/comparing these procedures with real data from intervention studies for depression.

Determining Clinical Significance

The first method proposed for assessing clinical significance (Jacobson et al., 1984), with slight modifications by Jacobson and Truax (1991), suggested a two-step criterion to assess clinically significant change (also see Jacobson, Roberts, Berns, &. McGlinchey, 1999, for a discussion of potential modifications and alternatives to the method). First, a cutoff point is established that the client has to cross at the time of the post-treatment assessment to be classified as changed to a clinically significant degree. Next, a reliable change index (RCI) is calculated to determine how much change has occurred during the course of the therapy and whether this change is reliable and not due to measurement error.

Although the Jacobson and Truax (1991) method may be the most popular method for assessing clinical significance (Ogles et al., 2001), it is not the only way of conceptualizing the quantification of clinical significance. The Jacobson and Truax method determines whether there is individual change in functioning relative to a comparison group after treatment. Although in many cases researchers compute proportions of clients who improved, recovered, remained unchanged or deteriorated, or even conduct tests (e.g., chi-square goodness of fit test) comparing the proportions in each category, this procedure does not directly compare the treated group to a normal comparison group. Furthermore, the method assumes normal distributions when establishing cutoff points and the RCI. In their paper, Jacobson and Truax concede that such an assumption is a problem that limits the generalizability of their method, especially given decades of research demonstrating that parametric tests of the form of the RCI are not robust to violations of the assumption of normality (e.g., Cressie & Whitford, 1986) and that cutoffs relative to distributions that are not normal can be very misleading.

Clinical Significance using Normative Comparisons

A more recent approach that addresses the practical issue of whether the client returns to normal functioning after treatment is the normative comparisons method. In this approach, clinical significance is conceptually defined as end-state functioning that falls within a normative range on important measures (Kendall, Marrs-Garcia, Nath, & Sheldrick, 1999). In order to determine whether the treated group falls within a normative range, Kendall et al. (1999) suggest using a two-independent-samples test of equivalence (Schuirmann, 1987) on the normative and treated populations. As Kendall et al. (1999) mention, when using traditional hypothesis testing to demonstrate equivalence between groups, the investigator attempts to confirm rather than reject the null hypothesis of no population mean difference. In other words, if a researcher is interested in demonstrating that two groups are equivalent, it would be inappropriate to use a traditional difference-based test (e.g., traditional *t* or *F* test) because the goal would be to not reject the null hypothesis that the population means are equivalent. The first step of the normative comparison method requires a range to be established whereby the treated and the comparison group will be considered equivalent (−δ, δ). In the second step, one-sided *t* tests are conducted on two null hypotheses, H_{01} and H_{02}, where:

$$H_{01}: \mu_1 - \mu_2 \geq \delta;\ H_{02}: \mu_1 - \mu_2 \leq -\delta$$
$$H_{11}: \mu_1 - \mu_2 < \delta;\ H_{12}: \mu_1 - \mu_2 > \delta$$

H_1 represents the alternate hypothesis, μ represents the population mean, and δ represents the smallest difference between the groups that would be considered nontrivial. A significant result for both tests would result in the researcher concluding that the difference between the means lies within the predefined range and that the treated sample is clinically equivalent to the normative sample (Kendall et al., 1999).

An issue with this approach is that it is a parametric statistical test that assumes that the population variances are equal and the population distributions are normal in form. It is well known, however, that sample sizes and variances of the normative and clinical groups are regularly very disparate (Cribbie & Arpin-Cribbie, 2009). Furthermore, as Gruman, Cribbie, and Arpin-Cribbie (2007) show, empirical Type I error rates for Schuirmann's test of equivalence have been found to deviate substantially from the nominal α level when sample sizes and variances are unequal. The implications of these findings are that researchers may mistakenly declare populations equivalent or may mistakenly declare populations not equivalent. Kendall et al. (1999) were aware of this limitation; however, they failed to account for it as their approach uses the original Schuirmann procedure.

An Equivalence Test that is Robust to Heterogeneous Variances and Non-Normality

To account for the regularly occurring violations of the homogeneity of variance assumption, Gruman et al. (2007) expanded on Schuirmann's (1987) original equivalence testing approach by integrating Welch's (1938) heteroscedastic standard error and degrees of freedom into the original model. This modification deals with the sample size and variance inequality issues that affect the Schuirmann test of equivalence (which utilizes the same standard error and degrees of freedom as the independent-samples *t* test).

Gruman et al. (2007) found that Type I error rates for the Schuirmann-Welch test are maintained at approximately α even when sample sizes and variances are extremely unequal. There is also 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 (Cribbie & Arpin-Cribbie, 2009). Therefore, due to its robustness when the sample variances are heterogeneous, it was recommended that researchers evaluating clinical significance via equivalence testing routinely utilize the Schuirmann-Welch procedure.

Although the Schuirmann-Welch procedure addresses the heterogeneity of variance issue, there still are limitations in its application in normative comparisons. Classic parametric methods, such as the Schuirmann-Welch procedure, assume that the data being analyzed are normally distributed. Much research has been conducted to this end and the general consensus is that this assumption is rarely met in real data (Erceg-Hurn & Mirosevich, 2008; Golinski & Cribbie, 2009). For example, Micceri (1989) examined 440 large data sets from psychological and educational studies and concluded that none of the data were normally distributed and that only a few distributions remotely resembled the normal curve. Furthermore, it is well known that the usual group means and variances are greatly influenced by the presence of extreme observations in score distributions (Keselman, Wilcox, & Lix, 2003). Extreme observations or heavy tails can significantly increase the standard error of the mean. Consequently, using measures that are not robust to non-normality and outliers, according to Wilcox (1994), can distort the view of how a typical individual in one group compares to a typical individual in another, how well Type I errors are controlled, and the power of the test. To account for these problems, a popular approach has been to remove the atypical values by trimming the data (Wilcox, 2012). Although a wide range of robust estimators have been proposed in the literature, the trimmed mean and Winsorized variance, which are calculated by removing extreme observations, have been most appealing (Keselman et al., 2003; Wilcox, 1994, 2012). As Wilcox (2012) and Keselman, Othman, Wilcox, and Fradette (2004) establish, the Type I error rates and power to detect effects are much less affected by extremities and non-normality when trimmed means are substituted for the usual means and variances. In conjunction with trimmed means and Winsorized variances, Yuen (1974) suggested that Welch's (1938) statistic be used. In her study, Yuen found that the statistic based on trimmed means and Winsorized variances could adequately control the rate of Type I errors and resulted in greater power than a statistic based on the usual mean and variance for heavily tailed distributions (Keselman et al., 2003).

The Current Study

van Wieringen and Cribbie (in press) used the Schuirmann-Yuen method and compared its effectiveness to the Schuirmann and Schuirmann-Welch tests of equivalence using a series of Monte Carlo simulations with different distribution shapes, sample standard deviations, and sample sizes. Their results indicate that when two groups with non-normal (and potentially different) distribution shapes are being compared, the empirical Type I error control of the Schuirmann-Yuen is substantially better than that of the original Schuirmann or the Schuirmann-Welch. The power rates also differed between these tests, with rates generally 10% to 30% higher for the Schuirmann-Yuen than for the Schuirmann or Schuirmann-Welch procedures for non-normal distributions. For example, they reported that when sample sizes were large and one distribution was skewed and one distribution contained outliers in one tail, that power was recorded at .204, .293, and .509 for the Schuirmann, the Schuirmann-Welch, and the Schuirmann-Yuen, respectively. While the effectiveness of the Schuirmann-Yuen under conditions that violate the assumptions of homogeneity of variance and normal distributions has been previously established, the present study attempts to extend its application to real data gathered from a sample of published studies that examined interventions for depression. This extension doubly augments current literature and further allows one to capture why these robust measures of clinical significance are so valuable when attempting to discern the therapeutic value of interventions. Firstly, given the possibility for disparate conclusions to arise when assessing clinical significance as opposed to statistical significance, the value of evaluating this possibility using data from previously published studies is pronounced since differing conclusions could impact whether a given intervention is purported to be efficacious or not. Secondly, using real data allows us to screen for violations of parametric assumptions to determine whether factors such as non-normality or variance heterogeneity could contribute to any different findings when comparing conclusions on the efficacy of treatments using Schuirmann-Yuen, Schuirmann-Welch, and Schuirmann tests of equivalence. The focus of this study, therefore, is to demonstrate the analytical and practical power of robust normative comparison methods for making statements regarding the efficacy of therapies for depression.

Method

In order to quantify clinical significance with the Schuirmann-Yuen normative comparison method, we collected a sample of published studies that tested the effectiveness of interventions for depression. Although there are a number of different alternatives that could have been considered, depression was selected because of its prevalence and the fact that a vast amount of research has been published on the effectiveness of existing therapies (e.g., Driessen et al., 2007; Gaynor et al., 2003; Nasiakos, Cribbie, & Arpin-Cribbie, 2010). It is important to point out though, as highlighted by a reviewer of the paper, that different depression scales are sensitive to different forms, symptoms, and severities of depression and therefore caution is needed in interpreting the results.

Collection of Past Studies

A specific set of search criteria was used in PsycINFO to collect studies of potential interest: "Effectiveness," "efficacy," "effect," or "evaluation" in the document title, "intervention," "therapy," or "treatment" in the document title, "depression" in the document title, and "psychotherapy," "cognitive therapy," or "behavioral therapy" in any other field. These search criteria were used in an effort to identify studies that evaluated the effectiveness of interventions for depression. The date range of our search was restricted from 2000 to 2011 to ensure that the data we gathered from the studies were current. Of the original studies found, an inclusion criterion for further consideration was the availability of pre- and post-treatment data for clinical groups on indicators of depressive symptomatology. The principal investigators of studies that met these criteria were then sent a request, by e-mail, for access to the raw data from the intervention study. It is important to highlight that the Schuirmann-Yuen cannot be computed with summary descriptive information (e.g., means and standard deviations), and thus in order to compute the procedure we needed to collect raw data. A total of 99 authors were contacted, 40 of whom initially responded. Only two of these authors, however, provided us with their datasets and assented to using them for our analyses. A follow-up email was then sent to the 59 authors who had not responded, yielding four more pertinent datasets. A summary of each of the corresponding studies for the six datasets is provided in Table I.

Normative Comparison Data

Normative data were gathered for the outcome measures utilized in the published studies we collected. As Nasiakos et al. (2010) mention, in order to determine clinical significance, treatment groups

Table I. Description of the intervention studies.

Study	Description
Goldman et al. (2006)	The effectiveness of client-centered therapy was compared to emotion-focused therapy in treating depressive symptoms. Pre-post measures were taken using the BDI.
Topolovec-Vranic et al. (2010)	The effectiveness of an internet-based cognitive behavior program (Mood GYM) in treating depressive symptoms in clients with a traumatic brain injury was evaluated. Pre-post measures were taken using the CES-D.
Henkel et al. (2010)	This study examines the efficacy of sertraline and cognitive behavioral therapy (CBT) among depressed patients with atypical features. Pre-post measures were taken using the IDS and HAM-D.
Stice et al. (2010)	High-risk adolescents with elevated depressive symptoms were randomized to a brief group cognitive-behavioral (CB) intervention, group supportive-expressive intervention, bibliotherapy, or assessment-only control conditions. Pre-post measures were taken using the BDI.
Mohr et al. (2005)	This study tested the efficacy of a 16-week telephone-administered cognitive-behavioral therapy (T-CBT) against a strong control for attention and nonspecific therapy effects.
Koch et al. (2007)	The study investigated the specific effects of a dance intervention on the decrease of depression symptoms. Pre-post measures were taken using the HBS.

must be compared with a normative group and for a most appropriate comparison, it is advised that either representative normative data be collected by the researchers or that normative data matching important characteristics of the sample be gleaned from published sources (Kendall & Sheldrick, 2000). Kendall and Sheldrick list a set of studies for different outcome measures for depression and note that if any of the identified outcome measures in their paper are analyzed in a given treatment outcome study, the respective sources of normative data can be found in their paper and be directly applied in normative comparisons. For our purposes, it would be impractical to collect normative data that matched the sample characteristics of the different treatment groups in each of the six datasets (e.g., traumatic brain injury patients). Furthermore, it was necessary for our analyses to collect complete normative data (i.e., not just means, standard deviations, sample sizes, etc. that are reported in most papers and in review articles such as Kendall and Sheldrick), as we needed the individual scores for each participant on the outcome measures in order to compute the trimmed means and Winsorized variances. Normative data were gathered from a sample of 184 undergraduate university students (29.9% male) aged 17 to 33 years old (mean age was 19.5). These participants were enrolled in introductory psychology courses and took part in the study in exchange for class credit. The sample varied vastly in self-identified race/ethnicity, being composed of 36% white, 7% black, 15% East and/or South-East Asian, 4% Hispanic, 11% Middle Eastern, 22% South Asian and 5% mixed. Although we freely acknowledge that our normative sample is not a perfect match to each of the populations in the empirical studies obtained, as demonstrated below our normative sample was found to be equivalent to published general population normative data and thus we believe that our comparisons are meaningful.

Statistical Analyses

The Schuirmann-Yuen (van Wieringen & Cribbie, in press), Schuirmann-Welch, and Schuirmann (Cribbie & Arpin-Cribbie, 2009) equivalence tests for normative comparisons were modeled with the R statistical software application (R Foundation for Statistical Computing, 2011). For our normative comparison analyses, we defined δ in the equivalence interval (δ, δ) as one standard deviation of the normative mean. Although this value was somewhat arbitrarily selected, one standard deviation has been used in the past in normative comparison examples (e.g., Kendall et al., 1999). We also included comparisons to the traditional paired-samples *t*-tests comparing pre- and post-test scores and an analysis using the Jacobson and Truax (1991) method for assessing clinical significance.

Results

All of the empirical studies tested the effectiveness of therapies for depression. Some, however, included multiple groups, tested the effectiveness of multiple therapies, and had multiple outcome variables. For the purposes of our study, we used the post-test data for all non-control groups on the outcome measure of depression. Therefore, a total of 18 clinical samples were included in our analyses.

Normative Data

Table II presents a summary of the normative data we gathered in order to conduct the normative comparisons. Normative data were collected for the five measures of depression used in the six studies (the order of appearance was randomized when participants completed the scales): Center for Epidemiological Studies – Depression Scale (CES-D), Inventory of Depressive Symptomatology (IDS), Hamilton Depression Scale (HAM-D), Heidelberger Befindlichkeitsskala (HBS), and the Beck Depression Inventory (BDI). Table II also presents published normative data for these depression measures for comparison purposes (although we were not able to locate published normative data for the HBS). It is important to note that the psychometric properties (e.g., reliability, validity) differ across the set of outcome measures that were used. Five different Schuirmann-Welch equivalence tests showed that the scores on the outcome measures of depression for our sample of undergraduate students were equivalent to those found in sources of normative data from previously published studies (Campo-Arias, Diaz-Martinez, Rueda-Jaimes, Cadena-Afanador, & Her-nandez, 2007; Crawford, Cayley, Lovibond, Wilson, & Hartley, 2011; Gonzalez, Boals, Jenkins, Schuler, & Taylor, 2013).

Table II. Summary of the normative data used in the normative comparison analyses (N = 184) and published normative data.

	Normative data from this study			Published normative data	
Scale	*M*	*SD*	Shapiro-Wilk *p*	*M*	*SD*
CES-D	15.66	10.37	.000	17.16	9.88
IDS	15.99	12.35	.000	15.03	11.18
HAM-D	7.88	7.17	.000	6.25	4.24
BDI	9.58	8.75	.000	6.25	6.94
HBS	3.31	1.96	.000	N/A	N/A

Note. N/A indicates that the data were not available.

Table III. Summary of the depression intervention studies obtained for analysis.

Study	Primary outcome	Intervention groups	Pre – Post *N*	Pre – Post *M*	Pre – Post *SD*
Goldman et al. (2006)	BDI	Process-experiential	36 – 36	26.11 – 6.17	6.96 – 5.33
		Client-centered	36 – 36	24.56 – 9.53	6.54 – 7.48
Topolovec-Vranic et al. (2010)	CES-D	Online cognitive behavioral therapy program	13 – 13	30.69 – 23.38	8.24 – 11.24
Henkel et al. (2010)	HAM-D	Sertraline	22 – 22	17.18 – 11.59	4.25 – 8.14
		CBT	22 – 22	16.50 – 11.23	4.11 – 6.39
		GSG	26 – 26	14.46 – 14.15	3.82 – 5.72
	IDS	Sertraline	22 – 22	28.55 – 18.27	7.61 – 13.05
		CBT	22 – 22	27.41 – 18.18	5.79 – 10.42
		GSG	26 – 26	24.50 – 23.00	6.15 – 10.67
Stice et al. (2008)	BDI	CB intervention	89 – 86	20.24 – 10.89	10.41 – 9.17
		Group supportive-expressive intervention	88 – 83	20.23 – 14.54	9.84 – 10.74
		Bibliotherapy	80 – 74	17.92 – 14.38	7.14 – 9.00
Mohr et al. (2005)	BDI	Experiential	62 – 62	27.39 – 18.48	7.06 – 10.28
		CBT	60 – 60	27.87 – 15.00	8.71 – 10.84
	HAM-D	Experiential	62 – 62	21.44 – 14.52	3.63 – 6.88
		CBT	60 – 60	21.45 – 12.28	3.77 – 5.73
Koch et al. (2007)	HBS	Dance-only	11 – 11	5.18 – 6.41	1.40 – 1.20
		Music-only	10 – 10	6.25 – 6.00	1.83 – 2.40

Note. CBT = cognitive behavioral therapy; GSG = guided self-help group.

Normality and Variance Homogeneity Assumptions

Regarding the assumption of normality within our normative data, all of the distributions of the outcome measures were non-normal (all *p*s < .001), using the Shapiro-Wilk test of normality (Table II). Table III presents the clinical measures of depression used within each study and also provides the pre- and post-test descriptive data for each clinical group. Specifically, the table provides pre- and post-test sample sizes, means and standard deviations. Of the 18 clinical groups, four pre-treatment distributions and 10 post-test distributions were determined to be non-normal using the Shapiro-Wilk test, with all being extremely skewed. Bartlett's test of homogeneity of variance was conducted to compare the variances of the normative and post-test clinical groups, and of the 18 comparisons six had heterogeneous variances (see Table IV). It is also important to highlight that the sample sizes (N = 10 to N = 89) in these studies were small to moderate, and thus statistically significant tests of normality and

Table IV. Summary of the variance homogeneity and normal distribution assumptions.

Study	Primary outcome	Intervention groups	Bartlett's *p*	Pre – Post Shapiro-Wilk *p*
Goldman et al. (2006)	BDI	Process-experiential	0.0007328	.008 – .005
		Client-centered	0.2462	.205 – .002
Topolovec-Vranic et al. (2010)	CES-D	Online cognitive behavioral therapy program	0.6954	.529 – .097
Henkel et al. (2010)	HAM-D	Sertraline	0.4234	.474 – .217
		CBT	0.4958	.150 – .171
		GSG	0.1586	.580 – .489
	IDS	Sertraline	0.733	.342 – .212
		CBT	0.3535	.645 – .606
		GSG	0.05627	.913 – .861
Stice et al. (2008)	BDI	CB intervention	0.6536	.014 – .000
		Group supportive-expressive intervention	0.02681	.389 – .001
		Bibliotherapy	0.7345	.612 – .019
Mohr et al. (2005)	BDI	Experiential	0.1185	.148 – .044
		CBT	0.03844	.001 – .013
	HAM-D	Experiential	0.6957	.146 – .261
		CBT	0.04303	.243 – .007
Koch et al. (2007)	HBS	Dance-only	0.06813	.092 – .436
		Music-only	0.374	.349 – .063

Note. CBT = cognitive behavioral therapy; GSG = guided self-help group.

Table V. Summary of the equivalence-based clinical significance analyses, pre-post differences, and pre-normative differences for the intervention studies.

			Normative comparison				
Study	Primary outcome	Intervention groups	Schuirmann and Schuirmann-Welch	Schuirmann-Yuen	ΔPre-post $p < .05$	$\eta^2_{\text{pre-post}}$	ΔPre-normative $p < .05$
Goldman et al. (2006)	BDI	Process-experiential	E	E	Yes	0.85	Yes
		Client-centered	E	E	Yes	0.80	Yes
Topolovec-Vranic et al. (2010)	CES-D	Online CBT	NE	NE	Yes	0.31	Yes
Henkel et al. (2010)	HAM-D	Sertraline	E	NE	Yes	0.33	Yes
		CBT	E	E	Yes	0.33	Yes
		GSG	NE	NE	No	0.003	Yes
	IDS	Sertraline	E	E	Yes	0.40	Yes
		CBT	E	E	Yes	0.44	Yes
		GSG	E	NE	No	0.02	Yes
Stice et al. (2008)	BDI	CBT	E	E	Yes	0.40	Yes
		Group supportive-expressive intervention	E	E	Yes	0.27	Yes
		Bibliotherapy	E	E	Yes	0.18	Yes
Mohr et al. (2005)	BDI	Experiential	NE	NE	Yes	0.30	Yes
		CBT	E	E	Yes	0.44	Yes
	HAM-D	Experiential	NE	NE	Yes	0.45	Yes
		CBT	E	E	Yes	0.67	Yes
Koch et al. (2007)	HBS	Dance-only	NE	NE	Yes	0.71	Yes
		Music-only	NE	NE	No	0.07	No

Note. CBT = cognitive behavioral therapy; GSG = guided self-help group; E = equivalent; NE = not equivalent.

variance heterogeneity are unlikely to be due to inflated sample sizes. Further, and more likely, is that the non-significant tests were due to low power for detecting assumption violations.

Clinical Significance Tests

Before presenting the results of the normative comparisons, we first provide results using a traditional paired *t*-test to compare the pre- and post-test

Table VI. Summary of the Jacobson and Truax (1991) clinical significance analyses.

Study	Primary outcome	Intervention groups	Percentage of individuals recovered
Goldman et al. (2006)	BDI	Process-experiential	86%
		Client-centered	56%
Topolovec-Vranic et al. (2010)	CES-D	Online cognitive behavioral therapy program	8%
Henkel et al. (2010)	HAM-D	Sertraline	50%
		CBT	36%
		GSG	14%
	IDS	Sertraline	27%
		CBT	23%
		GSG	4%
Stice et al. (2008)	BDI	CB intervention	31%
		Group supportive-expressive intervention	17%
		Bibliotherapy	9%
Mohr et al. (2005)	BDI	Experiential	39%
		CBT	45%
	HAM-D	Experiential	45%
		CBT	65%
Koch et al. (2007)	HBS	Dance-only	N/A
		Music-only	N/A

Note. CBT = cognitive behavioral therapy; GSG = guided self-help group; N/A indicates that we were unable to calculate this. This is due to a lack of test-retest reliability for the HBS.

means and also the results using the Jabobson and Truax (1991) method to determine the proportion of clients who reliably improved and moved beyond two standard deviations from the clinical mean (see Tables V and VI). The paired *t* results indicated that 15 clinical samples showed statistically significant improvement from pre-test to post-test, although we caution the reader that these tests were simple paired comparisons that were not in reference to a control group and did not control for any other variables. Eta squared was also calculated for each of the pre-test to post-test differences as a measure of effect size. The results of the Jacobson and Truax (1991) analyses indicated that there was substantial variability in the percentages of clients who reliably improved and moved outside the range of the clinical population.

Table V summarizes the results from the normative comparisons using the Schuirmann, Schuirmann-Welch, and Schuirmann-Yuen equivalence tests. As outlined by Cribbie and Aprin-Cribbie (2009), the first step of their normative comparison method requires comparing the pre-test and normative means. This comparison was conducted using the heteroscedastic trimmed Welch procedure (Yuen, 1974) since the distributions were often not normal and the variances often differed substantially. As expected, all of the 18 clinical samples had pre-test means that differed significantly from the normative means (see Table V). The normative comparison results for the 18 clinical samples show that 12 exhibited equivalence to the normal comparison group using both the Schuirmann and Schuirmann-Welch methods at ± 1.0 *SD* interval. These results differed from normative comparisons using the Schuirmann-Yuen method (at the same equivalence interval), where only 10 clinical samples were considered equivalent and so represented clinically significant change. Although the normative comparison, paired *t*, and Jacobson and Truax results were definitely related (e.g., for those declared equivalent to the normal comparison group the average percentage of clients meeting Jacobson and Truax criteria was 40%, and for those not declared equivalent to the normal comparison group the average percentage was 27%), they also provide distinct information about the clinical significance of the treatment effects.

Discussion

Measures of clinically significant improvement for groups of patients influence the degree to which treatments are generally considered to be effective or in need of modification (Bauer et al., 2004). Because of the increasing value psychologists place on evidence-based therapeutic methods (e.g., Norcross, Beutler, & Levant, 2005), there have been many methods proposed to assess clinical significance. A recent approach that has become the recommended method for assessing group-level clinical significance is Kendall et al.'s (1999) normative comparison method that uses the Schuirmann (1987) test of equivalence. Two important issues with the method, however, are that it assumes normal distributions and homogeneity of variances between the clinical and normative groups. Ignoring these assumptions and utilizing tests that are not robust to these assumptions can result in empirical conclusions that are inaccurate; depending on the nature of the distributions populations can mistakenly be declared equivalent or not equivalent. In this study, consistent with past research (e.g., Micceri, 1989), most of the clinical samples (61%) had non-normal distributions at post-test and none of the normative distributions were normally distributed (more specifically, they were all extremely skewed). Six of the 18 clinical samples also had variances significantly different from the normative sample. Given that van Wieringen and Cribbie (in press) found that only the Schuirmann-Yuen was robust to simultaneous violations of the normality and variance homogeneity assumptions, these results definitely provide support for the recommendation that researchers should always use the Schuirmann-Yuen test for normative comparisons.

This study extended the findings of van Wieringen and Cribbie (in press) to real data in order to provide information about the clinical significance of the interventions and also to test whether differences in the conclusions would be found when using the Schuirmann-Yuen test compared to the Schuirmann and Schuirmann-Welch tests. We found that, of the 18 clinical samples, the Schuirmann-Yuen test determined equivalence for about 11% fewer clinical samples than the Schuirmann and Schuirmann-Welch tests with an equivalence interval of ± 1.0 *SD*. The value of these intervention studies lies in their ability to accurately distinguish therapies that are efficacious for the everyday functioning of the client. Even though only a handful of clinical samples were analyzed, the results of van Wieringen and Cribbie and the current study highlight the importance of using an appropriate test statistic, one that overcomes violations of parametric assumptions in order to ensure that the conclusions are accurate.

The results of this study also illustrate why determining clinical significance has important implications when considering which intervention to use. For example, 15 of the 18 clinical samples had statistically significant changes from pre- to post-test, whereas normative comparisons using a Schuirmann-Yuen test determined equivalency for 10 clinical samples (however, it is important to highlight again that these analyses were conducted with paired-samples *t*-tests and did not include the control groups, covariates or other predictors that may have been present in the full models analyzed in the published papers). Therefore, although a given intervention may render a statistically significant difference from pre- to post-test data for the treatment group, the intervention does not necessarily return the group to a state of normal functioning. As Nasiakos et al. (2010) mention, methods that determine clinical significance, such as normative comparisons, provide supplementary information regarding whether the treated group returned to a state of "normal" functioning, a common goal of intervention studies for depression. By conducting robust normative comparisons on real data, this study demonstrates that different conclusions indeed do arise when clinical significance tests are conducted as opposed to statistical significance tests.

Although the findings of this study are important for showing the value of applying robust measures of clinical significance, there are a couple of limitations of the current research to consider. Firstly, intervention studies for depression were chosen for conducting our robust clinical significance analyses, therefore it is unclear how our results would compare to results of the analyses when applied to other outcomes such as anxiety. Furthermore, we were only able to gather a handful of the numerous published intervention studies for depression so the conclusions of this study may not necessarily extend to the population of studies. However, the results of this study mirror previous quantitative studies, as discussed above, that have highlighted that distributions are often non-normal, population variances are often not equal, and that tests based on robust estimators outperform tests based on standard estimators such as the usual mean and variance. Lastly, the results of this paper were summarized across different measures of depression that may deviate in the extent to which they are appropriate for undergraduate students (i.e., our normal comparison group). Although we have shown that the central tendencies of the undergraduate and general populations are equivalent, the differences that remain could be meaningful.

To conclude, this study collected data from published intervention studies that tested the effectiveness of therapies for depression and analyzed the data using the Schuirmann, Schuirmann-Welch, and Schuirmann-Yuen equivalence-based normative comparison methods. We found that most clinical samples had violations of the homogeneity of variance and/or normality assumptions, and therefore we recommend that researchers conducting normative comparisons utilize the Schuirmann-Yuen procedure as it provides the most reliable method available for determining whether a treated clinical group is equivalent to a normative comparison group. An R (The R foundation for Statistical Computing, 2011) function for conducting the normative comparison tests discussed in this article is available at http://www.psych.yorku.ca/cribbie.

References

Bauer, S., Lambert, M. J., & Nielsen, S. L. (2004). Clinical significance methods: A comparison of statistical techniques. *Journal of Personality Assessment*, *82*, 60–70. doi:10.1207/s15327752jpa8201_11

Campo-Arias, A., Diaz-Martinez, L. A., Rueda-Jaimes, G. E., Cadena-Afanador, L. P., & Her-nandez, N. L. (2007). Psychometric properties of the CES-D scale among Colombian adults from the general population. *Revista Colombiana de Psiquiatria*, *36*(4), 664–674.

Crawford, J. R., Cayley, C., Lovibond, P. F., Wilson, P. H., & Hartley, C. (2011). Percentile norms and accompanying interval estimates from an Australian general adult population sample for self-report mood scales. *Australian Psychology*, *46*, 3–14. doi:10.1111/j.1742-9544.2010.00003.x

Cressie, N. A. C., & Whitford, H. J. (1986). How to use the two sample *t*-test. *Biometrical Journal*, *28*, 131–148. doi:10.1002/bimj.4710280202

Cribbie, R. A., & Arpin-Cribbie, C. A. (2009). Evaluating clinical significance through equivalence testing: Extending the normative comparisons approach. *Psychotherapy Research*, *19*, 677–686. doi:10.1080/10503300902926554

Driessen, E., Henricus, L., van,Schoevers, R. A., Cuijpers, P., van Aalst, G., Don, F. J., … Dekker, J. M. J. (2007). Cognitive behavioral therapy versus short psychodynamic supportive psychotherapy in the outpatient treatment of depression: A randomized controlled trial. *BMC Psychiatry*, *7*, 58. doi:10.1186/1471-244X-7-58

Erceg-Hurn, D. M., & Mirosevich, V. M. (2008). Modern robust statistical methods: an easy way to maximize the accuracy and power of your research. *American Psychologist*, *63*, 591–601. doi:10.1037/0003-066X.63.7.591

Gaynor, T. S., Weersing, R. V., Kolko, D. J., Birmaher, B., Heo, J., & Brent, A. D. (2003). Adolescent therapy for depression: a comparison across cognitive-behavioral, family, and supportive therapy. *Journal of Consulting and Clinical Psychology*, *71*(2), 386–393. doi:10.1037/0022-006X.71.2.386

Goldman, N. R., Greenberg, S. L., & Angus, L. (2006). The effects of adding emotion-focused interventions to the client-centered relationship conditions in the treatment of depression. *Psychotherapy Research*, *16*, 537–549. doi:10.1080/10503300600589456

Golinski, C., & Cribbie, R. A. (2009). The expanding role of quantitative methodologists in advancing psychology. *Canadian Psychology*, *50*, 83–90. doi:10.1037/a0015180

Gonzalez, D. A., Boals, A., Jenkins, S. R., Schuler, E. R., & Taylor, D. (2013). Psychometrics and latent structure of the IDS and QIDS with young adult students. *Journal of Affective Disorders*, *149*, 217–220. doi:10.1016/j.jad.2013.01.027

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

Henkel, V., Mergl, R., Allgaier, A. K., Hautzinger, M., Kohnen, R., Coyne, J. C., Moller, H. J., & Hegerl, U. (2010).

Treatment of atypical depression: Post-hoc analysis of a randomized controlled study testing the efficacy of sertraline and cognitive behavioral therapy in mildly depressed outpatients. *European Psychiatry*, *25*(8), 491–498. doi:10.1016/j.eurpsy.2010.01.010

Jacobson, N. S., Follette, W. C., & Revenstorf, D. (1984). Physiotherapy outcome research: Methods for reporting variability and evaluating clinical significance. *Behavior Therapy*, *15*, 336–352. doi:10.1016/S0005-7894(84)80002-7

Jacobson, N. S., Roberts, L. J., Berns, S. B., &. McGlinchey, J. B. (1999). Methods for defining and determining the clinical significance of treatment effects: Description, application, and alternatives. *Journal of Consulting and Clinical Psychology*, *67*, 300–307. doi:10.1037/0022-006X.67.3.300

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, *59*, 12–19. doi:10.1037/0022-006X.59.1.12

Kazdin, A. E. (1999). The meanings and measurement of clinical significance. *Journal of Consulting and Clinical Psychology*, *67*, 332–339. doi:10.1037/0022-006X.67.3.332

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. doi:10.1037/0022-006X.67.3.285

Kendall, P. C., & Sheldrick, R. C. (2000). Normative data for normative comparisons. *Journal of Consulting and Clinical Psychology*, *68*, 767–773. doi:10.1037/0022-006X.68.5.767

Keselman, H. J., Othman, A. R., Wilcox, R. R., & Fradette, K. (2004). The new and improved two-sample *t* test. *Psychological Science*, *15*, 47–51. doi:10.1111/j.0963-7214.2004.01501008.x

Keselman, H. J., Wilcox, R. R., & Lix, L. M. (2003). A generally robust approach to hypothesis testing in independent and correlated group designs. *Psychophysiology*, *40*, 586–596. doi:10.1111/1469-8986.00060

Koch, C. S., Morlinghaus, K., & Fuchs, T. (2007). The joy dance: Specific effects of a single dance intervention on psychiatric patients with depression. *The Arts in Psychotherapy*, *34*, 340–349. doi:10.1016/j.aip.2007.07.001

Kraemer, H. C., Morgan, G. A., Leech, N. L., Gliner, J. A., Vaske, J. J., & Harmon, R. J. (2003). Measures of clinical significance. *Journal of the American Academy of Child & Adolescent Psychiatry*, *42*, 1524–1529. doi:10.1097/00004583-200312000-00022

Lambert, M. J., Shapiro, D. A., & Bergin, A. E. (1986). The effectiveness of psychotherapy. In S. L. Garfield & A. E. Bergin (Eds.), *Handbook of psychotherapy and behavior change* (3rd ed., pp. 157–211). New York: Wiley.

Lunnen, K. M., & Ogles, B. M. (1998). A multiperspective, multivariable evaluation of reliable change. *Journal of Consulting and Clinical Psychology*, *66*, 400–410. doi:10.1037/0022-006X.66.2.400

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*, 156–166. doi:10.1037/0033-2909.105.1.156

Mohr, C. D., Hart, L. S., Julian, L., Catledge, C., Honos-Webb, L., Vella, L., & Tasch, T. E. (2005). Telephone-administered psychotherapy for depression. *Archives of General Psychiatry*, *62* (9), 1007–1014. doi:10.1001/archpsyc.62.9.1007

Moleiro, C., & Beutler, L. E. (2009). Clinically significant change in psychotherapy for depressive disorders. *Journal of Affective Disorders*, *115*, 220–224. doi:10.1016/j.jad.2008.09.009

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. doi:10.1080/10503307.2010.501039

Norcross, J. C., Beutler, L. E., & Levant, R. F. (Eds.). (2005). *Evidence-based practices in mental health: Debate and dialogue on the fundamental questions*. Washington DC: American Psychological Association.

Ogles, B. M., Lunnen, K. M., & Bonesteel, K. (2001). Clinical significance: History, application, and current practice. *Clinical Psychology Review*, *21*, 421–446. doi:10.1016/S0272-7358(99)00058-6

Schuirmann, D. J. (1987). A comparison of the two-sided tests procedure and the power approach for assigning equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, *15*, 657–680. doi:10.1007/BF01068419

Stice, E., Rohde, P., Seeley, J. R., & Gau, J. M. (2008). Brief cognitive-behavioral depression prevention program for high-risk adolescents outperforms two alternative interventions: A randomized efficacy trial. *Journal of Consulting and Clinical Psychology*, *76*(4), 595–606. doi:10.1037/a0012645

Topolovec-Vranic, J., Cullen, N., Michalak, A., Ouchterlony, D., Bhalerao, S., Masanic, C., & Cusimano, M. D. (2010). Evaluation of an online cognitive behavioral therapy program by patients with traumatic brain injury and depression. *Brain Injury*, *24*(5), 762–772. doi:10.3109/02699051003709599

van Wieringen, K., & Cribbie, R. A. (in press). An improved test of equivalence for conducting normative comparisons.

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

Wilcox, R. R. (1994). A one-way random effects model for trimmed means. *Psychometrika*, *59*, 289–306. doi:10.1007/BF02296126

Wilcox, R. R. (2012). *Introduction to robust estimation and hypothesis testing* (3rd ed.). San Diego, CA: Academic Press.

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