Curr Psychol (2015) 34:672–680
DOI 10.1007/s12144-014-9280-z

Group Level Clinical Significance: An Analysis of Current Practice

Robert A. Cribbie[1] · **Chantal A. Arpin-Cribbie**[2] · **Rebecca Vendittelli**[1] · **Erica Tucciarone**[1]

Published online: 14 October 2014

Abstract Measures of clinical significance offer important information about psychological interventions that cannot be garnered from tests of the statistical significance of the change from pretest to posttest. For example, post-intervention comparisons to a nonclinical group often offer valuable information about the practical value of the change that occurred. This study explored the manner in which researchers conduct clinical significance analyses in an effort to summarize the effectiveness of an intervention at the group level. The focus was on the use of the original Jacobson and Truax (*Journal of Consulting and Clinical Psychology, 59*, 12–19, 1991) method and the normative comparisons method due to Kendall et al. (*Journal of Consulting and Clinical Psychology, 67*, 285–299, 1999). The results highlight that although the Jacobson and Truax method is routinely adopted for summarizing group-level clinical significance, advanced strategies for summarizing the results are very infrequently applied. Further, the Kendall et al. method, which provides valuable and distinct information regarding how the treated group is performing relative to a normal comparison group, is rarely adopted and even when it is it is often not conducted appropriately. Recommendations are provided for conducting group-level clinical significance analyses.

Keywords Clinical significance · Normative comparisons · Equivalence testing · Jacobson and Truax

✉ Robert A. Cribbie
cribbie@yorku.ca

[1] Quantitative Methods Program, Department of Psychology, York University, Toronto, Ontario, Canada

[2] Department of Psychology, Laurentian University, Barrie, Ontario, Canada

The past few decades have seen an upward trend in the importance placed on understanding the practical value of an intervention, leading to significant improvements in the appreciation of the relative benefit of different forms of intervention. The practical value of an intervention, namely its ability to reduce the symptoms being targeted and thus improve the overall quality of life of the client, has been labeled clinical significance (Jacobson, Follette, and Revenstorf 1984). The use of measures of clinical significance has increased dramatically and has greatly improved the ability of researchers to evaluate and compare treatments. For example, the Jacobson and Truax (1991) method (JT) of evaluating clinical significance, where the interest is in determining whether the treated clients have experienced meaningful change and whether this change has improved their standing relative to a clinical or normal comparison group, has become very popular and provides valuable information that cannot be obtained from a global statistical test of pre-treatment to post-treatment mean change (even if the change is evaluated relative to a control group) (Bauer, Lambert, and Nielsen 2004). A noteworthy distinction in terms of the use of measures of clinical significance is whether the interest is in evaluating clinical significance at the individual level (i.e., determining whether the intervention has had an important effect on each client separately) or at the group level (i.e., dealing with the treated group as a whole, and asking whether the intervention has had an important effect).

Many of the popular measures of clinical significance (e.g., JT) initially assess clients at the individual level, although group-level summaries of the results are a natural extension of individual level information. Thus, although the intended goal was to see the different ways in which individual subjects responded to the treatment, readers of clinical research studies, who normally have no relationship to the clients in the study, are rarely interested in individual measures of clinical

significance and instead are interested in some form of summary of the individual-level statistics. These summary statistics can be obtained from the individual level clinical significance statistics (e.g., proportions of clients who improved) or can be derived from methods using the group level data (e.g., posttest means of the intervention group). An example of a procedure that uses the group level data is the normative comparisons approach (NC) proposed by Kendall, Marrs-Garcia, Nath and Sheldrick (1999) that compares posttest means to normative sample means.

The purpose of this paper is to investigate the different ways in which clinical researchers are summarizing group-level clinical significance statistics. More specifically, this study explores how researchers use the JT and NC methods to make summary statements regarding the effectiveness of an intervention. The JT method was selected because it is not only the most popular, but also the most recommended, approach for conducting clinical significance statistics (e.g., Atkins, Bedics, McGlinchey, and Beauchaine 2005; Maassen 2000). The NC method was selected because it provides a theoretically distinct approach to assessing group-level clinical significance; instead of starting with individual level clinical significance information the procedure directly compares the treated group to a normal comparison group. To begin, an introduction to the JT and NC methods will be provided. Second, a review of treatment studies that have adopted the JT or NC methods will be conducted in order to better understand how these methods are being used to summarize the clinical significance of interventions. Lastly, the results of the review will be discussed in conjunction with recommendations for quantifying group level clinical significance.

Jacobson and Truax Method of Assessing Clinical Significance

Jacobson and Truax (1991) define clinically significant change as change that brings the client's level of functioning closer to the 'functional' population. In order to quantify how the level of functioning can be closer to normal, JT cite three potential cut-off points for clinically significant change: 1) the post-treatment score lies at least two standard deviations away from the 'dyfunctional' mean (labeled cutoff A); 2) the post-treatment score lies within two standard deviations of the 'functional' mean (labeled cutoff B); or 3) the post-treatment score lies closer to the mean of the 'functional' population than the mean of the 'dysfunctional' population (labeled cut-off C) (Jacobson and Truax 1991; Ogles, Lunnen, and Bonesteel 2001). Cutoff A can be established using published dysfunctional means, or by using sample measures of central tendency (i.e., pretest scores on the measure of interest) to represent the population parameters (although the latter is more common). Cutoff B requires access to a normal comparison group (or at least the mean of a normal comparison group), which, preferably, is similar demographically to the clinical sample. This could be from published information or data collected on a normal comparison group used in the study. Cutoff C requires both a 'functional' and 'dysfunctional' population, and is often preferable if this information is available since it uses information from both populations and thus allows for a more precise estimate of which population the individual belongs to. JT also present a 'reliable change index' (RCI) that determines whether the standardized change from pretest to posttest for each individual can be considered statistically significant. The RCI is designed to ensure that a post-test score that crosses the 'functional' cut-off point is indeed statistically reliable. The RCI is calculated as:

$$RCI = \frac{X_{post} - X_{pre}}{s_{diff}}$$

Here X_{pre} represents a client's pretest score, X_{post} represents that same client's posttest score, and S_{diff} is the standard error of difference between the two test scores. Although there have been several alternative and modified JT methods that have been proposed (e.g., Hageman and Arrindell 1999; Hsu 1999; Speer 1992; Speer and Greenbaum 1995), the original method remains the most popular (Bauer, Lambert, and Nielsen 2004), and is often recommended because it provides clinical significance results that are asymptotically equivalent to the modified procedures and does not require (possibly inaccurate) estimates of unknown population parameters.

As the JT statistics initially evaluate clients at the individual level, an important question that arises is how researchers conducting JT analyses summarize the RCI and cut-point results across all clients receiving a particular intervention in order to gain a more thorough understanding of the practical value of that intervention at the group level (for a detailed discussion of using individual level data to make group level statements see Cella, Bullinger, Scott, and Barofsky 2002). An obvious choice would be to simply compute proportions for each possible outcome from the RCI and cut-point results. For example, a researcher could calculate the proportion of clients who are recovered (i.e., met RCI and cut-point cut-offs), improved (met RCI but not cut-point cut-offs), deteriorated (met RCI cut-point but in the wrong direction), or unchanged (did not meet RCI or cut-point cut-offs). However, more sophisticated analyses could be conducted by comparing proportions (e.g., comparing category proportions for a single intervention or comparing category proportions for one intervention to those for a control group or another intervention group). In the study, we will identify and quantify the use of different methods for summarizing individual level clinical significance data.

Kendall et al. Method of Assessing Group Level Clinical Significance

Kendall et al.'s (1999) NC method approaches the problem of measuring clinical significance from a very different angle than the previously discussed individual level statistics. Instead of initially summarizing the data at the individual level (e.g., RCIs for each client), the group (i.e., intervention) level posttest means are computed and compared to a similar normal comparison group on the outcome variable of interest (e.g., depression, perfectionism). More specifically, the NC method assesses whether the treated clinical group is *equivalent to* a normal comparison group. An important advantage of the NC method is that it directly assesses the question of whether the intervention being investigated is able to return clients to a state of normal functioning. Kendall et al. accomplished the goal of assessing the equivalence of the treated and normal comparison groups by incorporating the equivalence testing methods proposed originally in the field of biopharmaceuticals (e.g., Schuirmann 1987; Westlake 1981) and introduced to psychology by Rogers, Howard and Vessey (1993), Seaman and Serlin (1998), and others. Unlike traditional null hypothesis test procedures that are designed to investigate a difference in population parameters (e.g., means), equivalence testing methods are designed to investigate the statistical equivalence of population parameters. With regard to the statistical equivalence of group means, equivalence testing methods, more specifically, seek to answer the question of whether the difference in means (e.g., differences in the treated and normal comparison means) is so small that it can be considered inconsequential. The amount of difference that is considered inconsequential is called the equivalence bound (δ) and is usually symmetric (i.e., the equivalence interval spans from $-\delta$ to δ). This bound is an important part of equivalence testing and needs to be selected based on the specific nature of each study. For detailed discussions of setting this bound see Rogers et al. or Cribbie and Arpin-Cribbie (2009).

In terms of understanding the difference between traditional and equivalence tests, recall that with traditional methods (e.g., *t* or *F* test) the research hypothesis relates to difference and thus the alternative hypothesis (H_a: $\mu_1 \neq \mu_2$ for a two-tailed test) appropriately also relates to difference, whereas for equivalence tests the research hypothesis relates to equivalence and thus the alternative hypothesis (H_a: $-\delta \leq \mu_1 - \mu_2 \leq \delta$) also appropriately relates to equivalence. Although it might be tempting to assess the equivalence of the treated and normal comparison means by looking for nonsignificance with a traditional difference-based test (e.g., *t* test) this would not be appropriate because, as any introductory statistics textbook teaches us, not rejecting H_o: $\mu_1 = \mu_2$ cannot be used to conclude that the population means are equal.

Kendall et al. (1999) proposed assessing the equivalence of the treated and normal comparison groups using the two one-sided tests (TOST) approach due to Schuirmann (1987). With this method, two null hypotheses are tested, H_{o1}: $\mu_T - \mu_{NC} \leq -\delta$ and H_{o2}: $\mu_T - \mu_{NC} \geq \delta$, where μ_T represents the population mean of the treated population and μ_{NC} represents the population mean of the normal comparison population. H_{o1} is rejected if $t_1 \leq t_{\alpha,\nu}$ and H_{o2} is rejected if $t_2 \geq t_{1-\alpha,\nu}$, where $t_1 = [M_T - M_{NC} - \delta] / s_{MT\text{-}MNC}$, $t_2 = [M_T - M_{NC} - (-\delta)] / s_{Mt\text{-}MNC}$, α represents the nominal significance level, M represents the sample mean, and ν and $s_{M1\text{-}M2}$ represent the degrees of freedom and standard error for the traditional *t* test, respectively. Cribbie and Arpin-Cribbie (2009) proposed the use of the heteroscedastic Schuirmann-Welch statistic (Gruman, Cribbie, and Arpin-Cribbie 2007), instead of the original Schuirmann statistic, in order to account for the often unequal variances (and sample sizes) of the treated and normal comparison groups, which biases the original Schuirmann test.

The purpose of the present paper is to determine the manner in which researchers are summarizing the clinical significance of interventions at the group level. In order to examine this issue, two literature reviews will be conducted. The first literature review evaluates how researchers are summarizing the results of individual level JT analyses (RCI, cutoffs). The second literature review investigates the frequency and nature of use of the NC approach. Together, these reviews are intended to provide information regarding current practices for assessing clinical significance at the group level. It is important to point out that this study is not intended to be a meta-analysis of the clinical effects; it was felt that these added results would only detract from the primary purpose of exploring the manner in which the analyses were conducted by turning the attention of readers to the study outcomes. However, a summary of the results of the clinical significance assessments for each study is available by contacting the first author.

Method

Literature Review 1: JT Method

A review of intervention studies that utilized the JT method was conducted in order to determine the manner in which individual level JT statistics (RCI, normative cutoff analyses) are being used to determine the clinical significance of specific interventions (i.e., group level analyses). The Google Scholar database was used to gather studies for this review. It is important to point out that the results returned from Google Scholar and PsycINFO were very similar, and thus we chose Google Scholar for its ease of access and replicability. Studies were included if they were peer-review journal articles published in 2010 or 2011, with the search requiring studies to have one of 'treatment', 'intervention' or 'therapy' in the title.

The search was limited to 2010 and 2011 because that time frame provided us with our target number of articles (150, which was deemed by the authors to provide an optimal balance between precision and time). Further, the article needed to reference the original Jacobson and Truax (1991) manuscript. Using these search criteria, 194 articles were identified. Thirty two articles were excluded because an intervention was not conducted, did not use the JT method or were not published journal articles (e.g., dissertations), leaving 162 articles that met inclusion criteria. The following information was extracted from the articles: 1) primary outcome variable; 2) type of intervention; 3) result of the statistical significance of the pre-post change (either raw or relative to a control group); 4) use or nonuse of the RCI statistic; 5) use or nonuse of a cutoff for moving closer to the functional population; 6) if a cutoff was adopted, which cutoff was applied; 7) method of summarizing the individual level statistics; and 8) inclusion/non-inclusion of an effect size if a summary statistic was utilized. Although Ronk, Hooke and Page (2012) highlight the importance of the selection of an appropriate outcome measure when conducting clinical significance analyses, given the wide variety of treatment outcomes and treatment outcome measures, we limited our investigation to only the primary outcome measure in each study.

Literature Review 2: NC Method

A review of intervention studies that utilized the NC method was conducted in order to determine the frequency and manner in which NC analyses are being applied to summarize group-level clinical significance. The Google Scholar database was also used to gather studies for this review (and again the results were very similar to those obtained using PsycINFO). The studies were obtained from the years 2000–2011, with the search requiring studies to be intervention studies published in peer reviewed journals with one of 'treatment', 'intervention' or 'therapy' in the title, and further that the article referenced the original Kendall et al. (1999) manuscript. It was necessary to include more years for the NC review than the JT review since there were fewer articles available that referenced the NC methodology than the JT methodology. This search resulted in 83 papers. Sixty articles were excluded because the authors did not utilize the Kendall et al. NC method and 11 studies were excluded because the authors used the NC method incorrectly. This resulted in 12 articles that met the inclusion criteria of appropriately using the NC method. The following information was extracted from these articles: 1) primary outcome variable; 2) type of intervention(s); 3) result of the statistical significance of the pre-post change (either raw or relative to a control group); 4) nature of the normative sample; 5) nature of the equivalence interval; 6) test statistic used for the NC analysis; and 7) result of NC analysis.

Results

JT Method Review Results

The results of the JT review are summarized in Table 1. It was found that of the 162 studies that met inclusion criteria, the median sample size was 36 for the primary intervention group, about a quarter of the studies used a control group, a third conducted some form of cognitive behavioral therapy (only three studies used non-psychological forms of intervention, e.g., pharmacological), and most (87 %) had a significant intervention effect. Further, the most common primary outcome variables were mood-related (e.g., depression, hopelessness; 22 %) and anxiety-related (e.g., generalized anxiety, post-traumatic stress, social phobia; 35 %), with many other outcomes including addictive behavior, eating disorder-related behavior, etc. accounting for the remaining 43 %. Note that this remaining 43 % also includes outcomes such as "well-being" which could cross many categories including the anxiety-related and mood-related categories above. As expected given the search criteria, almost all conducted individual level clinical significance analyses. Further, almost all studies also conducted some form of group level clinical significance analyses. For the individual level tests, the most frequently used cutoff was A (relative to the mean of the dysfunctional population), which is not surprising given that this is probably the easiest method for which to obtain a comparison level (e.g., using published clinical means or using the existing pre-intervention means) since normative data is often not available (Jacobson and Truax 1991).

For conducting the group level analyses, about an equal number of studies used both RCI and cutoff information and RCI information alone. Less than 10 % of studies solely relied on whether the client met normative cutoffs to summarize the group level analyses. About three-quarters of the group level analyses simply summarized the proportion of clients who improved, deteriorated, etc. However, 15 % of studies compared the proportions across treatment groups (or across treatment and control groups) using, for example, a χ^2 test of independence. Four studies (3 %) compared the proportion of clients that met each standard within only the treatment group (e.g., conducted a χ^2 goodness of fit test).

NC Method Review Results

The results of the NC review are summarized in Table 2. What is initially interesting is that about an equal number of studies correctly and incorrectly adopted the Kendall et al. (1999) NC method. The primary fault in applying the NC method was that the researchers did not conduct an equivalence test, the primary test necessary for concluding whether the treated population is equivalent to the normal comparison population. This finding may not be surprising given that equivalence

Table 1 Conclusions from the Jacobson-Truax (JT) review study

Review item	Results
Total Number of Studies[a]	162
Mean, Median Group Sample Size[b]	134, 36
With a Control Group	26 %
With a Significant Treatment Effect[c]	87 %
Conducted Individual Level Clinical Significance Analyses	99 %
Conducted Group Level Clinical Significance Analyses	97 %
Used JT Cutoff A, B, C [d]	46 %, 14 %, 23 %
Used RCI and Cutoffs for Group Level Test	49 %
Used only RCI for Group Level Test	42 %
Used only JT Cutoffs for Group Level Test	9 %
Group Level Test based only on Proportions	80 %
Group Level Test based on Between-Group Chi-Square Test	15 %
Group Level Test based on Within-Group Chi-Square Test	3 %
If Chi-square Conducted, % Including an Effect Size	7 %

[a] Studies that applied the JT method

[b] Based on the primary treatment group

[c] Significant change, or significant change relative to control group

[d] Of studies that used a cutoff and the cutoff used could be determined; one study used both cutoffs A and C

RCI reliable change index

testing methods are relatively novel and not available with most statistical software packages, however it definitely highlights the need for more coverage in the literature on the appropriate use of the NC method. Of the studies that correctly used the NC method, the median sample size was 30 for the primary intervention group, most (83 %) used some form of cognitive behavioral therapy (all interventions were psychological in nature), two thirds had anxiety as the primary outcome, almost all had a significant intervention effect, and slightly less than half had a control group. However, we caution that these numbers are based on a very small set of articles and thus it is important to not read too much into the specific proportions. The normative sample descriptive statistics used in these studies were all derived from previously published

Table 2 Conclusions from the normative comparisons (NC) review study

Review item	Results
Total Number of Studies that Purported to use the NC Approach	23
Number of Studies that Correctly used the NC Approach[a]	12
Number of Studies that used the NC Approach Incorrectly[b]	11
Mean, Median Group Sample Size[c]	79, 30
Mean, Median Normative Sample Size	907, 233
With a Control Group	42 %
With a Significant Treatment Effect[d]	92 %
Used Schuirmann's Two One-Sided Test Procedure[e]	100 %
Used 1 sd, 2 sd from the Normative Mean as δ[f]	86 %, 14 %
Treated sample declared equivalent to normative sample	50 %
Studies using published normative data[g]	100 %

[a] Studies that correctly applied the equivalence-based normative comparisons approach

[b] These studies were not used to calculate any of the statistics below

[c] Based on the primary treatment group

[d] Significant change, or significant change relative to control group

[e] One study did not indicate which test was used and thus was excluded from this calculation

[f] Five studies did not indicate what equivalence interval was used

[g] Instead of collecting normative data that was representative of the treated sample; two studies did not provide information regarding the nature of the normative sample and were excluded from this calculation

studies (i.e., none collected the normative data), and the median sample size of these studies was 233. All of these studies used Schuirmann's two one-sided test procedure (the procedure described in the original Kendall et al. article) and all but one used one standard deviation of the normative sample as the equivalence interval. Half of the studies found that the intervention group following treatment could be considered equivalent to a normal comparison group.

Discussion and Recommendations

As was expected, almost all studies conducted some sort of group level summary, with many more studies utilizing the JT method (162 studies used the JT method over just a 2 year period) than the NC method (only 12 studies correctly used the NC method over an 11 year period). An interesting finding is that across both reviews (JT/NC) less than 30 % of studies had a control group. This is very serious given that statistical analyses investigating change over time as a function of the interventions can be very misleading when they are not conducted relative to a control group (due to regression to the mean, placebo/waitlist effects, etc.).

Although most of the studies that adopted the JT method provided a straightforward summary of how many people met predetermined cutoffs (e.g., how many changed reliably and met criteria for moving towards normal functioning, i.e., were 'recovered'), there are some important points regarding how these summaries were obtained and what follow-up tests are available. First, most researchers (83 %) that used cutoffs used one of the cutoffs suggested by Jacobson and Truax (1991), namely falls outside the dysfunctional range (A), falls within the normative range (B), or falls closer to the mean of the normative population (C) (one study used both cutoffs A and C). As Kazdin (1999) discusses, it is important to consider what cutoff provides the best balance between specificity and sensitivity (i.e., correctly classifying recovered clients and not recovered clients). As an introduction to the problem, Jacobson and Truax explain that when norms are available B or C are recommended, with C preferable when the functional and dysfunctional populations overlap, and B preferable when the distributions do not overlap (when normative data is not available, A is the only cutoff available). It is a little disconcerting that almost half of the studies used cutoff A, given the preference in the literature for cutoff C (e.g., Jacobson and Truax 1991; Bauer et al. 2004) since most distributions overlap. The preference for cutoff A may relate to the fact that normative data is relatively scarce for many measures and most studies did not have a control group. It is important though that researchers acknowledge that clinical and normative statistics are available for many common psychological measures, and the incorporation of this information will allow for potentially more meaningful comparisons (e.g., use of Cutoff C). Very few studies had any discussion of how the particular cutoff used was selected, and thus we recommend that researchers consider what the most appropriate cutoff might be that maximizes sensitivity and specificity.

Secondly, less than 20 % of the studies compared the proportions in each category (e.g., using a chi-square test), and only 7 % of those that statistically compared proportions included an effect size. Although the chi-square test can simplify the reporting of group level statistics, it is important that these analyses are conducted in a logical manner. For example, imagine that for the treatment group 30 % were 'recovered', 40 % 'improved', 20 % were 'unchanged' and 10 % 'deteriorated', and for the control group 30 % were 'recovered', 40 % 'improved', 1 % were 'unchanged', and 29 % deteriorated. If a researcher were to run a chi-square test of independence on this data, the conclusion would likely be a significant effect indicating that the proportions differ across the categories. However, in this instance, it is clear that this result is only indicating a difference in the number of clients who were 'unchanged' or 'deteriorated' (a result that may be of different importance or relevance to a researcher).

Thus, there are a few recommendations regarding using group-level comparisons of proportions following JT (or similar) methods: 1) Follow-up tests comparing groups across all categories would be meaningful and would add important information. For example, if there was a control group, comparing the proportion of 'recovered', 'improved', 'unchanged', and 'deteriorated' clients across the treatment and control group would be very informative. If there was not a control group, comparing these categories across the treated group would also add useful information. 2) Related to the first point, specific comparisons would provide more precise information about the nature of the effect than a global test of independence. For example, comparing those improved/recovered to those unchanged/deteriorated in the treatment and control groups using a 2×2 chi-square test of independence (or just the treatment group using a chi-square goodness of fit test), would provide more specific information about how the treatment response differed across groups. 3) Effect size measures applied to meaningful comparisons (such as that proposed in the previous point) would be recommended as complements to traditional null hypothesis testing methods (such as those proposed in the previous two points) because we do not want the sample size of the study to have such a large influence on the calculation of the practical significance of the intervention (Wilkinson et al. 1999). Examples of informative effect size measures are odds ratios and correlation-based measures such as Cramer's V. 4) Lastly, the JT method assumes normal distributions when establishing cutoff points and the RCI, and thus researchers should be aware of the potential effects of nonnormality. In their paper, Jacobson and Truax concede that such an assumption is a problem that limits the generalizability of their method. This is important

since nonnormality is frequently encountered with psychological variables (Golinski and Cribbie 2009; Micceri 1989; van Wieringen and Cribbie 2014).

An alternative approach for summarizing the clinical significance of an intervention at the group level is to ask whether the treated group is equivalent to a normal comparison group following the intervention. For example, the NC method proposed by Kendall et al. (1999) uses equivalence testing to determine if the treated group can be considered statistically equivalent to a normal comparison following the treatment. Of the 23 studies that purported to adopt this method, only about one-half actually conducted the method accurately. In most cases where the method was applied incorrectly, the researchers did not use an equivalence test to actually compare the treated and normal comparison groups; for example, some simply determined the proportion of clients that fell close to (e.g., one sd from) the normative mean (i.e., a sort of modified JT method). Of those that correctly used the method, all used the Schuirmann (1987) two one-sided testing approach for assessing the equivalence of the treated and normal comparison groups. One concern with the use of this method is that it is not robust to unequal sample sizes and variances (Cribbie and Arpin-Cribbie 2009; Gruman, Cribbie, and Arpin-Cribbie 2007). This is especially problematic given that the group sizes and standard deviations of normal comparison and treated clinical groups often differ substantially. For example, the median sample size of the clinical groups was 30, whereas the median sample size of the normal comparison group was 233. Since methods exist that do not rely on the assumption of variance homogeneity (e.g., Gruman et al. 2007; Koh and Cribbie 2013), there is no reason to rely on outdated approaches.

Further, given past studies that have found that the distributions of variables in psychology are often nonnormal (see above), it is also important to use a test statistic that is robust to violations of the normality assumption. Robust tests for evaluating the equivalence of independent groups have been discussed by Gruman, Cribbie and Arpin-Cribbie (2007), Cribbie and Arpin-Cribbie (2009), Koh and Cribbie (2013), van Wieringen and Cribbie (2014), among others, and are recommended over the original Schuirmann procedure since it is expected that the assumptions of variance homogeneity and/or normality will be violated across clinical and normal comparison groups.

Another concern with the studies that adopted the NC method was that all used published normative data instead of collecting normative data that was more representative of the clinical sample. In some cases it is possible to find a normative sample that is somewhat representative of the characteristics of the clinical sample, however in many cases this will not be the case. As expected, comparing groups that differ in ways other than just having or not having the clinical disorder of interest (e.g., differing in age, culture, education, etc.) severely limits the validity of the equivalence-based test. Thus, although it can be time consuming, in some situations it is necessary for researchers to collect representative normative data in order to ensure that the comparisons conducted are meaningful.

Lastly, most studies used an equivalence bound of one standard deviation (of either the control or clinical group) when conducting the test of equivalence. Although this has become standard practice following its use by Kendall et al. (1999) in their examples, it is important, as Kendall et al., Rogers, Howard and Vessey (1993), and others, highlight, that the equivalence interval is tailored to the specific nature and goals of the study.

Before concluding, it is important to consider potential limitations of the study. First, we used a specific set of search criteria for finding articles for this review. It is possible that a different set of criteria may have uncovered more or different articles than we found, and this could potentially have affected the results and conclusions. Second, and related to the first limitation, is that the number of articles found that correctly utilized the NC method was very low. Thus, we caution the reader about making any general conclusions regarding the manner in which the NC method is being used in practice. As use of the relatively new method increases, much more reliable results based on a larger set of articles will be available. Third, we have only considered two methods for assessing group-level clinical significance; specifically, we have considered two popular statistical approaches for summarizing group level clinical significance. Further, these methods generally use popular scales/questionnaires to measure the behavior of interest. However, the degree to which changes in scores on a specific inventory (e.g., Beck Depression Inventory) translates into real-world/everyday improvements in functioning can vary from behavior to behavior. Further research should be dedicated to studying the measurement of these important behavioral changes that are (often) more difficult to measure. For example, measuring the clients' subjective level of functioning, assessments of functioning from others close to the client (Cella et al. 2002), including assessments related to functioning in occupational or other life activities, may prove to be a valuable tool for understanding clinical significance at the group level. Lastly, and not necessarily a limitation of the study but of the incorporation of the methods discussed in the study, is the difficult task of finding an appropriate normal comparison group and measurement instrument. There has been a lot of discussion surrounding what constitutes an appropriate normal comparison group (e.g., Rogers et al. 1993), however many issues still exist. For example, as an anonymous reviewer of this paper asked, is it appropriate to compare a clinical group to a normal group that contains individuals who would meet the criteria for the clinical group or are even actively involved in treatment. Once the definition of an appropriate normal comparison group is

established, an even more difficult task might be to find data for this group. Many researchers have highlighted the paucity of normal comparison data for popular scales in the literature, and collecting valid normal comparison data is often beyond the scope of most investigations. Further, in some cases the scales may not have been validated on normal comparison groups and there are questions regarding the appropriateness of the scale for use with a normal comparison population. A related issue is whether there is a linear relationship between scale scores and the level of pathology.

To summarize, this study looked at two common approaches for summarizing the group level effect of an intervention. One strategy explores (and/or compares) the proportion of subjects that reliably changed and met cutoffs for moving towards normal functioning (JT method), while the other assesses whether the treated group is equivalent to a normal comparison group following the intervention (NC method). While these are effective and highly recommended strategies for addressing different questions regarding the clinical significance of an intervention, the validity of both approaches relies to a great extent on the methodology used. Our primary recommendation is that researchers carefully ponder what measure of clinical significance is most appropriate for their given research. The JT method more directly incorporates individual results; however, the importance of selecting an appropriate cutoff for determining if the client progressed toward normal functioning cannot be overestimated (e.g., simply comparing the treated sample to their pretest mean may not be an appropriate contrast). The NC method more directly assesses whether the treated group has been returned to a state of normal functioning, however individual results have less impact, it can often be difficult to obtain appropriate normal comparison data, and it is important that researchers adopt an appropriate test statistic given the frequent violations of parametric test assumptions. It should be clear that numerous factors much be considered in order to ensure reliable and valid information regarding group-level clinical significance.

References

Atkins, D. C., Bedics, J. D., McGlinchey, J. B., & Beauchaine, T. P. (2005). Brief report: Assessing clinical significance: Does it matter which method we use? *Journal of Consulting and Clinical Psychology, 73*, 982–989.

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

Cella, D., Bullinger, M., Scott, C., & Barofsky, I. (2002). Group vs individual approaches to understanding the clinical significance of differences or changes in quality of life. *Proceedings of the Mayo Clinic, 77*, 384–392.

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

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

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.

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

Hsu, L. M. (1999). Caveats concerning comparisons of change rates obtained with five methods of identifying significant client changes: Comment on Speer and Greenbaum (1995). *Journal of Consulting and Clinical Psychology, 67*, 594–598.

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.

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.

Kazdin, A. E. (1999). The meanings and measurement of clinical significance. *Journal of Consulting and Clinical Psychology, 67*, 332–339.

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.

Koh, A., & Cribbie, R. A. (2013). Robust tests of equivalence for *k* independent groups. *British Journal of Mathematical and Statistical Psychology, 66*, 426–434.

Maassen, G. H. (2000). Principles of defining reliable change indices. *Journal of Clinical and Experimental Neuropsychology, 22*, 622–632.

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

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

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

Ronk, F. R., Hooke, G. R., & Page, A. C. (2012). How consistent are clinical significance classifications: When calculation methods and outcome measures differ? *Clinical Psychology: Science and Practice, 19*, 167–179.

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

Seaman, M. A., & Serlin, R. C. (1998). Equivalence confidence intervals for two-group comparisons of means. *Psychological Methods, 3*, 403–411.

Speer, D. C. (1992). Clinically significant change: Jacobson and 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.

van Wieringen, K., & Cribbie, R. A. (2014). Robust normative comparison tests for evaluating clinical significance. *British Journal of Mathematical and Statistical Psychology, 67*, 213–230.

Westlake, W. J. (1981). Response to T.B.L. Kirkwood: Bioequivalence testing — a need to rethink. *Biometrics, 37*, 589–594.

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