Replication of a comparison of statistical techniques using clinical significance methods

Kiryl N. Shada
Pacific University
Replication of a comparison of statistical techniques using clinical significance methods

Abstract
Clinically significant change is an important aspect of the therapeutic process. Computing significant change allows for measurement of client progress beyond clinical judgment. In this study, four methods of reliable change and clinical significance were compared using a sample of 395 individuals who attended psychotherapy in a clinical setting. Differences in classification were found between methods; this suggests that the method chosen to determine reliable change and clinical significance plays a factor on estimates of meaningful change. Consistent with prior outcome studies, the Jacobson and Truax method (Jacobson & Truax, 1991) provided a moderate estimate of reliable change and clinical significance and therefore is recommended for use in future outcome studies. Future research is necessary to combine client satisfaction with treatment outcome using a standardized measure in order to consider more factors when considering treatment outcomes.

Degree Type
Thesis

Rights
Terms of use for work posted in CommonKnowledge.

Comments
Library Use: LIH

This thesis is available at CommonKnowledge: http://commons.pacificu.edu/spp/469
Copyright and terms of use

If you have downloaded this document directly from the web or from CommonKnowledge, see the “Rights” section on the previous page for the terms of use.

If you have received this document through an interlibrary loan/document delivery service, the following terms of use apply:

Copyright in this work is held by the author(s). You may download or print any portion of this document for personal use only, or for any use that is allowed by fair use (Title 17, §107 U.S.C.). Except for personal or fair use, you or your borrowing library may not reproduce, remix, republish, post, transmit, or distribute this document, or any portion thereof, without the permission of the copyright owner. [Note: If this document is licensed under a Creative Commons license (see “Rights” on the previous page) which allows broader usage rights, your use is governed by the terms of that license.]

Inquiries regarding further use of these materials should be addressed to: CommonKnowledge Rights, Pacific University Library, 2043 College Way, Forest Grove, OR 97116, (503) 352-7209. Email inquiries may be directed to: copyright@pacificu.edu

This thesis is available at CommonKnowledge: http://commons.pacificu.edu/spp/469
REPLICATION OF A COMPARISON OF STATISTICAL TECHNIQUES USING CLINICAL SIGNIFICANCE METHODS

A THESIS
SUBMITTED TO THE FACULTY
OF
SCHOOL OF PROFESSIONAL PSYCHOLOGY
PACIFIC UNIVERSITY
HILLSOBORO, OREGON

BY
KIRYL N. SHADA

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN CLINICAL PSYCHOLOGY

JULY 26TH 2013

APPROVED:

Lisa R. Christiansen, PsyD
Abstract

Clinically significant change is an important aspect of the therapeutic process. Computing significant change allows for measurement of client progress beyond clinical judgment. In this study, four methods of reliable change and clinical significance were compared using a sample of 395 individuals who attended psychotherapy in a clinical setting. Differences in classification were found between methods; this suggests that the method chosen to determine reliable change and clinical significance plays a factor on estimates of meaningful change. Consistent with prior outcome studies, the Jacobson and Truax method (Jacobson & Truax, 1991) provided a moderate estimate of reliable change and clinical significance and therefore is recommended for use in future outcome studies. Future research is necessary to combine client satisfaction with treatment outcome using a standardized measure in order to consider more factors when considering treatment outcomes.

Keywords: Clinical significance, reliable change index, reliable change, treatment outcomes
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>i</td>
</tr>
<tr>
<td>INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>Reliable Change</td>
<td>2</td>
</tr>
<tr>
<td>Clinical Significance</td>
<td>3</td>
</tr>
<tr>
<td>Review of Current Methods</td>
<td>4</td>
</tr>
<tr>
<td>Jacobson and Truax Method</td>
<td>5</td>
</tr>
<tr>
<td>Process of the Jacobson and Truax Method</td>
<td>6</td>
</tr>
<tr>
<td>Review of the Jacobson and Truax Method</td>
<td>8</td>
</tr>
<tr>
<td>Gulliksen-Lord-Novick Method</td>
<td>10</td>
</tr>
<tr>
<td>Edwards-Nunnally Method</td>
<td>11</td>
</tr>
<tr>
<td>Hageman-Arrindells Method</td>
<td>11</td>
</tr>
<tr>
<td>COMPARISON OF METHODS</td>
<td>12</td>
</tr>
<tr>
<td>PURPOSE OF THE CURRENT STUDY</td>
<td>15</td>
</tr>
<tr>
<td>METHOD</td>
<td>16</td>
</tr>
<tr>
<td>Participants</td>
<td>16</td>
</tr>
<tr>
<td>Outcome measure</td>
<td>17</td>
</tr>
<tr>
<td>Procedure</td>
<td>18</td>
</tr>
<tr>
<td>RESULTS</td>
<td>19</td>
</tr>
<tr>
<td>DISCUSSION</td>
<td>22</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>28</td>
</tr>
<tr>
<td>APPENDIX</td>
<td>33</td>
</tr>
</tbody>
</table>
Introduction

Clinical and practical meaning of reliable change became important in the late 1970’s as researchers became determined to scientifically demonstrate that therapy in fact helps individuals improve in a way that was deemed observable (Ogles, Lunnen, & Bonesteel, 2001). Prior to popular methods used today, standard differences between groups and effect sizes were used as a means to detect individual change as a result of treatment. Jacobson, Roberts, Berns, and McGlinchey (1999) described the original method of detecting change in an individual during treatment as limited. They stated that although a large standard error makes it more likely for clinical significance, it is not a guarantee, thus the original method was not highly reliable. Maason (2001) provided one perspective of the history of clinically significant change; his account of the history of the Reliable Change Index is as follows:

As early as 1962, McNemar (1962) proposed a method to establish whether a difference score can be considered *dependable* — the term reliable change had not yet been introduced. The central criterion in this method was the ratio of the observed change and the standard error of measurement of the difference. I am not sure that he was the first to propose this method, nor that this was the first publication of his proposal. As a precaution, I call this method the *classic approach*. Jacobson, Follette and Ravenstorf (1984) introduced the term *Reliable Change Index* (RCI) and proposed an index with a denominator slightly different from the classic approach. Christensen and Mendoza (1986) showed that their formula was wrong and proposed a formula in a rather ill-chosen notation, which boiled down to the classic approach. Jacobson and his colleagues (Jacobson & Truax, 1991) acknowledged their mistake and ever since, in the field of
psychotherapeutic research, the classic approach has rather undeservedly been known as Jacobson’s RC. (pp. 495-496)

**Reliable Change**

Hsu (1996) defined reliable change as “a change which cannot reasonably be attributed to measurement error” (p. 374). Change can occur as a result of a number of things, including true change, measurement error, practice effects, or bias. Researchers and clinicians in the field of psychotherapy are only interested in true change (Maassen, 2000). Jacobson, Follette, and Revenstorf (1984a) originally proposed the Reliable Change Index as a process to evaluate change in pretest and posttest scores that are considered true change; true changes are changes that reflect an individual’s observable progress. Evans, Margison, and Barkham (1998) described the meaning of reliable change with a simple question: “Has the patient changed sufficiently to be confident that the change is beyond that which could be attributed to measurement error?” (p. 70)

Determining the amount an individual has changed after an intervention is a very important aspect of psychotherapy; it allows clinicians to track client progress that shows how an individual is responding to treatment allowing for proper termination or change in therapeutic approach. The current study explores five different methods used to calculate reliable change, which are discussed in detail in the following respective sections. These distinct methods can be reduced to the same fundamental equation using a calculated difference between pretest and posttest scores resulting in a standard Z score.

The fundamental equation is as follows:

$$\frac{Y - Y'}{SE}$$
In the numerator, \( Y - Y' \) indicates the difference between the pretest score and the posttest score. The calculation of the standard error in the denominator of the reliable change equation is the main source of debate among authors of the different methods for determining clinical significance (Hinton-Bayre, 2010). Once the value for the reliable change is computed, the next step is determining if that value representing individual change is significant. As a result, clinical significance is an important component when determining reliable change. The following sections will include a definition of clinical significance, as well as discussions of the history of clinical significance and current theories of clinical significance.

**Clinical Significance**

Jacobson, Roberts, Berns, and McGlinchey (1999) stated that the goal of clinical significance, routinely defined as returning to the nonclinical population, “was to report psychotherapy data in a way that was clinically meaningful, given the expectations that consumers have about psychotherapy” (p. 305). The authors also noted that operationalizing clinically significant change provided meaningful data for clinicians beyond standardized group comparisons as well as provided information to the client as a “consumer of mental health services” (p. 300). Evans, Margison, and Barkham (1998) developed a question explaining the meaning of clinically significant change: “How does the end state of the patient compare with the scores observed in socially and clinically meaningful comparison groups?” (p. 70). There are three statistically derived indices that can be used to determine clinical significance (these are discussed in greater detail in a section discussing the Jacobson and Traux [JT] method).

Throughout the history of clinical significance, the methodology has changed quite a bit but the question has remained the same: “How likely is it that a client with a particular problem will leave therapy without that problem?” (Jacobson et al., 1999, p. 306)
Through the literature a number of proposed statistical outcome methods have been revealed; each method’s author(s) claim to have the best method to classify individuals in psychotherapy. All four analyses discussed in detail in the present study are based on the three assumptions of classical test theory as follows, as stated by McGlinchey, Atkins, and Jacobson (2002). First, “both error components are mutually independent and independent of the true pretest and posttest scores” (McGlinchey et al., 2002, p. 519). Second, both error components have a normal distribution with the mean falling at zero. Third, “the standard error of the difference of the two error components is equal for all participants” (McGlinchey et al., 2002, p 519).

**Review of Current Methods**

Current methods for evaluating reliable change can be grouped into two types: estimation interval methods and null hypothesis methods. The different methods differ in two distinct ways; the way true change is estimated for an individual and the way that standard error is used. The null hypothesis method has an advantage of using the observed change as the unbiased estimate of true change, whereas the estimation interval methods use normative data from the population or sample (Maasson, Bossema, & Brand, 2009). Hinton-Bayer (2010) made the point that “agreement between RC models will depend on practice effect, reliability, variance inequality, and the individual cases’ relative position to the control group at initial testing” (p. 251). Different RC methods can lead to different conclusions in regards to the effect of a treatment or interventions on an individual. In addition, the content validity of the measure in questions plays an important role in the accuracy of determining significant and reliable change. Reliable change is important because in psychotherapy researchers are only interested in true change, whether or not the individual change has clinical significance (Maassen, 2000). The

**Jacobson and Truax Method**

The Jacobson and Truax (JT) method, also known as the ‘classical approach’ in empirical research, is a null hypothesis method of evaluating clinically significant change (Maassen, 2001). It is also one of the most frequently used and applied methods for analysis of clinically meaningful change (Maasseen, 2004). Jacobson et al. (1999) explained the premises of the JT method as the following:

The final product is a twofold criterion for clinically significant change: (a) The magnitude has to be statistically reliable and (b) by the end of therapy, clients have to end up in a range that renders them indistinguishable from well-functioning people. By applying our metric to a population of treated clients, one can determine the percentage of clients who improved but did not recover, the percentage of clients who recovered, and the percentage of clients who remained unchanged or who deteriorated in each treatment condition. (p. 300)

The JT method was established after researchers proposed that a change in therapy is clinically significant when an individual moves from a clinical range to a nonclinical range during the course of therapy. The JT method provided a statistical approach to a classification system for client progress divided into four groups: deteriorated, unchanged, improved, or recovered (Jacobson, Follette, & Ravenstrof, 1984b). This method was instituted to rule out measurement error as a means to explain an observed change in an individual. Following the original proposal, a second step was added to make the JT method a two-step criterion for
clinically significant change (Jacobson & Truax, 1991). This adaptation of the original JT approach accounts for differences between initial and final variance and is empirically supported; it is preferred above using the standard error in the equation (Maasseen, 2004). The addition also accounts for the unreliability of both the pre- and posttest scores by looking at both scores as distinct distributions (Hagman & Arrindell, 1993).

**Process of the Jacobson and Truax Method**

The first part of the two-step process of the JT method is to establish a cutoff point for a measure of functioning that separates the ‘nonclinical’ population from the ‘clinical’ population. When establishing cut off points there are three reasonable options to consider as identified by Jacobson and Traux (1991). The first, **Cutoff A**, is defined as “two standard deviations beyond (in the direction of functionality) the mean for that population” (Jacobson & Traux, 1991, p. 13). That is, Cutoff point A is obtained when the score falls out of the range of the clinical population towards the range of the nonclinical population. This cutoff typically results in an individual correctly identified as nonclinical when scoring below the cutoff; Cutoff A has high sensitivity. The equation for Cutoff A is as follows:

\[ M_{\text{clinical}} - 2SD_{\text{clinical}} \]

The second, **Cutoff B**, is defined as “the point 2 SD within a recognized nonclinical mean” (Jacobson & Traux, 1991, p. 13). Note that this cutoff score can only be utilized when nonclinical data is also available. This cutoff is the most lenient because of overlap between functional and nonfunctional group distributions; it also has high specificity. The equation for Cutoff B is as follows:

\[ M_{\text{nonclinical}} - 2SD_{\text{nonclinical}} \]
The last, Cutoff C, is a weighted midpoint between the means of a nonclinical and clinical population (Jacobson & Traux, 1991). This cutoff point is based on the relative probability that an individual’s score will end up in one population instead of the other population (Jacobson et al., 1999). When normative data is available, Cutoffs B or C are preferable to Cutoff A. Cutoff C is preferred when nonclinical data is available and there is overlap between the clinical and nonclinical distributions; if there is not overlap in the distributions and normative data is available, Cutoff B is recommended. When normative data is not available, Cutoff A is the only option (Jacobson & Traux, 1991).

The second part of the JT method is used to determine whether the change found is truly a result of the individual changing and not the result of measurement error. The JT Method is considered an assessment of statistically reliable change that uses a Reliable Change Index (RCI) (Maasseen, 2004). An RCI is calculated using psychometric properties of the measurement or assessment tool used. When an individual passes the statistically determined RCI in a positive direction, change that is not due to chance is demonstrated. Originally the RCI was calculated as the difference in pretest and posttest scores divided by the standard error of measurement (SE), which consisted of the test-retest reliability estimate. The RCI was designed to account for the magnitude of change and measurement error. Thus, the less reliable an instrument, the more change required to achieve a statistically reliable change (Wise, 2004). Christensen and Mendoza (as cited in Wise, 2004) proposed a change or more stringent correction to the original JT RCI formula. They proposed using “the amount of difference which one could expect between two scores, obtained on the same test by the same individual, as a function of measurement error alone” (p. 52) ($S_{diff}$) in place of SE in the denominator of the RCI.

The equation for the JT method is as follows:
Once the two-step JT method is completed, one may determine in which classification group the individual belongs; this is determined by using the cutoff point and the RCI combined. For example, Atkins Bedics, McGlinchey, and Beauchaine (2005) used Cutoff A and the procedure was as follows: the individual was considered recovered if they passed Cutoff A and the RCI in the positive direction; the individual was considered improved if they passed the RCI in the positive direction but not Cutoff A; the individual was considered unchanged if they passed neither criterion; or the individual was considered deteriorated if they passed the RCI in the negative direction. The method for classifying the individual to the appropriate group would be the same when using Cutoff B or Cutoff C.

**Review of the Jacobson and Truax Method**

Although there is no agreement for which method of calculating clinically significant change is the most useful in clinical settings, research discussed in more detail below indicates that the JT method is preferred. For example, McGlinchey and Jacobson (1999) gave credit to the newer methods but indicated that the results were very similar using the JT method, thus they deemed it unnecessary to use a more complex method than the JT method. The newer methods tend to be more complex as a result of controlling for more error but have not shown any remarkable advantages over the JT method (Maassen, 2001; 2004). McGlinchey, Atkins, and Jacobson (2002) agreed upon using the JT method and argued that this method and other existing methods should be improved upon before any new methods are created. Ferguson, Robinson, and Splaine (2002) reported that the JT method had advantages over using the Standard Error of Measure (SEM) and estimating an effect size in change scores when considering clinically significant change. Most recently, in a study for reliable change in evaluation of treatment for
substance abuse, researchers found that the JT method was the most accurate as well as the most conservative for assessing improvement when compared to EN, GLN, and HA (Marsden et al., 2010). According to Maassen (2004), advantages of new approaches are “more apparent than real” (p. 888) and he and other researchers recommended stopping the use of alternative statistical methods and using instead the classical approach (JT method).

Not all authors agree with the assessment that the JT method should be the method of choice. For example, Maason (2000) argued that other methods use information that allow for more precise estimates. In addition, Hsu (1996) presented three main criticisms of the JT method. These criticisms included: lack of realistic assumptions, arguments against using the standard error in the denominator, and failure to account for regression to the mean. Hsu (1999a) pointed out another limitation when using the JT method, which is that pretest scores falling within the extreme dysfunction range on measures can result in an overestimation of positive treatment effects. Tingey, Lambert, Burlingame, and Hansen (1996) reported problems with the use of normative samples defining Clinical Significance. The authors described three criticisms of the JT method, including the inability to use a normative sample across studies, restriction of the social validation method by limiting data to only one clinical sample and one nonclinical sample, and a lack of a means to identify sample distinctness. The JT method was recommended in situations where practice effects can be ignored (McGlinchey et al., 2002); this is because the JT method does not clearly differentiate practice effects from true change and may result in inflated rates of improvement or deterioration in individuals who have not demonstrated true change (Temkin, 2004). Two additional limitations are that the JT method assumes that pretest and posttest are parallel measures as well as estimates the reliability coefficient and the variance using the test retest correlation and the variance of the pretest (McGlinchey et al., 2002).
According to Bauer, Lambert, and Nielson (2004), extensions for assessing clinical significance have been proposed, but there have been very few studies in the literature that utilize them.

Some authors, such as Spear and Greenbaum (1995), Hsu (1999), and Hageman and Arrindell (1993), argue that alternative methods that compute clinically significant change have advantages over the traditional JT method. One area that authors and supporters of new methods, including the GLN method, the EN method, and the HA method, attempt to account for is regression to the mean as a result from pretest-posttest regression effects. The HA method further differs from the JT method by including a focus on different rates of change for a group versus an individual; this results in a more complex approach. One last reliable change method, the Hierarchical Linear Method (HLM), differs from the JT method by incorporating numerous data points in the equation as opposed to the traditional approach using a pretest and posttest score (Bauer et al., 2004). These more recently proposed methods indicate that procedures used to estimate true change should be improved (Maassen, 2000). Hsu (1999b) pointed out that the methods are based off different assumptions and, as a result, it is not surprising that each method may arrive at a different conclusion for the same individual. Each additional statistical method utilized in this study method is described in more detail in the following sections.

**Gulliksen-Lord-Novick Method**

The Gulliksen-Lord-Novick (GLN) method was proposed as an alternative ‘regression estimate’ method following Louis Hsu’s criticism of JT method (Atkins et al., 2005; Marsden et al., 2010). The GLN method assumes that when no treatment effects are present, pretreatment and posttreatment scores can be viewed as parallel measurements. In other words, “clients’ true scores do not change and standard errors of measurements are equal for pretest and posttest scores” (Hsu, 1999a, p. 594). This method differed from the original JT method by
controlling for regression to the mean; in order to achieve this, the hypothesized population mean of a relevant group was used. If the population mean is not available it is recommended that the pretest scores of the participants be used (Bauer et al., 2004). A limitation of this method is the fact that population means and Standard Deviations are rarely known (Atkins et al., 2005). In a previous review of reliable change this method has been predicted as more conservative than the JT reliable change index (Marsden, 2010).

**Edwards-Nunnally Method**

The Edwards-Nunnally method (EN) was proposed by Speer (1992) after he and a group of colleagues advocated for the formation of confidence intervals when looking at reliable change. Confidence intervals of about 2 standard deviations and centered on the individual’s true score are used as the criteria for determining deterioration or improvement (Jacobson et al., 1999). Like GLN, the EN method was originally intended to minimize the risk of regression to the mean. However, for the EN method, an estimated true score based on the obtained pretest score is used (Wise, 2004). The EN method establishes reliable change using an individual’s posttest score relative to an established confidence interval around the estimate of the individual’s pretest score. This was proposed to improve the original clinical significance method by minimizing the influence of regression to the mean (Bauer et al., 2004). Multiple authors concluded that using this method would result in more conservative results; meaning that fewer individuals would be categorized as reliably improved and more individuals would be categorized as reliably deteriorated (Marsden et al., 2010; Wise, 2004). Hsu (1999) indicated that this method was flawed because it uses the standard error of measurement as opposed to the standard error of prediction.

**Hageman–Arrindells Method**
The Hageman–Arrindells method (HA) was proposed in 1992 as a modification to the traditional RC Index by taking into account regression to the mean; this was intended to improve the pre-post difference score (Hageman & Arrindell, 1999). HA was developed to expand the EN method by adjusting for error and estimating the underlying true score as opposed to using observed scores (Wise, 2004). According to Bauer, Lambert, and Nielsen (2004), the HA method involves the most significant changes to the JT method; it provides a method that allows a differential analysis of clinically meaningful change at an individual level and group level. The two levels are identified as follows: an individual level dictates which classification group the score falls within, and a group level looks at the proportion of improvement estimations (Jacobson et al., 1999). The HA method uses reliability of the difference scores ($r_{dd}$), which compares an individual’s change to the change of the group to which the individual belongs. This procedure may give greater precision to the estimate by requiring that the reliability of difference score reach a minimum threshold (Marsden et al., 2010). This method has some benefits, including that it allows researchers to present cumulative results for group based analyses in addition to individual based analyses (Atkins et al., 2005).

Comparison of Methods

Statistical controversies surrounding the RCI have had a long existence in empirical study history. Spear and Greenbaum (1995) presented one of the first empirical evaluations of different statistical methods for computing reliable change. In this study the authors proposed a fifth method, HLM, which represents a growth curve approach to clinically significant change. Spear and Greenbaum used data from 3-5 self-report measurement tools from 73 older adults in an outpatient setting to compare five different methods: JT, GLN (then called HLL), EN, Nunnally-Kotsch (NK), and HLM. Overall, the HLM and EN methods produced significantly
higher improvement rates than the HLL and NK methods, but did not produce higher rates than the JT method. With exception of the JT method, the results indicated that the HLM method resulted in the highest improvement rate and lowest deterioration rate. The EN method resulted in both the second highest improvement and the second highest deterioration rate. Spear and Greenbaum concluded that when a large enough sample size is used, the HLM method is recommended. In situations where the HLM method is unobtainable, the authors suggested the use of the JT method for the following reasons: problems associated with residualized true score adjustments are avoided, the method is straightforward, and the JT method at this time was the only method used in published literature.

McGlinchey and Jacobson (1999) compared the HA method to the JT method using the same data that Jacobson and Traux (1991) used in support of the JT method. This data set consisted of couples in marital therapy treatment using the outcome measurements the Dyadic Adjustment Scale (DAS; Spanier, 1976) and the global distress scale of the Marital Satisfaction Inventory (GDS; Snyder, 1979). The HA method, being relatively new, was also evaluated by McGlinchey and Jacobson (1999) to discuss its implications in future research; they commended it, stating that “Hageman and Arrindell have provided valuable insights in the search for obtaining increased precision in clinical significance, though clearly more work needs to be done” (p. 1216). The authors concluded that both the JT and HA methods produced similar results. However, the authors recommended the JT method as the preferred method because of the complexity of the computations used in the HA method.

McGlinchey, Atkins, and Jacobson (2002) (as cited in Wise, 2004) compared five different RCI methods for determining clinical significance. Four RCI methods used in Speer and Greenbaum (1995), JT, HLL, EN, and NK, were compared. The authors also included the
HA method previously studied by McGlinchey and Jacobson (1999). The authors also expanded the original methodology and evaluated the methods for accuracy in prediction with respect to relapse; each method predicted this equally well. The authors found that the HA method was the most conservative and, therefore, had the least individuals classified as Recovered. Also, the authors indicated that the EN method had a high rate of change. McGlinchey et al. (2002) concluded that the JT method has not been improved upon further by any additional method; the authors pronounced support for the JT method.

Bauer, Lambert, and Nielsen (2004) aimed to compare five different methods used to calculate clinically significant change as well as increase the understanding and the impact of the different proposed methods of clinically significant change. The authors studied 386 outpatients who were in treatment that included routine clinical practice in a student-training center. Results indicated that significant differences existed between all methods except GLN and JT. The HLM method was found to produce relatively low clinically significant change and the EN method was found to provide the most liberal estimates for clinically significant change. The HA method characterized more individuals as deteriorated than any other method and was especially unlikely to categorize many individuals as improved. In conclusion, three of the five methods (GLN, JT and HLM) were comparable and two (EN and HA) differed from the rest. The authors claimed that the JT method provided a moderate estimation of meaningful change; they also recommended that the JT method should be used in outcome studies.

The most recent study to date comparing clinical significance methods used a simulation to obtain data. Atkins Bedics, McGlinchey, and Beauchaine (2005) aimed to assist researchers in the field of psychotherapy to make well-informed choices about what method of clinical significance to use in research. The authors were not attempting to determine which method was
better or more accurate, but instead were aiming to provide an understanding of how the methods differ in their classification groupings. Four methods were compared, including the JT, GLN, HN, and EN methods. The simulation data included pre-therapy and post-therapy estimates generated from several clinical trials. Reliabilities ranged from .60 to .95 and Cohen’s $d$ effect sizes ranged from .10 to 1.00. Results indicated considerable agreement between methods, which increased as the reliability of the measurement increased. At the reliability level most used in psychotherapy research, .90, very few differences were found between methods. Differences found at this level included the EN method being the most “certain,” resulting in the smallest SE, and the HA method being the most complex and conservative of the methods, resulting in the fewest recovered cases. The authors reported that the JT and GLN methods resulted in nearly identical classifications.

**Purpose of Current Study**

The current literature shows that there is still a general disagreement about which method to calculate clinical significance is preferred. Although newer methods attempt to be more accurate by controlling for error, such as regression to the mean, the results throughout the literature indicate that the newer methods are similar to the more original and simpler approach, the JT method. The current study was a replication of Bauer, Lambert, and Nielson’s (2004) ‘Clinical Significance Methods: A Comparison of Statistical Techniques’ and used a similar student training clinic population. The aim of the current study was to evaluate and compare difference methods of clinically significant reliable change. Also, this study aimed to increase clinicians’ understanding of the impact of the different methods of clinically significant change. Though the literature differed for which method was preferred, the hypotheses of the current study reflect thematic findings in the literature. It was hypothesized that the GLN method would
be most similar to the JT method than the EN and HA methods. Also, the EN method would provide the most lenient results meaning that this method would result in a higher rate of improvement and a higher rate of deterioration. It was hypothesized that there would be a general agreement between all four methods with minor differences in improved and deteriorated rates.

**Method**

**Participants**

This study utilized data from 395 clients who had previously sought treatment at a university-based student-training clinic in the US Northwest. Clients ranged in age from 8 to 75 years (M= 33 SD=11.4). Clients’ sex included female (60.8%), male (37%), and transgender (<1%). The majority of clients identified as white (84.3%), with other self-identified race or ethnicities including biracial (4.3%), Mexican (4.3%), and Black or African American (1.3%). A small number of clients identified as disabled (6.8%). At intake 64% of clients were diagnosed by the attending clinician using the Diagnostic and Statistical Manual of Mental Disorders 4th edition (DSM – APA, 1994), while 36% received a deferred diagnosis or never had a diagnosis entered into the database. Of those clients receiving a diagnosis, 35.5% were given a mood disorder diagnosis, 20.6% were diagnosed with a V code, 19.5% were given an anxiety disorder diagnosis, 8.4% were diagnosed with adjustment disorder, 7.6% were given a PTSD diagnosis, and the remainder were diagnosed with other conditions. The number of sessions attended ranged from 2 sessions to 86 sessions, with an average of 11 sessions. The number of treatment weeks ranged from 2 weeks to 189 weeks, with an average of 16 treatment weeks. The clients in this study started treatment less disturbed (OQ-45.2 M = 74.45 SD = 24.67) than those in routine
outpatient care (M=80.98, SD=24.34) but had scores similar to other university counseling centers according to Bauer et al. (2004).

Therapists included psychology interns and student clinicians attending a doctoral psychology program who were supervised by licensed doctoral level psychologists on a weekly basis. Therapists endorsed a variety of treatment orientations and content approaches including Cognitive Behavioral Therapy (30.1%), Psychodynamic (14.9%), Gestalt (12.9%), Alcohol and Drug (11.6%), integrative (11.6%), Behavioral (7.6%), Dialectical Behavior Therapy (2%), mindfulness/Acceptance and Commitment Therapy (1.5%), couples counseling (1%), Latino/Bilingual (.5%), and other (.3%). Psychotherapy was offered in both Spanish and English to the community on a sliding scale fee schedule and free to certain contracted university student clients or veterans. Duration of therapy was determined by the needs and preferences of the clients as well as based on the therapist’s and supervisor’s judgment.

**Outcome Measure**

Psychological distress was assessed using the Outcome Questionnaire (OQ-45.2; Lambert et al., 2004). This scale was created and refined for use in a clinical setting. It has become essential, especially in clinical settings like a training center, because of the efficiency and effectiveness it demonstrates (Lambert et al., 2004). It entails 45 self-report questions that measure a person’s subjective experience as well as how they function in the world using a Likert scale. In addition to a Total score, the OQ-45.2 yields three subscale scores: Symptom Distress, Interpersonal Relations, and Social Roles. For the purpose of this study, the total score on the first visit (pretreatment) and the final visit (posttreatment) of each individual were the values used to assess change over the course of treatment. Normative statistics have been found for a number of settings the OQ-45.2 is used in including community mental health settings,
employee assistance programs, inpatient, and outpatient. It is also used for a variety of purposes including as a screening tool, to inform treatment decisions and recommendations, and to monitor progress and regression. Internal consistency, Chronbach’s alpha, used in this study was .93; this is the same value used in the study by Bauer et al. (2004). According to the manual, a cut off total score of 63 should be used to distinguish between clinical and nonclinical samples; if an individual scores 63 or below they are considered part of the community, or nonclinical, sample. For the purpose of this study the OQ-45.2 provided the measure of change as well as determined the classification of patients into outcome groups (Recovered, Improved, No Change, Deteriorated). Clients completed the OQ-45.2 before each scheduled therapy session.

**Procedure**

OQ-45.2 scores from 395 clients who had at least one therapy session in addition to pretest and posttest scores were used to compare four methods for determining clinically significant change. All of the methods assume continuous data. For the purpose of this study, four methods, HA, GLN, EN, and JT use the pretest and posttest scores exclusively; all have been previously examined in different studies. The HLM method was left out due to insufficient information about additional data points. Formulas, additional details, and a detailed calculation demonstrating the RCI can be found in the Appendix.

The average OQ-45.2 pretest score was 74.45 (SD = 24.67). The average posttest score for the sample was 60.13 (SD = 27.73). Following the procedure of Bauer et al. (2004), Cutoff C was used for calculation of clinical significance. This cutoff is based on information about both the clinical and nonclinical samples. The original OQ 45.2 cut off score (63) was found in the manual. Because our mean pretest score (M=74.45) differed from the mean pretest score in the manual (M=79.8), cut off C was modified. Using the current study’s pretest data and the
The calculation of $C=60$ was done as follows:

$$C_{\text{utoff}} = \left( \frac{SD_{\text{clinical}} M_{\text{nonclinical}}} {SD_{\text{clinical}}} \right) + \left( \frac{SD_{\text{nonclinical}} M_{\text{clinical}}} {SD_{\text{nonclinical}}} \right)$$

The Reliable Change Index (RCI) was calculated for all participants using the JT, GLN, EN, and HA methods as described above. Based on this data, participants were categorized into Improved (those who evidenced reliable change in a positive direction), unchanged (those who did not evidence reliable change), and Deteriorated (those who evidenced reliable change in a negative direction). Table 1 shows the overall rates of classifying reliable change for the four methods for the sample of 395 individuals.

<table>
<thead>
<tr>
<th>Method</th>
<th>Deteriorated</th>
<th>Unchanged</th>
<th>Improved</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>% Frequency</td>
<td>% Frequency</td>
<td>% Frequency</td>
</tr>
<tr>
<td>Jacobson-Truax</td>
<td>4.8 19</td>
<td>57.5 227</td>
<td>37.7 149</td>
</tr>
<tr>
<td>Gulliksen-Lord-Novick</td>
<td>4.8 19</td>
<td>54.7 216</td>
<td>40.5 160</td>
</tr>
<tr>
<td>Edwards-Nunnally</td>
<td>6.3 25</td>
<td>44.8 177</td>
<td>48.9 193</td>
</tr>
<tr>
<td>Hageman-Arrindell</td>
<td>4.1 16</td>
<td>48.1 190</td>
<td>47.8 189</td>
</tr>
</tbody>
</table>

*Note: N=395*

Consistent with Bauer, Lambert, and Nielson (2004), and as hypothesized, the EN method classified the least number of individuals as reliably unchanged (44.8%), the most individuals as reliably improved (48.9%), and the most as reliably deteriorated (6.3%). Again consistent with Bauer et al. (2004), the JT method and GLN method resulted in the most similar percentages across the three categories. However, this study differed from Bauer et al.’s (2004) results in other ways, with the JT and GLN method categorizing the most individuals as reliably
unchanged (57.5% and 54.7%, respectively) and the least number of individuals as reliably improved (37.7% and 40.5%, respectively). Bauer et al. (2004) found that the HA method categorized the most individuals as unchanged. In contrast to Bauer et al. (2004), this study found that the HA method classified the least number of individuals as reliably deteriorated (4.1%); Bauer et al. (2004) found the HA method had the smallest number improved.

Table 2
Paired comparisons Between Reliable Change Classifications

<table>
<thead>
<tr>
<th>Method</th>
<th>JT</th>
<th>GLN</th>
<th>EN</th>
<th>HA</th>
</tr>
</thead>
<tbody>
<tr>
<td>JT</td>
<td>-</td>
<td>.95</td>
<td>.77</td>
<td>.80</td>
</tr>
<tr>
<td>GLN</td>
<td>p=.001</td>
<td>-</td>
<td>.82</td>
<td>.85</td>
</tr>
<tr>
<td>EN</td>
<td>p&lt;.001</td>
<td>p&lt;.001</td>
<td>-</td>
<td>.88</td>
</tr>
<tr>
<td>HA</td>
<td>p&lt;.001</td>
<td>p&lt;.001</td>
<td>ns</td>
<td>-</td>
</tr>
</tbody>
</table>

Note: Significance levels for paired comparisons (Wilcoxon test) between the reliable change classification of the four approaches are in the lower left quadrant. Kappa coefficients for the agreement between methods are in the right upper quadrant. JT= Jacobson-Truax; GLN=Gulliksen-Lord-Novick; EN = Edwards-Nunnally; HA= Hageman-Arrindell.

Kendall’s coefficient of concordance statistic ($W$) was computed to compare differences in classification rates between all four methods. This omnibus test is distributed as a chi-square. $W$ was highly significant ($W = .080), \chi^2(3, N = 395) = 94.40, p < .001. Pairwise comparisons of all methods were calculated using the signed ranks Wilcoxon test. This nonparametric test compared each method against each of the others separately to measure the level of agreement between the methods. The results are presented in lower left section of Table 2. Statistically significant differences at the specified .001 level were found between 5 of the 6 pairs. The methods that did not significantly differ from each other included the HA and EN methods.

To quantify the extent of agreement across methods, Cohen’s kappa coefficient ($\kappa$) was calculated for each combination of two methods. Table 2 shows the results in the right upper quadrant. All 6 coefficients reached statistical significance at the .001 level. Similar to Bauer et al. (2004) and as hypothesized, the highest agreement was found between the JT and GLN methods ($\kappa=.95$). In contrast to Bauer et al. (2004) who found the lowest level of agreement
between the EN and HA methods, this study found that the lowest level of agreement was found between EN and JT methods ($\kappa = .77$).

Table 3

<table>
<thead>
<tr>
<th>Method</th>
<th>CS Deteriorated</th>
<th>Deteriorated</th>
<th>Unchanged</th>
<th>Improved</th>
<th>Recovered</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>% Frequency</td>
<td>% Frequency</td>
<td>% Frequency</td>
<td>% Frequency</td>
<td>% Frequency</td>
</tr>
<tr>
<td>JT</td>
<td>1.8</td>
<td>7</td>
<td>3.0</td>
<td>12</td>
<td>57.5</td>
</tr>
<tr>
<td>GLN</td>
<td>1.8</td>
<td>7</td>
<td>3.0</td>
<td>12</td>
<td>54.7</td>
</tr>
<tr>
<td>EN</td>
<td>2.0</td>
<td>8</td>
<td>4.3</td>
<td>17</td>
<td>44.8</td>
</tr>
<tr>
<td>HA</td>
<td>1.8</td>
<td>7</td>
<td>2.3</td>
<td>9</td>
<td>48.1</td>
</tr>
</tbody>
</table>


Table 3 shows the overall rates of classifying the client’s clinically significant change for the four methods for the sample of 395 individuals. Similar to Bauer et al. (2004), the EN method classified the most amount of individuals as recovered (35.2%) and the HA method classified the least amount of individuals as recovered (23.3%). Kendall’s coefficient of concordance statistic ($W$) was again computed to compare differences in classification rates. This omnibus test is distributed as a chi-square. $W$ was highly significant ($W=.056$), $\chi^2(3, N=395) = 66.76, p < .001$.

Table 4

<table>
<thead>
<tr>
<th>Method</th>
<th>JT</th>
<th>GLN</th>
<th>EN</th>
<th>HA</th>
</tr>
</thead>
<tbody>
<tr>
<td>QT</td>
<td>-</td>
<td>.95</td>
<td>.80</td>
<td>.70</td>
</tr>
<tr>
<td>GLN</td>
<td>ns</td>
<td>-</td>
<td>.84</td>
<td>.74</td>
</tr>
<tr>
<td>EN</td>
<td>p&lt;.001</td>
<td>ns</td>
<td>-</td>
<td>.75</td>
</tr>
<tr>
<td>HA</td>
<td>ns</td>
<td>ns</td>
<td>p&lt;.001</td>
<td>-</td>
</tr>
</tbody>
</table>

Note: Significance levels for paired comparisons (Wilcoxon test) between the clinically significant change classification of the four approaches are in the lower left quadrant. Kappa coefficients for the agreement between methods are in the right upper quadrant. JT= Jacobson-Truax; GLN=Gulliksen-Lord-Novick; EN = Edwards-Nunnally; HA= Hageman-Arrindell.

Table 4 shows pairwise comparisons using the signed ranks Wilcoxon test, which showed statistically significant differences between three of the six total combinations of methods.
excluding: JT and GLN, JT and HA, & GLN and HA. To quantify the extent of agreement across methods for clinical significance, Cohen’s kappa coefficient (κ) was calculated for each combination of two methods. Table 4 shows the results in the right upper quadrant. All 6 coefficients reached statistical significance at the .001 level. Similar to Bauer et al. (2004), this study found the highest agreement between the JT and GLN methods (κ = .95). The lowest level of agreement was found between HA and JT methods (κ = .70), whereas Bauer et al. (2004) found the lowest level of agreement between the EN and HA methods.

Table 5

<table>
<thead>
<tr>
<th>Method</th>
<th>JT</th>
<th>GLN</th>
<th>EN</th>
<th>HA</th>
</tr>
</thead>
<tbody>
<tr>
<td>JT</td>
<td>-</td>
<td>95.6</td>
<td>73.4</td>
<td>74.6</td>
</tr>
<tr>
<td>GLN</td>
<td>95.3</td>
<td>-</td>
<td>82.7</td>
<td>73.1</td>
</tr>
<tr>
<td>EN</td>
<td>79.6</td>
<td>84.1</td>
<td>-</td>
<td>60.9</td>
</tr>
<tr>
<td>HA</td>
<td>70.0</td>
<td>73.9</td>
<td>75.4</td>
<td>-</td>
</tr>
<tr>
<td>Average Agreement</td>
<td>81.6</td>
<td>84.4</td>
<td>79.7</td>
<td>73.1</td>
</tr>
<tr>
<td>Bauer et al. (2004)</td>
<td>81.2</td>
<td>83.8</td>
<td>72.3</td>
<td>69.5</td>
</tr>
</tbody>
</table>

Note: Results from the current study are underlined. The results in the upper right quadrant and bottom row are from Bauer, Lambert, and Nielson (2004). Bauer et al. (2004) average agreement was recalculated without including the statistics from the HLM method. JT= Jacobson-Traux; GLN=Gulliksen-Lord-Novick; EN = Edwards-Nunnally; HA= Hageman-Arrindell.

Table 5 shows the rates of agreement between the different methods in addition to the average rate of agreement for each when method compared to the other three individually. The current study shows the highest agreement between the JT and GLN methods (95%); this is consistent with Bauer et al. (2004) who also found the highest rate of agreement between the JT and GLN methods (96%). In the current study the HA method had the lowest average agreement (73%) in addition to the largest range of agreement between the HA method and any one other method (70%-75%). Again, this was consistent with Bauer et al. (2004) who found the HA method had the lowest average rate of agreement (70%) and largest range (61%-75%).
compared to Bauer et al. (2004), the current study present nearly identical average rates of agreement for the JT method (81% and 82%, respectively).

Discussion

Throughout the history of psychological research there have been numerous debates about the meaning of reliable change as well as what constitutes true change. In other words, the importance of clinical significance and reliable change has existed in the field of psychology for quite some time. Ogles, Lunnen, and Bonesteel (2001) looked into the history of clinical significance and suggested that the “earliest studies of psychotherapy implicitly focused on demonstrating that clients made clinically meaningful change” (p 422). The literature showed that researchers initially looked at group change; over the years and throughout the literature, there was a shift from looking at group change, or statistical significance, to looking at individual change, also known as clinical significance (Ogles et al., 2001). This is similar to Bauer et al. (2004) where the authors mentioned examining classification frequency differences between methods as an important aspect of psychological research. The authors continued saying “estimates of clinically significant improvement for groups of patients affect the degree to which treatments are generally considered to be effective or in need of modification.” (p. 66). This not only impacts the individual clients, but can influence researchers, clinicians’ practice, and policy makers, as well as how and when recommendations are used (Bauer et al., 2004). Using a reliable assessment measure as well as categorizing the client based on the test data as improved, unchanged, and deteriorated, allow a clinician to monitor client progress; thus determining if the proper course or intensity of treatment is having a positive (improved), adverse (deteriorated), or little (unchanged) effect. These categories, as demonstrated by this study, depend on the statistical methods used to calculate clinically significant change.
There are important aspects to determining clinically significant change as Kazdin (1999) points out; treatment goals are typically based off of clinical significance measures. Hinton-Bayre (2010) stated that monitoring change in test scores over time could be useful for different reasons including monitoring an individual’s state after an injury or insult, or progress in response to an intervention. Along the same lines, Maassen (2000) stated that progress or deterioration could be due to a fit or mismatch between the client and the particular treatment modality in addition to a number of factors in the individual’s functioning. Hsu (1999b) argues an opposing viewpoint of Maassen (2000) with this statement:

*The conclusion that a change is clinically significant (as determined by the JT, HA, or GLN methods) is therefore by no means equivalent to the conclusion that is it the psychotherapy that caused this change, and the conclusion that a therapy is efficacious or effective, which implies that the psychotherapy caused the change, can in general NOT be justified by the empirical finding of large proportions of clinically significant client changes.* (p. 1201)

It is clear that there is disagreement about the methods for determining clinical significance throughout psychological research; this disagreement is not limited to the meaning of clinical significance but also includes what the current study addresses: which method of reliable change and clinical significance most accurately represents change that is actually taking place within the client.

In this study I attempted to increase clinicians’ understanding of the impact of the different methods of clinically significant change by comparing and contrasting rates of change associated with different methods of calculating clinically significant change. Another advantage to the current study is it provided information on client classification rates not
otherwise evaluated routinely in a student-training center. Like in Bauer et al. (2004), a large data set was used; this differed from studies such as Speer and Greenbaum (1995), McGlinchey, Atkins, and Jacobson (2002), and McGlinchey & Jacobson (1999) and allowed for a greater possibility of finding differences if they existed (Bauer et al., 2004). The results across the past studies provide examples of differences that may arise due to differences in statistical methods used for reliable change and clinical significance.

Results from this study indicate that differences in estimates of clinically significant change exist between all methods except the JT and GLN methods, JT and HA methods, and HA and GLN methods. Discrepancies between the current study and Bauer et al. (2004) for the HA method will be discussed in greater detail in the following paragraph. In terms of individual classification categories (unchanged, improved, etc…), as hypothesized, the JT and GLN method appear the most similar in regard to the rate of classification of clinically significant change. In fact, these two methods found the same percent of individuals deteriorated (3%) as well as were within less than three percent of each other for unchanged (58% & 55%, respectively), improved (10% & 12%, respectively), and recovered (27% & 29%, respectively). Bauer et al. (2004) found 96% agreement and McGlinchey et al. (2002) was reported as having over 90% agreement between the JT and GLN method (Bauer et al., 2004); similarly, the current study found 95% agreement between these two methods.

Contrary to Atkins et al. (2005), Bauer et al. (2004), and McGlinchey (2002), who found that the HA method had the lowest classifications of improvement, the current study found that the HA method had the highest percentages of individuals classified as improved (25%), but the least percentage of individuals classified as recovered (23%). McGlinchey and Jacobson (1999) had a similar experience when comparing the JT method and the HA method; like the current
study, the authors reported that those two methods produced similar results. While computing the clinical significance for the HA method I found a discrepancy between the equation for individual clinical significance ($CS_{INDIV}$) provided in Hageman and Arrindell (1999b; p 1176) and the equation provided in the appendix of Bauer et al. (2004, p 70). Specifically, the denominators are different; for the purpose of the current study the equation in Hageman and Arrindell (1999b) was used (see Appendix). Due to this discrepancy, the results of the HA method beyond the reliable change (Table 1 and Table 2) should be interpreted with caution.

In the current study, as hypothesized, the EN method resulted in the largest number of individuals classified as clinically significantly deteriorated (4.3%), clinically significantly recovered (35.2%), and reliably improved (48.9%). This is similar to Bauer et al. (2004) who stated, “The EN method provided the most liberal estimates of clinically significant change and/or reliable change… (21%)” (p 67). A study by Speer and Greenbaum (1995) showed that the EN method resulted in the second highest level of improvement and deterioration. Similarly McGlinchey et al. (2002) found that use of the EN method resulted in a high rate of change. Interestingly, although the classification rates of the EN method were similar between the current study and Bauer et al. (2004), the average agreement for the EN method in the current study (80%) was the most different when comparing average agreement for all four methods with that of Bauer et al. (2004; 72%).

Further research is needed in order to explore the validity of clinically significant change methods. This study is limited to simply showing how the four methods differ for this one data set; it does not allow for an evaluation as to which method is more accurate in representing meaningful change. Bauer et al. (2004) stated, “validity data are needed to evaluate which statistical estimate most accurately reflects meaningful change” (p 68). Studies such as the
current one are often difficult because there is no ‘gold standard’ in terms of what method is considered to be the best. Along the same lines, Hinton-Bayer (2010) stated that there is no one method that is “universally more sensitive or conservative” and the classification bias will be dependant on the individual case and the nature of the data collection (p 252). Studies that couple both clinically significant change and client satisfaction may be the future direction for studies looking to find the best method for classifying reliable and clinically significant change.

Overall, as predicted, the four methods provided comparable results for both reliable change and clinically significant change. Also, this study had higher rates of agreement between the four methods than Bauer et al. (2004). In the current study the two methods that resulted in the most similar classification frequencies were the JT and GLN methods. This study found different results for the HA method than authors of previous studies mentioned including Speer and Greenbaum (1995) and McGlinchey et al. (2002); discrepancies in the denominator of the equation may be the reason for these differences. Trends found in the literature that were continued in the current study included the characterization of the EN method as the most liberal method, which provided the highest number of individuals classified as reliably changed, with or without or clinical significance. Due to a substantial history found in the literature across disciplines, the relatively straightforward computations required, and overall ease of classifying individual using the results Bauer et al. (2004) concluded that the JT method should be the preferred method. Also, Bauer et al. (2004) stated that the JT method is already widely known, and “cutoff estimates are available for a number of widely used instruments” (p 68). The current study agrees with the body of literature and authors like Bauer et al. (2004), McGlinchey and Jacobson (1999), Maassen (2004), and Jacobson et al. (2002) who also endorsed the JT method as the most preferred method.
Acknowledgements

This research was possible because of data collected over time from the Pacific Psychology Clinic via Clinic Director Dr. Lisa Christiansen, PsyD. Also, I would like to thank Dr. Paul Michael, PhD. for help with statistical analysis.
References


“classical approach” leads to poor estimates. *Journal of the International Society.* 10,
899-901 DOI: 10.10170S1355617704106115


significance, reliable change and recommendations for future directions. *Journal of
Personality Assessment* 82(1) 50-59.
Appendix

Information relevant for one example calculation

Information based on one individual:
Pretreatment OQ 45.2 score $X_{\text{pre}} = 58$
Posttreatment OQ 45.2 score $X_{\text{post}} = 11$

Information based on the Sample:
Pretreatment Mean= $M_{\text{pre}} = 74.45$
Pretreatment Standard Deviation= $SD_{\text{pre}} = 24.67$
Posttreatment Mean= $M_{\text{post}} = 60.13$
Posttreatment Standard Deviation = $SD_{\text{post}} = 27.73$
Internal Consistency: Cronbach’s $\alpha = .93$
Standard error of measurement ($S_E$): $S_E = \sqrt{1 - r_x} = 6.527$
Correlation between pretreatment and posttreatment scores $r_{\text{pre*post}} = .672$

Jacobson and Truax (1991)

RCI:

\[
\frac{(X_{\text{post}} - X_{\text{pre}})}{\sqrt{2}S_E^2} = \frac{(11 - 58)}{\sqrt{2} \times 6.527^2} = -5.09 
\]

The patient is classified as reliably improved because the score is smaller than -1.96

Gulliksen, Lord, Novick (Hsu, 1989)

RCI:

\[
\frac{(X_{\text{post}} - M_{\text{pre}}) - r_{xx}(X_{\text{pre}} - M_{\text{pre}})}{SD_{\text{pre}} \sqrt{1 - r_{xx}^2}} = \frac{(11 - 74.45) - .93(58 - 74.45)}{24.67 \sqrt{1 - .93^2}} = -5.18 
\]

The patient is classified as reliably improved because the score is smaller than -1.96

Edwards and Nunnally (Speer, 1992)

RCI:

\[
[r_{xx}(X_{\text{pre}} - M_{\text{pre}}) + M_{\text{pre}}] \pm 2SD_{\text{pre}} \sqrt{1 - r_{xx}} 
\]

\[
= [.93(58 - 74.45) + 74.45] \pm 2 \times 24.67 \sqrt{1 - .93}, 
\]

Where the upper boundary = 59.15 + 13.05 = 72.21 and the lower boundary = 59.15 – 13.05 = 46.10.

The patient is classified as reliably improved because the posttreatment score ($X_{\text{post}} = 11$) is below the lower boundary.

Hageman and Arrindell (1999b)

Individual RCI:
\[ (X_{\text{post}} - X_{\text{pre}}) r_{dd} + (M_{\text{post}} - M_{\text{pre}})(1 - r_{dd}) = \sqrt{r_{dd}} \sqrt{2 S_E^2} \]

\[ (11 - 58).806 + (60.13 - 74.45)(1 - .806) = -4.91 \]

This individual is classified as reliably improved because the score is smaller than -1.65.

\[ r_{dd} = \frac{S D_{\text{pre}}^2 \times r_{xx(\text{pre})} + S D_{\text{post}}^2 \times r_{xx(\text{post})} - 2 S D_{\text{pre}} S D_{\text{post}} r_{xx^*} r_{\text{pre}^* \text{post}}}{S D_{\text{pre}}^2 + S D_{\text{post}}^2 - 2 S D_{\text{pre}} S D_{\text{post}} r_{xx^*} r_{\text{pre}^* \text{post}}} \]

\[ \frac{24.67^2 \times .93 + 27.73^2 \times .94 - 2 \times 24.67 \times 27.73 \times .672}{24.67^2 + 27.73^2 - 2 \times 24.67 \times 27.73 \times .672} = 0.806 \]

\[ r_{xx(\text{pre})} = \frac{S D_{\text{pre}}^2 - S_E^2}{S D_{\text{pre}}^2} = \frac{24.67^2 - 6.527^2}{24.67^2} = .93 \]

\[ r_{xx(\text{post})} = \frac{S D_{\text{post}}^2 - S_E^2}{S D_{\text{post}}^2} = \frac{27.73^2 - 6.527^2}{23.73^2} = .94 \]

Individual clinical significance index:

\[ M_{\text{post}} + (X_{\text{post}} - M_{\text{post}}) r_{xx(\text{post})} - TRC \]

\[ \sqrt{r_{xx(\text{post})} S_E} \]

\[ TRC = \frac{S D_{\text{norm}} \sqrt{r_{xx(\text{norm})} M_{\text{pre}} + S D_{\text{pre}} \sqrt{r_{xx(\text{pre})} M_{\text{norm}}}}}{S D_{\text{norm}} \sqrt{r_{xx(\text{norm})} + S D_{\text{pre}} \sqrt{r_{xx(\text{pre})}}}} \]