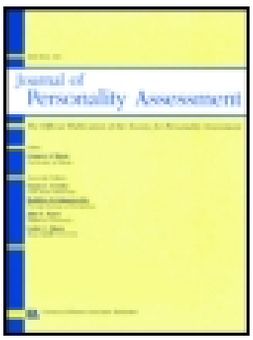

Routledge
Taylor & Francis Group

Journal of Personality Assessment

ISSN: 0022-3891 (Print) 1532-7752 (Online) Journal homepage: https://www.tandfonline.com/loi/hjpa20

Methods for Analyzing Psychotherapy Outcomes: A Review of Clinical Significance, Reliable Change, and Recommendations for Future Directions

Edward A. Wise

To cite this article: Edward A. Wise (2004) Methods for Analyzing Psychotherapy Outcomes: A Review of Clinical Significance, Reliable Change, and Recommendations for Future Directions, Journal of Personality Assessment, 82:1, 50-59, DOI: 10.1207/s15327752jpa8201_10

To link to this article: https://doi.org/10.1207/s15327752jpa8201_10

JOURNAL OF PERSONALITY ASSESSMENT, *82*(1), 50–59

Methods for Analyzing Psychotherapy Outcomes: A Review of Clinical Significance, Reliable Change, and Recommendations for Future Directions

Edward A. Wise
Mental Health Resources
Memphis, Tennessee

The purpose of this article is to provide a brief review of the history, development, and current status of the concepts of clinical significance (CS) and the reliable change index (RCI). I address issues regarding the development, criticisms, and applications of CS and RCI. I review the use of normative data, cutoff points, formula adjustments, and the comparative validity of various RCI methods. An examination of the convergence of multiple domains and multiple measures demonstrates ways to further develop the concepts of reliable change and CS. Finally, I make some recommendations and implications for future research and the development of assessment tools.

If pretest and posttest treatment scores are statistically significantly different, what does this tell program evaluators about how the individual fared in the treatment? Do patients have to exceed $p < .05$ for them to have obtained a "clinically significant" change? For that matter, what is a clinically significant change? Jacobson, Follette, and Revenstorf (1984) identified several problems interpreting traditional psychotherapy outcome studies. These include the use of pretest and posttest group scores that provide no information about treatment effects for specific individuals, the use of statistical significance tests that have little clinical relevance, the inability to determine the proportion of treated individuals that improve, and the fact that clinical significance (CS) is not consistently defined or utilized as an adjunct to statistical significance testing. Jacobson, Follette, and Revenstorf (1984) called for "agreed upon conventions for determining what constitutes improvement … [and] consensus as to what is meant by clinical significance" (pp. 338–339).

CS

Jacobson and Truax (1991) discussed the problem of differentiating statistical from CS and stated that "judgments regarding clinical significance are based on external standards provided by interested parties in the communities. … The clinical significance of a treatment refers to its ability to meet standards of efficacy set by consumers, clinicians and researchers" (p. 12). Jacobson, Follette, and Revenstorf (1984) "propose[d] that a change in therapy is clinically significant when the client moves from the dysfunctional to the functional range during the course of therapy" (p. 340), thereby providing a definition of CS. Jacobson, Follette, and Revenstorf (1984) operationalized the dysfunctional–functional continuum as the following:

> 1. Does the level of functioning at post-test fall outside the range of the dysfunctional population, where range is defined as extending to 2 standard deviations above (in the direction of functionality) the mean for that population? 2. Does the level of functioning at post-test fall within the range of the functional or normal population, where range is defined as beginning at two standard deviations below the mean for the normal population? 3. Does the level of functioning at post-test suggest that the subject is statistically more likely to be in the functional than in the dysfunctional population; that is, is the post-test score statistically more likely to be drawn from the functional than the dysfunctional distribution? (p. 340)

Jacobson and Truax further clarified the choice of norms by stating the following:

> First, when norms are available, either 2 or 3 [as defined previously] is often preferable to 1 as a cutoff point: In choosing between 2 and 3, when functional and dysfunctional populations overlap, 3 is preferable to 2; but when the distributions are nonoverlapping, 2 is the cutoff point of choice. When norms are not available, 1 is the only cutoff point available: To avoid the problem of different cutoff points from study to

> study, 1 should be standardized by aggregating samples from study to study so that dysfunctional norms can be established. (p. 14)

A normative change meeting either criterion 1, 2, or 3 would be considered CS. (An example of this procedure may be found in the Appendix.)

In calculating these cutoff points, Jacobson, Follette, and Revenstorf (1984) envisioned the use of both functional and dysfunctional distributions. Although Jacobson, Follette, and Revenstorf proposed a return to the functional distribution as signifying CS, they also acknowledged that "it is not always possible to identify an appropriate normative group" (p. 342) and that in some circumstances the return to normal criterion may be too stringent. For example, individuals treated in an inpatient psychiatric unit would not reasonably be expected to return to the normal distribution on completing stabilization, although they may experience enough change to be safely discharged. In continuing to develop the concept, Jacobson and Truax (1991) also suggested using confidence bands around the CS cutoff points to further reduce measurement error.

CRITICISMS OF CS

Wampold and Jenson (1986) were the first to raise the point that

> The assumption of two distinct distributions may be appropriate for some clinical problems … [but] a viable alternative assumption is that there is one population, which contains the functional and dysfunctional individuals, and the dysfunctional individuals are found in one of the tails of the distribution of scores for that population. (pp. 302–303).

Hsu (1996) also criticized the derivation of the cutoff scores, stating that the "cut scores provide no information about the probability that a client belongs to a Functional or Dysfunctional group" (p. 373) because the two standard deviations from each of these groups are unequal. Hence, the arithmetical mean between these scores might not be the true cut point differentiating one sample from the other. Kendall and Grove (1988) proposed that viewing symptoms on a continuum rather than as a bimodal distribution would solve the distribution problem. However, Jacobson, Follette, and Revenstorf (1986) argued that irrespective of the population distribution, there were two samples of interest: those seeking treatment and those who do not. Jacobson et al. stated, "As long as two distinguishable groups exist, it is possible to identify a point at which an individual is equally likely to be a member of either group. These are the necessary and sufficient conditions for establishing the cutoff points" (p. 309).

Tingey, Lambert, Burlingame, and Hansen (1996a) reported similar problems regarding the use of normative samples in defining CS. Tingey et al. noted

> 1) an inability to identify and use relevant normative samples across studies; 2) the restriction of the social validation methodology by the use of only one dysfunctional and one functional sample; and 3) the lack of a procedure to determine the distinctness of samples (p. 110).

Rather than use a return to normal criterion, Tingey et al. proposed the innovative use of multiple samples "organized around a rational or empirical continuum … corresponding [to] low to high levels" (p. 111) of the variable being measured. Establishing a functional to dysfunctional continuum addressed Wampold and Jenson's (1986) concerns regarding the assumption of a bimodal distribution. Using this method, to be classified as CS, an individual must move from one category on the normative continuum to another but not necessarily return to normal. Tingey et al. used the Jacobson and Truax (1991) formula for determining cutoff points to be used in defining adjacent samples:

$$Cutoff = \frac{(s_1 \times \bar{X}_2) + (s_2 \times \bar{X}_1)}{s_1 + s_2},$$

where X_1, s_1, X_2, and s_2 specify the means and standard deviations of two different samples. Tingey et al. clarified that the average normative sample variances be used in calculating the reliable change index (RCI) and that the normative samples be demonstrated to be distinct, as evidenced by a one-tailed t test ($p < .05$) and Cohen's $d \geq .50$. Tingey et al. then went on to provide an illustration using the Symptom Checklist–90–Revised (SCL–90–R; Derogatis, 1983) and combined the SCL–90–R normative data (Mildly Symptomatic and Severely Symptomatic) with two other distinct samples comprised of Asymptomatic and Moderately Symptomatic individuals. (Examples of CS and RCI calculations using the Tingey et al. [1996a] data may be found in the Appendix for the interested reader.) Noting that the use of confidence bands around the cutoffs resulted in unclear classifications and imposed an additional criterion to pass on an already conservative standard, they recommended only using cutoff scores and RCI to measure reliable CS change.

Martinovich, Saunders, and Howard (1996) generally welcomed Tingey et al's. (1996a) extension but noted several problems. These included Martinovich et al.'s observations that any underlying psychometric problems with the measures could be exacerbated by use of the method, and the method did not solve the problem of identifying and discriminating between functional and dysfunctional groups; they also noted difficulties involved with making distinctions involved with using Criterion 3 with nonnormal distributions (Jacobson, Follette, & Revenstorf, 1984). Follette and Collahan (1996) reiterated their position that the return to normal criterion is the necessary one and criticized the Tingey et al. method of determining if two distributions were unique and the use of intermediate normative data. Tingey, Lambert, Burlingame, and Hansen

(1996b) responded to Martinovich et al. by noting that measurement error was inherent in the instruments, and the method did not increase such error; that the use of intermediate normative categories reflected clinical realities of gradual change; and for a variety of reasons, disagreed that Jacobson and Truax's criterion c was inappropriately used. In their response to issues raised by Follette and Callaghan (1996), Tingey et al. reiterated their support for a continuum of normative data from the functional to dysfunctional and provided additional rationales for their method of using *t* and *d* to distinguish such distributions.

Kazdin (1999) stated that "clinically significant change can occur when there is a large change in symptoms, a medium change in symptoms, and no change in symptoms" (p. 332). Kazdin (1999) indicated that symptom change may not be the gold standard on which to base clinical change and when symptom change is the criterion, a continuum is appropriate. Kazdin (1999) also observed that "One may wish to judge treatments on the extent to which they change symptoms, but the results could be quite different if other criteria were used, such as impairment, quality of life, or impact on others" (p. 336). Kazdin (2001) also stated that "it is not clear that what we refer to as clinically significant on most measures or by most criteria reflects genuine or, indeed, any differences in the everyday lives of the clients" (p. 456). Kazdin (2001) questions the reliance on symptom reduction "as the sole or primary criterion, and the need to match the criteria and measures to … the clinical problems, treatment goals, and lives of the clients" (p. 455). Kazdin (2001) cogently argued that "it is difficult to find evidence to show that passing a threshold or entering a range means the client is better in any way that affects daily functioning or that a failure to pass this threshold means otherwise" (p. 461).

RCI

Recognizing that a CS result could be obtained that was unreliable, Jacobson, Follette, and Revenstorf (1984) introduced a second statistical requirement to ensure that any obtained CS change was reliable. Jacobson, Follette, and Revenstorf termed this the *reliable change index* (RCI). RCI was developed to account for measurement error and was originally defined as

$$\text{RCI} = \frac{\bar{X}_1 - \bar{X}_2}{SE},$$

where X_1 = pretest score; X_2 = posttest score; $SE = s_1\sqrt{1 - r_{xx}}$; s_1 = the standard deviation of control group, normal population, or pretreatment group; and r_{xx} = the test–retest reliability. If RCI ≥ 1.96, then it is likely that the change was reliable ($p < .05$). In general, the less reliable the instrument, the greater the difference required to achieve a statistically reliable change. RCI was designed to account for the magnitude of CS change and measurement error, such as when an individual crosses the threshold from the dysfunctional to functional population, but the amount of change may be minimal or unreliable. According to the Jacobson, Follette, and Revenstorf criteria, for a treatment effect to be considered a reliable and significant change, it must pass a two stage process in which (a) it must be proven to be statistically reliable (RCI) and (b) the individual must pass from the dysfunctional to the functional distribution (CS). Jacobson, Follette, and Revenstorf noted that if these criteria were adopted, "psychotherapy would look less effective" (p. 350) due to the very conservative nature of the criteria. Jacobson, Follette, Revenstorf, Baucom, et al. (1984) later supported the use of the criteria by stating that

> If a client changes to a degree that is both outside the range that would be explained by measurement error and places him or her with greater likelihood in the functional population this client can be considered unequivocally as a treatment success. (p. 498)

Using these criteria, one could then classify each individual in a treatment outcome study as *Recovered* (passed both CS normative and RCI criteria), *Improved* (passed RCI criteria alone) *Unchanged/Indeterminate* (passed neither), or *Deteriorated* (passed RCI in the negative direction). Examples of the Jacobson and Truax (1991) method of calculating CS and RCI is found in Appendix.

CRITICISMS OF THE RCI

Christensen and Mendoza (1986) noted that the RCI formula as originally proposed was based on an individual's obtained pretest score with inherent measurement error and not a "true" pretest score. To correct for this error, Christensen and Mendoza proposed the use of the S_{diff} in place of the denominator standard error (*SE*) to reflect "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. 307) where $S_{diff} = \sqrt{2(SE)^2}$. Christensen and Mendoza also noted that the magnitude of difference required using this denominator is even more stringent than the original RCI formula. Hageman and Arrindell (1993) believed that this formula improved the accuracy of both the pretest and posttest scores, the resulting pretest–posttest difference scores, and hence the RCI. Jacobson et al. (1986) acknowledged that Christenson and Mendoza were correct and adopted their formula.

In further highlighting the methodological complexities of measuring the RCI, Speer (1992) stated that "The more deviant the initial scores and the less reliable the instrument, the greater the regression based improvement that may occur" (p. 403). Although Speer noted that "regression to the mean is not universal [and] occurs only when the correlation be-

tween amount of change and the initial score is negative," he went on to state that determining "whether or not it is a factor that requires attention is an empirical matter and should not be ignored" (p. 403).

Speer (1992) clarified that the larger issue, as cited by Wampold and Jenson (1986), is whether one views the dysfunctional population as separate from but overlapping with the functional population or whether they represent a sample from the extreme end of the population distribution. The more extreme the scores, the greater the chance of the scores regressing to the mean and influencing RCI. Speer (1992) then introduced the Edwards–Nunnally method to minimize "the risk of improvement rates capitalizing on regression to the mean (capitalizing on error of measurement), in those circumstances in which regression to the mean is demonstrated" (p. 404). Rather than using the uncorrected pretest score, Speer used an estimated true score based on the obtained pretest score to minimize regression to the mean. This correction results in more conservative recovery rates than the previously revised RCI formula. Additionally, when regression to the mean is not demonstrated, this formula would produce low estimates of improvement and lower reliability among the more dysfunctional scores, thereby creating disparities in classification rates between this formula and the Jacobson and Traux (1991) method.

Hageman and Arrindell (1993) pointed out that "regression to the mean represents a *problem* only in as far as it is caused by measurement unreliability" (p. 695). Hageman and Arrindell (1993) went on to clarify that "Speer's (1992) finding of a negative correlation between initial score and change is indeed proof of the presence of regression to the mean, however not necessarily of a problematic kind" (p. 695). That is, a negative correlation may also be due to true individual change and not necessarily measurement error. Hageman and Arrindell (1993) summarized Speer's (1992) correction as only addressing error in the pretest scores, noting that posttest scores were equally as subject to unreliability. Subsequently, Hageman and Arrindell (1993) proposed a correction to take into account error for both the pretest and posttest scores and adjusted for regression to the mean. Hageman and Arrindell (1999b) refined their equation to estimate the underlying true scores instead of using observed scores in yet a further attempt to correct for regression to the mean ($RC_{indiv} = (X_2 - X_1)\, r_{DD} + (M_2 - M_1)(1 - r_{DD}) / \sqrt{r_{DD}}$ $\sqrt{SE^2}$, where r_{DD} = reliability of the difference scores).

Speer (1999) later criticized Hageman and Arrindell's (1999b) approach and discussed two relevant "myths" related to RCI. Essentially, Speer changed his opinion and indicated that it was not necessary to adjust for regression to the mean and that in a pretest–posttest design, unadjusted scores were acceptable. Speer went on, however, to state "There is *no* agreement or consensus among methodologists about the ubiquitousness of regression to the mean, it's effects on d-scores [difference-scores], whether or not d-scores are biased and/or unreliable and whether or not d-scores really need adjustment or correction in the analysis of two-wave data" (p. 1209).

COMPARISONS OF RCI METHODS

In light of the statistical controversies surrounding the RCI, the construct validity of the methods and their relative accuracies are of considerable importance. In one of the first empirical evaluations of change scores, Speer and Greenbaum (1995) examined four pretest–posttest methods (Jacobson–Truax, Edwards–Nunnally, Hsu–Linn–Lord, and Nunnally–Kotsch) along with a hierarchical linear model (HLM). These pretest–posttest methods used residualized pretreatment or difference scores, which were believed to be more accurate than the standard formula because of improved reliability. When Speer and Greenbaum omitted the pretest–posttest method with the lowest average rate of classification agreement, they found the average classification rate of agreement was 89% and the HLM was less sensitive in classifying clients as Deteriorated compared to the difference score methods. Among the four pretest–posttest difference score methods studied, Speer and Greenbaum

> Recommend[ed] use of the Jacobson and Truax (1991) pre-post difference method for the following reasons: (a) It avoids statistical problems associated with residualized true score adjustments; regression to the mean is neither inevitable nor, perhaps, as big a problem as previously thought; (b) it is computationally straightforward; and (c) there is a small literature reporting change rates produced by this method. (p. 1047)

McGlinchey and Jacobson (1999) compared the Jacobson and Truax (1991) method with one proposed by Hageman and Arrindell (1999b). McGlinchey and Jacobson concluded that there were no substantial differences between the two methods and that the complex computations used in the Hageman and Arrindell (1999b) method made the Jacobson and Truax formula the preferred method. Hageman and Arrindell (1999a) criticized the choice of norms used by McGlinchey and Jacobson, which Hageman and Arrindell (1999a) argued affected the standard error of measurement and hence the classifications. Hageman and Arrindell (1999a) pointed out the importance of using the correct value for the standard error of measurement to be able to compare studies.

McGlinchey, Atkins, and Jacobson (2002) subsequently compared five methods for determining RCI, including four used in the Speer and Greenbaum (1995) study as well as the Hageman and Arrindell (1999b) formula. In addition to examining differential classification rates, McGlinchey et al. also studied the methods for predictive accuracy with respect to relapse. Whereas McGlinchey et al. found some differences in classification rates, Hageman and Arrindell had the lowest number of Recovered patients and hence was the most conservative estimate. Nonetheless, "All five methods sig-

nificantly discriminated between participants who relapsed during the 2 years following therapy" (McGlinchey et al., 2002, p. 541). McGlinchey et al. concluded that "the evidence … supports the Jacobson Truax method as a 'null' method that has yet to be rejected by an alternative method of superior performance" (p. 542).

Bauer, Lambert, and Nielsen (2003) combined the methods used by both Speer and Greenbaum (1995) and McGlinchey et al. (2002) to compare the classification accuracy of five RCI methods (Jacobson–Truax [Jacobson & Traux, 1991]; Edwards–Nunnally [Speer, 1992]; Gulliksen–Lord–Novick [Hsu, 1989]; Hageman–Arrindell [Hagemann & Arrindell, 1999b]; and the HLM approach). Bauer et al. found that the Edwards–Nunnally and Hageman–Arrindell approaches demonstrated poor convergence with the other three methods. The Edwards–Nunnally method tended to produce the most liberal Recovery rates, whereas the Hageman–Arrindell method produced the most conservative. The HLM method was noted to require more than two data points (pretest–posttest) and produced relatively low convergence with the other methods. Of the remaining two methods, Bauer et al. found little difference between the Jacobson–Truax method and the Gulliksen–Lord–Novick method. Due to the widespread use of the Jacobson–Truax method and the relative ease of calculation, Bauer et al. (2003) recommended the Jacobson–Truax method. Thus, Speer and Greenbaum (1995), Speer (1999), Maassen (2001), McGlinchey et al. (2002), and Bauer et al. demonstrated the relative convergence of the Jacobson–Truax method, and these authors called for a moratorium on creating alternative RCI formulas to study the results derived from the Jacobson and Truax (1991) formula. Consensus appears to favor further study and development of the Jacobson and Truax approach.

APPLICATIONS AND INNOVATIONS

In a selected review of the RCI and CS literature, Ogles, Lunnen, and Bonesteel (2001) identified 74 published RCI-related articles in a 9-year period in the *Journal of Consulting and Clinical Psychology,* ranging from 3 to 14 publications per year. Ogles et al. found that across all outcome studies included, an RCI analysis was the most frequent method employed. Many prestigious journals have published special sections related to the use of and debate about RCIs (e.g., *Journal of Consulting and Clinical Psychology, Clinical Psychology: Research and Practice, Psychotherapy Research, Behaviour Research and Therapy, Behavioral Assessment,* etc.). The method has been used with a variety of clients including adults, children, people with personality disorders as well as medical and neuropsychological patients (e.g., Dolan, Evans, & Wilson, 1992; Ferguson, Robinson, & Splaine, 2002; Jacobson, Roberts, Berns, & McGlinchey, 1999; Lunnen & Ogles, 1998; Sheldrick, Kendall, & Heimberg, 2001; Temkin, Heaton, Grant, Dikmen, & Sureyya, 1999).

Jacobson and Revenstorf (1988) offered some methodological refinements by noting that when multiple measures of similar constructs were used to assess treatment outcomes, summary statistics of mean recovery rates across the measures could be computed to calculate CS and RCI. Alternatively, Jacobson and Revenstorf suggested a multivariate weighted composite score based on all measures that could be analyzed by separating the functional and dysfunctional populations, which would allow for the use of cutoff points. Discriminant functions were also mentioned when using multiple measures to classify individuals into functional or dysfunctional categories. When multiple measures are used to assess different constructs, however, Jacobson and Revenstorf advocated using the multiple measures and "accept[ing] the fact that no single index of clinically significant change will capture all components of the disorder under study" (p. 140).

Other innovative applications include Nietzel, Russell, Hemmings, and Gretter's (1987) meta-analytic study of patients with unipolar depression in which composite group mean scores were compared to normative groups to determine CS change. Nietzel et al. also recommended a normative change of 1 *SD* instead of 2 *SD*s as the cutoff point for CS. Abramowitz (1998) used the RCI formula in a meta-analytic review of exposure therapy in the treatment of obsessive–compulsive disorders and calculated a change score for each group instead of on each patient. Similarly, Sheldrick et al. (2001) employed the RCI formula with normative comparison methods in the treatment of children with conduct disorders. Sheldrick et al. derived change scores for each group and then utilized group scores to determine if treated individuals were returned to normal limits (1 *SD*) to determine CS. Seggar, Lambert, and Hansen (2002) developed a normative continuum using the Beck Depression Inventory (Beck, 1987) similar to that created by Tingey et al. (1996). Adding to the innovative work of Tingey (1989) and Tingey et al. (1996a), Wise (in press) extended the SCL–90–R normative continuum by adding a sample of 225 outpatients attending an Intensive Outpatient Program, thereby adding a more severe population of outpatients.

To address the multidimensional nature of clinical significance, the use of multiple measures or criteria have been useful. Ogles, Lambert, and Masters (1996) provided RCI graphs using various scales and instruments that facilitate the use of RCI by clinicians. Ogles, Lambert, and Sawyer (1995) demonstrated RCI convergence between a structured rating scale completed by clinicians and two client rated symptom measures on a national sample of outpatients with depression under four different treatment conditions. Lunnen and Ogles (1998) studied the RCI using the perspectives of therapists, clients, and spouses to rate symptom change, alliance, and satisfaction. Beckstead et al. (2003) demonstrated the use of the RCI on five different instruments on the same population and

found a 65% classification agreement rate between them. Beckstead et al. concluded that different measures of different domains will produce different RCI classification results. Wise (in press) examined the effects of varying RCI criteria (1.96, 1.28, and .84, corresponding to 95%, 90%, and 80% confidence levels, respectively) with pretest–posttest client symptom ratings, pretest–posttest clinician-rated level of functioning, client satisfaction, and discharge to a lower level of care and demonstrated that although different measures of different domains produce different results, convergence of outcome classification rates can also be demonstrated.

FUTURE DIRECTIONS AND RECOMMENDATIONS

Jacobson and colleagues (Jacobson, Follette, & Revenstorf, 1984; Jacobson & Revenstorf, 1988; Jacobson et al., 1999; Jacobson & Truax, 1991) have repeatedly contended that consumers and clinicians "expect to be as normal as their functional counterparts by the time therapy has ended" (Jacobson & Revenstorf, 1988, p. 134) and that this return to normal criterion should be the definition of CS. The use of a return to normal criterion seems unrealistic for many clinical practice contexts. This is most evident in assessing the treatment effects of inpatient programs, partial hospitalization programs, and intensive outpatient programs. The comparison may also not be appropriate for certain types of cases such as dual diagnosis disorders, personality disorders, psychiatric patients with comorbid medical problems, and so forth. For these patients, the natural course, chronicity, and recurrent waxing and waning of intensity of symptoms argues against a return to normal criterion. In fact, literature reviews and meta-analytic studies of outpatients with depression report that few achieve complete remission and despite significant improvements, some data indicate that many of these patients remain more depressed at the end of therapy than the general population (e.g., Hansen, Lambert, & Forman, 2002; Lecrubier, 2002; Nietzel et al., 1987; Robinson, Berman, & Neimeyer, 1990; Westin & Morrison, 2001). Additionally, a functional return to normal may be expected to take considerably longer than symptom remission, and it is not clear if particular functional impairments remit faster or slower than others (e.g., return to work vs. increased socialization). Because comparatively few patients actually achieve a full remission of depressive symptoms, and functional capacities are the last improvements, the return to normal criterion appears unrealistic.

The work of Norman, Sloan, and Wyrwich (2003), who conducted a review of the literature on minimally important differences (MID) for health-related quality-of-life instruments, is particularly relevant to the return to normal criteria. Based on 38 studies and 62 effect sizes, Norman et al. demonstrated that irrespective of the disease or instrument, .5 *SD* consistently detected reliable change in chronic medical patients. This consistency "corresponds almost exactly to the limit of human discrimination identified by Miller (1956) over 40 years ago" (Norman et al., 2003, p. 588). Norman et al. also found that otherwise "healthy people recovering from an episode of back or shoulder pain, referred to a therapist, with every expectation of complete recovery" (p. 589) demonstrated a higher threshold for minimal change than those with chronic conditions. This indicates that the patients' expectations for full or partial recovery influence their discrimination of minimal change. This suggests that psychotherapy patients' expectations regarding a return to normal outcome or a decrease but not elimination of symptoms would affect their judgments of change. Hence, Jacobson and Truax's (1991) return to normal criterion may accurately reflect the perspective of the patient who has transient situational disorders but may not be accurate for those with more chronic psychological disorders such as Dysthymia, refractory or recurrent Major Depression, patients with underlying Axis II disorders, and so forth. Whereas some researchers noted previously have recommended or utilized 1 *SD*[1] as a CS normative change criterion, the MID work of Norman et al. provided support for lessening this criterion to .5 *SD* based on the patient population and suggests that corresponding RCI adjustments may also be appropriate.

Note that the cost of lessening the standards for CS and RCI could be to increase the number of patients classified as Improved or Deteriorated. However, this would assume that the patients are normally distributed around a mean change of zero and that any change to the criterion would have a symmetrical effect. If, on the other hand, therapy is at all effective (e.g., Nietzel et al., 1987; Robinson et al., 1990; Shadish et al., 1997; Smith, Glass, & Miller, 1980; Wampold, 2001), then the change would not necessarily be symmetrical and could result in more people classified as Improved rather than Deteriorated. In any event, such increases are likely to be relatively small.

Additionally, varying the RCI allows the CS variables to have a greater influence in the determination of therapy outcomes in general and in individual improvement rates in particular. In addition to varying the number of standard deviation units required to demonstrate normative CS change, a symptom scale might also be used by applying varying confidence levels to the RCI formula (e.g., 1.96, 1.28, and .84, corresponding to 95%, 90%, and 80% confidence levels, respectively). Such varying RCI confidence levels could be used in conjunction with real-world measures such as discharge to a lower level of care, clinician ratings, functional ratings, client ratings of coping, ratings by signifi-

[1]In reviewing the meta-analytic studies of psychotherapy outcomes, Wampold (2001) noted that the average improvement was reflected by an effect size (ES) of .80. Because a change of 1 *SD* corresponds to an ES of 1.0, where .80 is considered to be a large ES, it would appear that a change of 1 *SD* is also a defensible indicator of CS.

cant others, and client satisfaction (e.g., see Wise, in press). In this example, varying the RCI cutoff from 1.96 to .84 might be used with CS variables such as discharge to a lower level of care, clinician Global Assessment of Functioning Scale (GAF; American Psychiatric Association, 2000) ratings that were 1 *SD* closer to the functional distribution, client reports that treatment helped them cope more effectively with their problems, and so forth. Alternatively, an RCI cutoff of 1.96 might be used with less rigorous or unstandardized CS variables, whereas .84 might be used with standardized measures such as symptom rating scales. The use of multiple measures with varying levels of stringency allows for the comparative and convergent validity of treatment effectiveness within studies. Altering the RCI or CS criteria in some settings under some conditions using various measures might be more informative indicators of outcome.

For example, when Wise (in press) reduced the RCI criterion from 1.96 to .84, the number of patients who Deteriorated remained unchanged, whereas the number who Improved increased by 4%. However, when the same changes in RCI were used with "Discharge to a Lower Level of Care" as the CS variable, the number of Improved increased by 12%, the number who Deteriorated increased from 3% to 11%, and all of those reclassified came from the Indeterminant/Unclassified[2] group. Inspection of the charts of the Deteriorated group revealed that 92% of these patients had experienced significant psychosocial stressors in the 2 weeks before termination. Not only were the majority of the patients who were reclassified from the Indeterminant/Unclassified group, but the finding that 92% of those who Deteriorated experienced psychosocial stressors just prior to discharge proved to be an important one with immediate, clinically relevant implications that otherwise would have been overlooked.

Irrespective of what normative variables or cutoffs are utilized, RCI imposes an upper limit on the number of cases classified as Recovered, Improved, Indeterminant/Unclassified or Deteriorated. Thus, if only 50% of a sample pass the RCI criteria for symptom improvement, individuals who demonstrate significant normative change on the CS variable of interest but fail to pass the RCI criteria, will not be included in these recovery rates. Additionally, although there is some differentiation between the Recovered and Improved categories in terms of CS, there is no such distinction within the Deteriorated classification. That is, the criterion of RCI change alone determines the classification of Deteriorated. Perhaps an additional category reflecting a reliably worse condition would be more accurate for those demonstrating only a negative RCI, whereas Deteriorated could be used to denote those who passed both the RCI and CS criteria in the negative direction. In any event, this degree of reliance on RCI seems to defeat the role and purpose of CS variables contributing to the classification of psychotherapy outcomes.

There can be little doubt that traditional RCI recovery rates are extremely conservative psychotherapy outcome measures and that those who cross into the functional range are "unequivocally … treatment success[es]" (Jacobson, Follette, Revenstorf, Baucom, et al., 1984, p. 498). Developing methods to further refine RCI and CS measures in the large proportion of patients in the Indeterminant/Unclassified range would be of considerable help in more accurately identifying and studying those who are not unequivocal treatment successes but who are nonetheless improving and on their way to a positive outcome as well as those who are not responding to treatment.

Examining treatment response classifications from the psychopharmacology literature (Lecrubier, 2002) indicates that *Remission* is defined as a symptom reduction of 75% to 100% lasting 2 weeks to 6 months, whereas *Recovery* is defined as the same amount of symptom reduction lasting > 6 months. A positive or negative *Response* is defined as ≥ 50% symptom reduction or increase, and a *Partial Response* is defined as a 25% to 49% reduction or increase in symptoms. Table 1 shows one way of combining these definitions with RCI, CS, and Norman et al.'s (2003) MID analyses in an effort to further identify and classify those who would traditionally be classified in the Indeterminant/Unclassified range. The gradations in Table 1 reflect the continuum of Recovery–Deterioration as well as the degree of confidence that can be placed in the classifications.

Additionally, note that in their meta-analytic review, Robinson et al. (1990) found posttreatment findings to be highly significantly predictive of follow-up findings. Another way to classify the Indeterminant/Unclassified group might be to analyze clinical variables that achieve either RCI or CS and

TABLE 1
Proposed Terms to Classify Reliable Change Index and Clinical Significance

Reliable Change Indexes	*Clinical Significance and Normative Change Criterion*			
	Confidence Levels (%)	*Change of 2 SDs*	*Change of 1 SD*	*Change of 0.5 SD*
1.96	95	Recovered	(+) Response	Minimal (+) Response
1.28	90	Remitted	(+) Response	Minimal (+) Response
0.84	80	Improved	(+) Partial Response	Minimal (+) Response
–0.84	80	Mildly Deteriorated	(–) Partial Response	Minimal (–) Response
–1.28	90	Moderately Deteriorated	(–) Response	Minimal (–) Response
–1.96	95	Deteriorated	(–) Response	Minimal (–) Response

[2]Statistically, this group shows No Reliable Change and historically has been called *Unchanged, No Change,* and *Unclassified.* However, these people may have Improved or Deteriorated below the threshold of statistical detection; therefore, I refer to them as *Indeterminant/Unclassified.*

to classify these individuals as Improved (e.g., see Hansen, Lambert & Forman, 2002, and Wise, in press).

Because control groups and random assignment are not typically appropriate in naturalistic studies, unique methodologies are often employed to accommodate clinically representative studies. Analyses of treatment outcomes that demonstrate RCI + CS results should be accorded greater weight in recognizing their demonstrated effectiveness. Similarly, individuals passing RCI or normative CS criteria with additional supporting real-world operationalized clinical measures or indicators (e.g., transfer to a lower level of care, functional ratings, clinician ratings, spouse ratings, etc.) are clearly passing a higher bar than traditional pretest–posttest group designs focused only on statistical significance and subsequently should be accorded such relative status in acknowledging their empirically validated treatment effectiveness.

Some clinically relevant measures may be indirectly related to functional status (e.g., symptom severity, GAF scale), whereas others reflect behavioral change (e.g., discharge to a lower level of care, return to work, absenteeism, number of social contacts, etc.). Although it has been demonstrated that improvement in functional status takes longer than symptom improvement and is directly related to the dose and phase of treatment (Howard, Leuger, Maling, & Martinovich, 1993), there are few measures that evaluate both symptoms and the multiple domains of behavioral functional status. Although some measures assess familial, social, work, leisure, and health status (e.g., Sheehan, 1986; Ware, 1993; Weissman, Klerman, Paykel, Prusoff, & Hanson, 1974), they do not measure all of these domains, are not in the public domain, are incomplete as stand alone outcome measures, or add too much time to be included in a pretest–posttest design in addition to a stand alone symptom measure in real-world settings. Recognizing the need for an instrument that was practitioner friendly and assessed multiple domains, Lambert et al. (1996) provided the OQ45, a brief 45-item instrument that assesses Symptom Distress, Interpersonal Relations, and Social Role Performance (employment, family, and leisure). Of particular relevance is the fact that Beckstead et al. (2003) demonstrated evidence supporting the construct validity of the OQ–45.2 cutoff scores for CS. Similarly, Grissom, Lyons, and Lutz (2002) reported on a new instrument, Treatment Evaluation and Management, that is based on the dose and phase model of psychotherapy. It is comprised of 92 items that assess Subjective Well-Being (emotional and physical health), Symptoms (depression, anxiety, somatization, substance abuse, etc.), Functional Disability (social, vocational and activities of daily living), Therapeutic Bond and Satisfaction With Care (working alliance, understanding and trust), along with inconsistency and malingering indicators. Although it is too early to know the parameters of the normative database, the initial psychometrics are promising, and this instrument is noteworthy for its emphasis on assessing multiple domains including alliance and case mix adjustment variables with relatively few items.

Clearly there is a need for real-world measures designed to assess both symptoms and functional capacities in intervals or increments that can be expected to respond to psychotherapy dosages, that are quantifiable, normed, and assess multiple functional domains. As mental health becomes integrated with medical care, these needs for multiple domain assessments will become even more evident. The RCI and CS methodology has withstood rigorous debate and survived stronger than originally conceived. Despite methodological limitations, studying RCI and CS has moved the outcomes paradigm from studying treatment groups to studying individual change within those groups. Similarly, assessment instruments must move beyond symptom focus and evaluate individuals with respect to the complex broader domains of their functional, real-world, lives in which clinically significant change is operationalized.

REFERENCES

Abramowitz, J. S. (1998). Does cognitive behavior therapy cure obsessive–compulsive disorder? A meta-analytic evaluation of clinical significance. *Behavior Therapy, 29,* 339–355.

American Psychiatric Association. (2000). *Diagnostic and statistical manual* (4th ed.). Washington, DC: Author.

Bauer, S., Lambert, M. J., & Nielsen, S. L. (2003). Clinical significance methods: A comparison of statistical techniques. Manuscript submitted for publication.

Beck, A.T., & Steer, R. A. (1987). *Beck Drepression Inventory manual*. San Antonio, TX: The Psychological Corporation, Harcourt Brace Jovanovich.

Beckstead, D. J., Hatch, A. L., Lambert, M. J., Eggett, D. L., Goates, M. K., & Vermeersch, D. A. (2003). Clinical significance of the Outcome Questionnaire (OQ–45.2). *The Behavior Analyst Today, 4,* 79–90.

Christensen, L., & Mendoza, J. L. (1986). A method of assessing change in a single subject: An alteration of the RC index. *Behavior Therapy, 15,* 305–308.

Derogatis, L. R. (1983). *SCL–90–R: Administration, scoring and procedures manual* (2nd ed.). Towson, MD: Clinical Psychometric Research.

Dolan, B. M., Evans, C., & Wilson, J. (1992). Therapeutic community treatment for personality disordered adults: Changes in neurotic symptomatology on follow-up. *International Journal of Social Psychiatry, 38,* 243–250.

Ferguson, R. J., Robinson, A. B., & Splaine, M. (2002). Use of the Reliable Change Index to evaluate clinical significance in SF–36 outcomes. *Quality of Life Research, 11,* 509–516.

Follette, W. C., & Callaghan, G. M. (1996). The importance of the principle of clinical significance—Defining significant to whom and for what purpose: A response to Tingey, Lambert, Burlingame, and Hansen. *Psychotherapy Research, 6,* 133–143.

Grissom, G. R., Lyons, J. S., & Lutz, W. (2002). Standing on the shoulders of a giant: Development of an outcome management system based on the dose model and phase model of psychotherapy. *Psychotherapy Research, 12,* 397–412.

Hageman, W. L., & Arrindell, W. A. (1993). A further refinement of the reliable change index by improving the pre–post difference score: Introducing the RC_{ID}. *Behaviour Research and Therapy, 51,* 693–700.

Hageman, W. L., & Arrindell, W. A. (1999a). Clinically significant and practical?: Enhancing precision does make a difference: Reply to McGlinchey and Jacobson, Hsu, and Speer. *Behaviour Research and Therapy, 37,* 1219–1233.

Hageman, W. L., & Arrindell, W. A. (1999b). Establishing clinically significant change: Increment of precision and the distinction between individ-

ual and group level analysis. *Behaviour Research and Therapy, 37,* 1169–1193.

Hansen, N. B., Lambert, M. J., & Forman, E. M. (2002). The psychotherapy dose-response effect and its implications for treatment delivery services. *Clinical Psychology: Science and Practice, 9,* 329–343.

Howard, K. I., Leuger, R., Maling, M., & Martinovich, Z. (1993). A phase model of psychotherapy: Causal medication of outcome. *Journal of Consulting and Clinical Psychology, 61,* 678–685.

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

Hsu, L. M. (1996). On the identification of clinically significant client changes: Reinterpretation of Jacobson's cut scores. *Journal of Psychopathology and Behavioral Assessment, 18,* 371–385.

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

Jacobson, N. S., Follette, W. C., & Revenstorf, D. (1986). Toward a standard definition of clinically significant change. *Behavior Therapy, 15,* 309–311.

Jacobson, N. S., Follette, W. C., Revenstorf, D., Baucom, D. H., Hahlweg, K., & Margolin, G. (1984). Variability in outcome and clinical significance of behavioral marital therapy: A re-analysis of outcome. *Journal of Consulting and Clinical Psychology, 52,* 497–504.

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

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

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

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

Kazdin, A. E. (2001). Almost clinically significant (p. < .10): Current measures may only approach clinical significance. *Clinical Psychology: Science and Practice, 8,* 455–462.

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

Lambert, M. J., Hansen, N. B., Umphress, V., Lunnen, K., Okiishi, J., Burlingame, G., et al. (1996). *Administration and scoring manual for the Outcome Questionnaire (OQ–45.2).* Wilmington, DE: American Professional Credentialing Services.

Lecrubier, Y. (2002). How do you define remission? *Acta Psychiatrica Scandinavica, 106*(Suppl. 415), 7–11.

Lunnen, K. M., & Ogles, B. M. (1998). A multiperspective, multivariable evaluation of reliable change. *Journal of Consulting and Clinical Psychology, 66,* 400–410.

Maassen, G. H. (2001). The unreliable change of reliable change indices. *Behaviour Research and Therapy, 39,* 495–498.

Martinovich, A., Saunders, S., & Howard, K. I. (1996). Some comments on "Assessing clinical significance." *Psychotherapy Research, 6,* 124–132.

McGlinchey, J. B., Atkins, D. C., & Jacobson, N. S. (2002). Clinical significant methods: Which one to use and how useful are they? *Behavior Therapy, 33,* 529–550.

McGlinchey, J. B., & Jacobson, N. S. (1999). Clinically significant but impractical?: A response to Hageman and Arrindell. *Behaviour Research and Therapy, 37,* 1211–1217.

Miller, G. A. (1956). The number seven plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63,* 81–97.

Nietzel, M. T., Russell, R. L., Hemmings, K. A., & Gretter, M. L. (1987). Clinical significance of psychotherapy for unipolar depression: A meta-analytic approach to social comparison. *Journal of Consulting & Clinical Psychology, 55,* 156–161.

Norman, G. R., Sloan, J. A., & Wyrwich, K. W. (2003). Interpretation of changes in health-related quality of life: The remarkable universality of half a standard deviation. *Medical Care, 41,* 582–592.

Nunnally, J. C., & Kotsch, W. E. (1983). Studies of individual subjects: Logic and methods of analysis. *British Journal of Clinical Psychology, 22,* 83–93.

Ogles, B. M., Lambert, M. J., & Masters, K. S. (1996). *Assessing outcome in clinical practice.* Boston: Allyn & Bacon.

Ogles, B. M., Lambert, M. J., & Sawyer, J. D. (1995). Clinical significance of the National Institute of Mental Health Treatment of Depression Collaborative Research Program data. *Journal of Consulting and Clinical Psychology, 63,* 321–326.

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

Robinson, L. A., Berman, J. S., & Neimeyer, R. A. (1990). Psychotherapy for the treatment of depression: A comprehensive review of controlled outcome research. *Psychological Bulletin, 108,* 30–49.

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

Shadish, W., Matt, G., Navarro, A., Siegle, G., Crits-Christoph, P., Hazelrigg, M., et al. (1997). Evidence that therapy works in clinically representative conditions. *Journal of Consulting and Clinical Psychology, 65,* 355–365.

Sheehan, D. V. (1986). *The anxiety disease.* New York: Bantam.

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

Smith, M., Glass, G., & Miller, T. (1980). *The benefits of psychotherapy.* Baltimore: Johns Hopkins University Press.

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

Speer, D. C. (1999). What is the role of two-wave designs in clinical research? Comment on Hageman and Arrindell. *Behaviour Research and Therapy, 37,* 1203–1210.

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.

Temkin, N. R., Heaton, R. K., Grant, I., Dikmen, S., & Sureyya S. (1999). Detecting significant change in neuropsychological test performance: A comparison of four models. *Journal of the International Neuropsychological Society, 5,* 357–369.

Tingey, R. C. (1989). Assessing clinical significance: Extension in methods an application to the SCL–90–R. *Dissertation Abstracts International, 50*(04B), 1659.

Tingey, R. C., Lambert, M. L., Burlingame, G. M., & Hansen, N. B. (1996a). Assessing clinical significance: Proposed extensions to the method. *Psychotherapy Research, 6,* 109– 103.

Tingey, R. C., Lambert, M. L., Burlingame, G. M., & Hansen, N. B. (1996b). Clinically significant change: Practical indicators for evaluating psychotherapy outcome. *Psychotherapy Research, 6,* 144–153.

Wampold, B. E. (2001). *The great psychotherapy debate: Models, methods and findings.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Wampold, B. E., & Jenson, W. R. (1986). Clinical significance revisited. *Behavior Therapy, 17,* 302–305.

Ware, J. E. (1993). *SF–36 Health Survey: Manual and interpretation guide.* Boston: Nimrod.

Weissman, M. M., Klerman, G. L., Paykel, E. S., Prusoff, B., & Hanson, B. (1974). Treatment effects in the social adjustment of depressed patients. *Archives of General Psychiatry, 30,* 771–778.

Westin, D., & Morrison, K. (2001). A multidimensional meta-analysis of treatments for depression, panic, and generalized anxiety disorder: An

empirical examination of the status of empirically supported therapies. *Journal of Consulting and Clinical Psychology, 69,* 875–899.

Wise, E. A. (in press). Psychotherapy outcome and satisfaction methods applied to intensive outpatient programming in a private practice setting. *Psychotherapy: Theory, Research and Practice.*

APPENDIX

To demonstrate the application of CS and RCI, Derogatis's (1983) SCL–90–R Global Severity Index (GSI) scale will be used as an example. Recall that Jacobson and Truax (1991) indicated that when two distributions are nonoverlapping, the cutoff point of choice is 2 *SD*s above the functional group. In this case, Derogatis's (1983) "non-patient normal" group obtained a *M* GSI = .31 with a SD =.31, whereas the Outpatient group had a *M* = 1.26 with a *SD* = .68. Hence, the cutoff point would equal 2 *SD*s above the nonpatient group, or (2 × .31) + .31 (*M*) = .93. An individual whose pretreatment GSI score = 1.4 and posttreatment GSI = .80 would be considered to have obtained a CS change because they crossed the .93 threshold.

To determine whether or not such a CS change is reliable requires calculation of the Jacobson & Truax (1991) RCI formula:

$$RCI = \frac{\bar{X}_1 - \bar{X}_2}{SE},$$

where X_1 = pretest score; X_2 = posttest score; $SE = s_1\sqrt{1 - r_{xx}}$; s_1 = the standard deviation of control group, normal population, or pretreatment group; and r_{xx} = the test–retest reliability. Again, using the SCL–90–R "Non-patient normal" norms where the GSI *SD* = .31 and the GSI test–retest correlation coefficient was reported by Tingey et al. (1996a) to be .939, $\text{RCI} = \frac{1.40 - .80}{.31\sqrt{1 - .939}} = 7.5$. Because 7.5 is ≥ 1.96, 7.5 results represent a reliable change.

Alternatively, using normative cutoff points between adjacent samples, the following formula is used:

$$\text{Cutoff} = \frac{(s_1 \times \bar{X}_2) + (s_2 \times \bar{X}_1)}{s_1 + s_2}.$$

where X_1, s_1, X_2 and s_2 specify the means and standard deviations of two different samples. The sample means (and standard deviations) for the SCL–90–R GSI normative continuum gathered by Tingey et al. (1996a) and calculated by this formula are Asymptomatic = 0.19 (0.16), Mildly Symptomatic = 0.31 (0.31), Moderate Symptomatic = 0.79 (0.45), and Severely Symptomatic = 1.30 (0.82). The cutoff scores between the Asymptomatic and Mildly Symptomatic would be

$$\frac{(0.16 \times 0.31) + (0.31 \times 0.19)}{0.16 + 0.31} = 0.23.$$

Similarly, computing the cutoff points for the Mildly Symptomatic:Moderately Symptomatic = 0.51 and Moderately Symptomatic:Severerly Symptomatic = 0.97. Tingey et al. (1996a) also required that the adjacent groups be differentiated by $t < .05$ and $d > .50$, which was demonstrated between these groups. Thus, an individual whose pretreatment GSI score was 1.4 and posttreatment GSI = 0.80 would be considered to cross the normative CS cutoff by passing the 0.97 cutoff. In this case, a normative shift from one group on the normative continuum to a less pathological normative group provides an alternative to the traditional CS method. To determine if this CS change is reliable, RCI must be calculated.

Using these figures with the RCI formula cited previously,

Then, because GSI $X_1 = 1.4$ and $X_2 = 0.80$,

$$\sqrt{1 - .939} = 0.20. \qquad \text{RCI} = \frac{1.4 - 0.80}{0.20} = \frac{0.60}{0.20} = 3.0. \qquad SE = 0.82$$

Because 3.0 ≥ 1.96, the CS normative change in score from 1.4 to 0.80 is considered to be reliable.

Edward A. Wise
Mental Health Resources
1027 South Yates Road
Memphis, TN 38119
E-mail: eawmhr@aol.com

Received June 15, 2003
Revised July 17, 2003