JOURNAL OF PERSONALITY ASSESSMENT, 82(1), 60–70

Clinical Significance Methods: A Comparison of Statistical Techniques

Stephanie Bauer
Center for Psychotherapy Research
Stuttgart, Germany

Michael J. Lambert
Department of Psychology
Brigham Young University

Steven Lars Nielsen
Counseling and Career Center
Brigham Young University

Clinically significant change refers to meaningful change in individual patient functioning during psychotherapy. Following the operational definition of clinically significant change offered by Jacobson, Follette, and Revenstorf (1984), several alternatives have been proposed because they were thought to be either more accurate or more sensitive to detecting meaningful change. In this study, we compared five methods using a sample of 386 outpatients who underwent treatment in routine clinical practice. Differences were found between methods, suggesting that the statistical method used to calculate clinical significance has an effect on estimates of meaningful change. The Jacobson method (Jacobson & Truax, 1991) provided a moderate estimate of treatment effects and was recommended for use in outcome studies and research on clinically significant change, but future research is needed to validate this statistical method.

Clinically significant change refers to changes in patient functioning that are meaningful for individuals who undergo psychosocial or medical interventions. This concept has considerable value in research aimed at classifying each individual patient's status with regard to normative functioning. In this regard, it allows researchers to focus on the functioning of each patient rather than on group averages and statistical significance of between-group comparisons. Research using operationalizations of clinical significance has been especially useful in estimating dose-response relationships (e.g., Anderson & Lambert, 2001) and in outcome management systems that use it as a marker for recovery and deterioration (Lambert et al., 2001). In addition, it has been used to estimate the relative value of empirically supported therapies as examined in clinical trials versus routine practice (Hansen, Lambert, & Forman, 2002).

In all these uses, the degree of change in the individual is of primary interest. Such a focus is not only thought to be of scientific importance but it leads to narrowing the gap between clinical research and clinical practice. Thus, the concept and its operationalization have generated considerable interest. Its introduction by Jacobson, Follette, and Revenstorf (1984) was regarded as an important advance in methodology (Lambert, Shapiro, & Bergin, 1986), and it has become an expected statistic in published outcome studies by some journal editors. The topic of clinically significant change has generated considerable attention in special journal sections devoted to the topic (e.g., Jacobson, 1988; Kendall, 1999; Kendall, Marrs-Garcia, Nath, & Sheldrick, 1999; Tingey, Lambert, Burlingame, & Hansen, 1996b).

The original proposal (Jacobson et al., 1984) with minor modifications (Jacobson & Truax, 1991) suggested a two-step criterion for clinically significant change. First, a cutoff point for a measure of psychological functioning is established that is conceptualized as a cutoff between two populations: a patient/dysfunctional population, and a nonpatient/functional population. To this end, Jacobson and Truax identified three reasonable cutoffs for consideration. The first, *Cutoff A*, was defined as the point 2 *SD*s beyond the range of the pretherapy mean (Cutoff A = $M_{clinical} - 2\ SD_{clinical}$). Cutoff A has high sensitivity, that is, an outcome score below this score is very unlikely to belong to the pa-

tient population, although it is not possible to draw conclusions about "recovery" without information on a functional comparison group. The second, *Cutoff B,* was defined as the point 2 *SD*s within a recognized functional mean (Cutoff B = $M_{nonclinical} + 2\ SD_{nonclinical}$) and should be calculated if only nonpatient data are available. Cutoff B has high specificity. It is not difficult for most clients to attain because most dysfunctional and functional distributions overlap. The third, *Cutoff C,* was a weighted midpoint between the means of a functional and dysfunctional population (Cutoff C = $[(SD_{clinical} \times M_{nonclinical}) + (SD_{nonclinical} \times M_{clinical})] / (SD_{clinical} + SD_{nonlinical})$). When both patient/dysfunctional and nonpatient/functional data sets are available, and there is overlap between the two distributions, C represents the best choice for a cutoff point. Compared to A and B, it is the least arbitrary score, as it is based on the relative probability of a particular score ending up in one population as opposed to another (Jacobson, Roberts, Berns, & McGlinchey, 1999). In contrast, A and B would be chosen if not enough information for the calculation of C is available. For example, if adequate norms are lacking, A must be used because neither B nor C can be calculated.

The second step of the Jacobson–Truax (JT; 1991) method is to determine whether a client's change from pretest to posttest is reliable rather than simply an artifact of measurement error. To assess this, Jacobson et al. (1984) proposed a reliable change index (RCI) that each participant has to pass to demonstrate that his or her change is not simply due to chance. RCIs are derived from the psychometric qualities of the outcome measure used to estimate change: The formula divides the difference between pretreatment and posttreatment scores by a variation of the standard error of measurement (S_E). In the discussion of how to calculate reliable change most accurately, a frequent topic concerns the question of which reliability estimate should be used to calculate the S_E. Most studies use either test–retest coefficients or internal consistency estimates. Martinovich, Saunders, and Howard (1996) gave a detailed description of the advantages and disadvantages of both alternatives. In the end, they recommended, especially for clinical populations, the use of a measure of internal consistency rather than test–retest reliability. The problem with test–retest reliabilities is that in clinical samples they are deflated by real individual differences in change, and there is no doubt that these changes occur even during very short periods of time and without the patients being in therapy during that period (Howard, Lueger, Maling, & Martinovich, 1993). Therefore, researchers often use test–retest reliability scores from nonpatient samples. Agreeing with the explanations of Martinovich et al. (1996), Tingey et al. (1996b) came to the conclusion that the use of internal consistency would be the better way to calculate rates of reliable change. This makes especially good sense because outcome scales typically attempt to measure characteristics that should change over time rather than personality traits that are, by definition, stable over time.

Based on the two criteria (cutoff and RCI), the JT method classifies individuals as *Recovered* (i.e., passed both cutoff and RCI criteria), *Improved* (i.e., passed RCI criterion but not the cutoff), *Unchanged* (i.e., passed neither criteria), or *Deteriorated* (i.e., passed RCI criterion but worsened).

Jacobson and colleagues (Jacobson et al., 1999) are aware that several difficulties are related to the approach of clinical significance. For instance, the cutoff point always depends on the specific samples used in a particular study as long as there are no carefully collected norms for both dysfunctional and normal populations. Without normative information, one cannot evaluate if the sample included in an actual study is representative or not. Furthermore, without any norms for normal populations, Cutoff C cannot be calculated. Instead, one has to use Cutoff A, which results in varying estimates of clinical significance because means and standard deviations vary from study to study. In contrast, if distribution information on both normative samples is available, and all participants score in the dysfunctional range at the beginning of treatment, and the distributions do not overlap, it would be senseless to use a cutoff score: In that case, a change from dysfunctional to functional range would always be reliable and never due to measurement error (Jacobson et al., 1999).

Another problem was addressed by Tingey, Lambert, Burlingame, and Hansen (1996a). Tingey et al.'s main point of criticism is that Jacobson et al. (1984) did not provide an operationalization of a comparative social standard. As a consequence, it is not possible to identify and use relevant normative samples across studies. Additionally, the social validation methodology is restricted by the use of only one dysfunctional and one functional sample. Finally, no procedure to determine the distinctness of samples is available. Due to these limitations, Tingey et al. proposed interesting extensions for assessing clinical significance, focusing on the derivation of relevant social standards. Even if these suggestions have been further extended and their value has been acknowledged (Martinovich et al., 1996), there have been hardly any studies really using them.

The "traditional" JT method for assessing clinically meaningful change is among the most frequently reported by researchers. In a review of outcome studies that reported clinically significant change published in the *Journal of Consulting and Clinical Psychology,* Ogles, Lunnen, and Bonesteel (2001) noted that the clinical significance method originally proposed by Jacobson et al. (1984) was used in 35% of studies that employed some form of clinical significance calculation. No other method came close in terms of frequency of use. Because of Jacobson and his colleagues' (Jacobson et al., 1984) original suggestions, a general consensus on a conceptual definition of *clinical significance* has developed: The status of a patient is characterized as clinically significantly changed when the client's level of measured functioning is located in the nonfunctional range at the beginning of treatment and in the functional range at the end of treatment, if that change is statistically reliable. From a

mathematical perspective, there are multiple ways to realize this definition. In this article, we compare the JT statistical approach to four alternative methods.

The JT method to calculate reliable changes has been challenged by authors who believe that alternative methods may be superior. As described previously, the RCI investigates the statistical significance of differences between pretreatment and posttreatment scores for each individual person. Criticism of the original RCI conceptualization was formulated by Hsu (1989) and Speer (1992) and focused on the phenomenon of regression toward the mean that was not taken into consideration by Jacobson and colleagues (Jacobson et al., 1984). Hsu (1989) introduced the Gulliksen–Lord–Novick (GLN) approach that includes the assumption that posttest–pretest regression effects are relevant to the interpretation of posttest scores. The same limitation was noted by Speer (1992). Speer (1992) suggested calculating confidence intervals around the pretreatment scores (± 2 *SD*s) and evaluate the posttreatment scores in relation to this interval. This approach is known as the Edwards–Nunnally (EN; Speer, 1992) method. The most recent approach to assess clinical significance was presented by Hageman and Arrindell (HA; 1999a). In contrast to all other methods, Hageman and Arrindell (1999a) argued that the rates of reliable or clinically significant change of a particular sample should not be calculated by summing up the results of the individual participants of that group. This would result in an underestimation of the true rates of change. As a consequence, Hageman and Arrindell (1999b) suggested different analyses for the individual versus the group rates of change. In addition, Hageman and Arrindell's (1999b) metrics attempt to correct for regression to the mean in providing a closer approximation of the underlying true scores. The postulated enhancement in precision (Hageman & Arrindell, 1999b) was questioned by McGlinchey and Jacobson (1999) who consider the HA approach as "too complex for its value, which has yet to be demonstrated" (p. 1216). Another aspect is formulated by Speer (1992) who generally criticizes the use of two-wave designs to assess change in psychotherapy research. Speer (1992) recommended the use of growth curve modeling (e.g., hierarchical linear modeling [HLM]; Bryk & Raudenbush, 1992) for the study of change. Besides the parameter estimation based on multiwave data, the advantages of HLM are the use of the empirical Bayes estimation and the possibility of handling missing data (Bryk & Raudenbush, 1992; Speer, 1992). Therefore, HLM is supposed to calculate clinically significant change more precisely than the other (two-wave) methods. Speer (1992) argued that if using only pretreatment and posttreatment scores, one should concentrate on the comparability of change rate data among different studies and therefore use the traditional RCI instead of creating new methods.

In addition to theoretical articles on the accuracy of different approaches, two research studies have been conducted to evaluate the degree to which the various methods are redundant or provide different estimates of clinically significant change for the individual patient.

Speer and Greenbaum (1995) were the first to compare Jacobson's (Jacobson & Truax, 1991) method with other approaches. Speer and Greenbaum compared the classification of patients based on the RCI (but not the functional/dysfunctional cutoff) of five methods within a sample of 73 outpatients who were diagnosed with a range of disorders and assessed with a self-report scale of well-being. Speer and Greenbaum used the RCI of the original Jacobson method as summarized by Jacobson and Truax (1991). They chose RCIs from three alternative methods: EN (Speer, 1992), Hsu–Linn–Lord (HLL; Speer & Greenbaum, 1995), and the Nunnally–Kotsch (NK; Nunnally & Kotsch, 1983). These methods, like the original, classify individuals based on pretreatment and posttreatment change scores. They differ from the JT method in that they use residualized pretreatment (or difference) scores rather than raw change scores. This presumably increases precision because of increased reliability (e.g., Rogosa, Brandt, & Zimowski, 1982). Speer and Greenbaum also included in their comparison a method based on HLM that has the presumed advantage of using scores for each patient from more than just two time points.

Speer and Greenbaum (1995) found relatively high rates of agreement between methods (from 77.7% to 81.2%) with the exception of the HLL method, which provided lower rates of recovery and higher rates of deterioration. The HLM method provided the highest improvement rates. Speer and Greenbaum recommended this approach for routine use.

Following this study, McGlinchey, Atkins, and Jacobson (2002) attempted a replication of Speer and Greenbaum (1995) by comparing five methods of estimating clinical significance. McGlinchey et al. included three methods used by Speer and Greenbaum (JT, EN, HLM) and replaced the HLL method with the related GLN approach. Additionally, McGlinchey et al. added a fifth approach, HA, that was proposed by Hageman and Arrindell (1999b). McGlinchey and Jacobson (1999), having previously compared the HA procedures with the Jacobson procedures in couple therapy, concluded that the methods were essentially equivalent. Hageman and Arrindell (1999a), however, criticized these findings and the way in which reliability estimates were selected.

McGlinchey et al. (2002) used outcome data from 128 patients diagnosed with major depressive disorder who participated in one of three cognitive behavioral therapy treatments and whose progress was followed up to 2 years after treatment. Because depressive symptoms were used to exclude potential participants, only those with high depression as measured by the Beck Depression Inventory (BDI; Beck et al., 1961) were included in the study. This analysis used the HA method in place of the NK method used by Speer and Greenbaum (1995). This was done because both methods take into account both pretest and posttest reliabilities in de-

termining reliable change, and the HA method was the newer approach. We followed the same procedure as McGlinchey et al. by including the HA method and excluding the NK method and by replacing the HLL with the GLN approach.

This study partially replicated the work of Speer and Greenbaum (1995) and McGlinchey et al. (2002) by examining five methods of calculating clinically significant change. It differs from both studies by using the Outcome Questionnaire (OQ–45) as the measure of patient functioning. Like the Speer and Greenbaum analysis, this study was based on a general outpatient sample with multiple diagnoses rather than a single disturbance (as was done by McGlinchey et al., 2002) Like McGlinchey et al., clinically significant change was examined rather than limiting the analysis to estimating reliable change, as was done by Speer and Greenbaum. We used a larger sample than either of the previous studies, increasing the likelihood of finding existing differences, but shared the goals of both studies by attempting to examine differences in estimates of clinically significant change that are a function of calculation methods for examining the consequences of using one method instead of another.

METHOD

Participants

In this study, we used data from 386 clients who had sought treatment at a university-based outpatient clinic. Clients ranged in age from 18 to 54 years (M = 22.88, SD = 3.54) and were 66% female, 86% White, 4.8% Latino/Latina, and 9.2% other or mixed ethnicity. Clients were diagnosed by their treating clinician without the benefit of research-based diagnostic evaluations. At intake, 74.6% of clients were diagnosed by their treating clinician with a *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.; American Psychiatric Association, 1994) disorder, whereas 25.4% had their diagnosis deferred and never had a formal diagnosis entered into the database. Those receiving a diagnosis had a mood disorder (29.2%), adjustment disorder (12.4%), anxiety disorder (10.1%), or eating disorder (7.0%). Thirty-five percent of clients had a V-code diagnosis, whereas the remainder (6.3%) received a variety of other diagnoses. Ninety percent of clients had 10 or fewer sessions, with a mean dosage of 8 sessions (range = 2 to 27). The clients in this study started treatment less disturbed (M = 68.35, SD = 22.34) than those in routine outpatient care (M = 80.98, SD = 24.84) but had scores similar to those reported in other university counseling centers (Lambert, Hansen, et al., 1996).

Therapists were 48 university counseling center staff consisting of 27 doctoral level psychologists and 21 doctoral trainees, including interns. Therapists had a variety of treatment orientations, most subscribing to an integration of two or more systems. Licensed therapists did not use manually guided treatments nor were their sessions recorded. They averaged about 14 years of post-doctoral experience. Trainees were supervised on a weekly basis. The most common orientations were cognitive behavioral (50%) and psychodynamic/interpersonal (20%). Psychotherapy was offered free to university student clients, with length of treatment based largely on client needs and preferences.

Outcome Measure

Psychological dysfunction was assessed using the OQ–45 (Lambert, Hansen, et al., 1996), which provided both the measure of weekly change and the criterion measure for classification of patients into outcome groups (Recovered, Improved, No Change, or Deteriorated). The OQ–45 was designed to measure patient progress in therapy by repeated administration during the course of treatment and at termination. In this study, it was administered to clients immediately prior to their first appointment. Each week the clients had a scheduled visit, they came 5 to 10 min early and completed the questionnaire again. The OQ–45 provides a total score based on all 45 items and three subscale scores: subjective discomfort (intrapsychic functioning), interpersonal relationships, and social role performance. Only the OQ–45 total score, which provides a global assessment of patient functioning, was used in our study.

The OQ–45 has been reported to have adequate reliability and validity across a number of settings, including both clinical and normative populations. Research has indicated that the OQ–45 is a psychometrically sound instrument with adequate test–retest reliability at a 3-week interval (r = .84; Lambert, Burlingame, et al., 1996), and excellent internal consistency (Cronbach's α = .93; Lambert, Hansen, et al., 1996). The OQ–45 has also been demonstrated to have acceptable concurrent validity coefficients ranging from .55 to .88 (all significant at $p < .01$) with the Symptom Checklist–90–Revised (Derogatis, 1977), BDI, Zung Self-Rating Depression Scale (Zung, 1965), Taylor Manifest Anxiety Scale (Taylor, 1953), State–Trait Anxiety Inventory (Spielberger, 1983), Inventory of Interpersonal Problems (Horowitz et al., 1988), and the Social Adjustment Scale (Weissman & Bothwell, 1976). Furthermore, the OQ–45 has been shown to be sensitive to change in patients over short time periods while remaining stable in untreated individuals (Vermeersch, Lambert, & Burlingame, 2000).

Using formulas developed by Jacobson and Truax (1991), clinical and normative data for the OQ–45 were analyzed by Lambert, Hansen, et al. (1996) to provide cutoff scores for the RCI and movement from dysfunctional to functional status. Patients who change in a positive or negative direction by at least 14 points are regarded as having made reliable change. This degree of change exceeds measurement error based on the reliability of the OQ–45 and is one of the two criteria posited by Jacobson and Truax as indicating clini-

cally meaningful change. The cutoff on the OQ–45 for demarking the point at which a person's score is more likely to come from the dysfunctional population than a functional population has been estimated to be 64. When a patient's score is 63 or lower, their functioning is considered more similar to nonpatients than patients at that point in time.[1] Passing this cutoff (from dysfunctional to functional) is the second criterion posited by Jacobson and colleagues (Jacobson et al., 1991) as an indicator of clinically significant change. Patients who show reliable change and pass the cutoff are considered Recovered, whereas those who only show reliable change are considered Improved. Support for the validity of the OQ–45's reliable change and clinical significance cutoff scores have been reported by Lunnen and Ogles (1998) and Beckstead et al. (2003).

Procedure

OQ–45s from 386 clients who had pretherapy and posttherapy scores plus at least one additional measurement were used to compare five methods for estimating clinically significant change: three approaches used by Speer and Greenbaum (1995)—the JT and EN, and the HLM approach—as well as the GLN and the HA approach as used by McGlinchey et al. (2002). Computational formulae and details for each method are described in the Appendix. To give a clearer picture of the single methods and to illustrate the calculation procedures, the Appendix includes an example of calculations based on data from a single patient. These calculations show in detail how the RCI is estimated with the four methods using pretreatment and posttreatment scores.

All of the methods assume continuous data. Four of the methods rely exclusively on the use of pretest and posttest scores, whereas the HLM method used pretest, posttest, and all available OQ–45 data points in between.

The average OQ–45 pretest score was 68.35 (SD = 22.33). The mean difference score was 10.9 (SD = 19.18), indicating average gains following treatment that were comparable to gains made by clients seen in similar settings that also have low treatment dosage (Hansen et al., 2002). The corresponding effect size (Cohen's d) for the pretest–posttest change was .48, a modest effect.

Lambert, Hansen, et al. (1996) reported a reliability coefficient for the OQ–45 of .93 (Cronbach's alpha) based on normative data from several patient and nonpatient samples. For reasons explained previously, this value was used for all clinical significance calculations. The resulting S_E amounts to 5.9 (computation formula is included in the Appendix). Just to illustrate the difference, the score was also calculated using test–retest reliability (.84). The corresponding S_E is 8.9. Using this much higher error would lead to more patients being classified as Unchanged.

Following the procedures of McGlinchey et al. (2002), Cutoff C was used for calculation of clinical significance. This cutoff is based on information about both the functional and dysfunctional samples. The HA method calculated a "C True" by multiplying the means by their reliability data, which yielded a cutoff of 57 for that method.

GLN method. In synthesizing the work of Lord and Novick (1968) and Linn and Slinde (1977), Hsu (1989, 1995, 1996) was the first researcher to provide an alternative method to Jacobson et al. (1984), referred to as the GLN method. Hsu (1989) suggested that the RCI index of the Jacobson et al. (1984) method did not take into account the phenomenon of regression to the mean that might considerably influence outcome classifications. Regression to the mean implies that more extreme scores because of imperfect reliability, will naturally tend to become less extreme over repeated assessments. The GLN method attempts to control for this potential confound by including a hypothesized population mean (and standard deviation) toward which scores would regress. These are suggested to be the scores of a "relevant group," for example, the group from which the participants of the study were selected (Hsu, 1999). If there is not such a population or their scores are not available, Hsu (1999) advocated using the pretreatment scores of the participants. This was done in our study.

EN method. The EN method was presented by Speer (1992). This method synthesizes the work of Edwards, Yarvis, Mueller, Zingale, and Wagman (1978) and Nunnally (1967, 1975) who advocated formulation of confidence intervals for calculating prechange to postchange rates. The EN method establishes reliable change by observing a participant's posttest score relative to an established confidence interval around the estimated true pretreatment score of the individual. Speer concluded that, like the GLN method, the EN approach would be an improvement on the original clinical significance method by minimizing the influence of regression to the mean in the calculation of improvement rates. Fur-

[1]The "original" cutoff (63 points) was calculated using the samples reported in the OQ–45 manual, that is, a clinical sample with a mean score of 79.8 (SD = 25.3) and a nonclinical sample with a mean score of 48.7 (SD = 20.2). In this study, we used two other samples in which the mean was 68.35 (SD = 22.33) for the clinical and 49.04 (SD = 17.3) for the nonclinical sample. The nonclinical data are comparable to those used in the OQ–45 manual, with similar impairment. However, the patients in our study were less impaired than the clinical sample in the manual. As a consequence of this lower impairment of the patient sample, Cutoff C needed modification and resulted in a lower cutoff score. Cutoff C is the weighted midpoint between the means of a functional and dysfunctional population and therefore always "reacts" directly to the samples used in a specific study.
The calculation of C = 57 was done as follows:

$$\text{Cutoff C} = \frac{(SD_{clinical} M_{nonclinical}) + (SD_{nonclinical} M_{clinical})}{SD_{clinical} + SD_{nonclinical}}$$
$$= \frac{2233 \times 49.04 + 17.3 \times 68.35}{22.33 + 17.3} = 57.4.$$

TABLE 1
Rates (Percentages and Frequencies) of Reliable Change Across Five Methods of Calculating Reliable Change Using the Total Sample

	Deteriorated		*Unchanged*		*Improved*	
Approach	*%*	*Frequency*	*%*	*Frequency*	*%*	*Frequency*
Jacobson–Truax	6.5	25	58.5	226	35.0	135
Gulliksen–Lord–Novick	5.2	20	59.3	229	35.5	137
Edwards–Nunnally	11.9	46	42.0	162	46.1	178
Hageman–Arrindell	8.8	34	66.6	257	24.6	95
HLM	4.9	19	65.3	252	29.8	115

Note. $N = 386$. HLM = hierarchical linear modeling.

thermore, the ease of presentation offered by confidence intervals is an additional benefit of this method.

HA method. This is one of the most recent clinical significance methods, developed by Hageman and Arrindell (1999b). Drawing on Cronbach and Gleser's (1959) use of the phi coefficient as a measure of discrimination, the HA method involves the most significant revisions to the JT method. Among its distinguishing features, the HA method differentially analyzes clinically meaningful change at the individual level (i.e., participant to participant) and at the group level (i.e., obtaining proportions of participants in the sample who have reliably changed and passed the cutoff point). The RCI (RC_{INDIV}) of the method is determined by incorporating both pretest and posttest reliabilities in its calculations, purporting to enhance precision further. In addition, the HA method is the first to modify the cutoff criterion, applying the same corrections for regression to the mean to the cutoff as are used in the RC_{INDIV}.

HLM. Speer and Greenbaum (1995) recently advocated a multiwave data approach using growth curve modeling (e.g., HLM; Bryk & Raudenbush, 1992). One of the advantages of a multiwave approach is that it uses more than two data points per individual; by doing so, it reflects the change that occurs between pretest and posttest assessments more precisely. Besides the parameter estimation based on multiwave data, further advantages of HLM are the use of empirical Bayes estimates, which are weighted estimates that combine information from the individual and the sample as a whole as well as the capability of handle missing data.

Following calculations of the clinical significance of change for each method, differences between methods were analyzed through the use of nonparametric statistics.

RESULTS

Table 1 presents the overall rates of classifying the client's reliable change for the five clinical methods for the total sample of 386 clients. The EN method classified the fewest number of clients as unchanged (42%) and the greatest percentage of clients as improved (46%) and deteriorated (12%). The HA and HLM methods classified the largest number of clients as unchanged (67% to 65%) and the smallest number as improved. There was little difference between the JT and GLN methods in classifying patients as reliably changed.

TABLE 2
Paired Comparisons Between Reliable Change Classifications

Approach	*JT*	*GLN*	*EN*	*HA*	*HLM*
JT	—	.92	.71	.76	.78
GLN	*ns*	—	.70	.72	.80
EN	$p < .01$	*ns*	—	.59	.59
HA	$p < .001$	$p < .001$	$p < .001$	—	.67
HLM	$p < .05$	$p < .001$	$p < .001$	$p < .001$	—

Note. Significance levels for paired comparisons (Wilcoxon test) between the reliable change classifications of the five approaches are underlined. 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; HLM = hierarchical linear modeling.

To compare differences in classification rates, Kendall's coefficient of concordance statistic (W) was computed. This omnibus test is distributed as a chi-square. W was highly significant for the five methods ($W = .064$), $\chi^2(4, N = 386) = 98.935$, $p < .001$, suggesting that overall the methods differ from one another in classifying the rate of reliable change. Pairwise comparisons of all methods were calculated using the Wilcoxon signed ranks test based on Z scores. This nonparametric test measured level of agreement by comparing perfect agreement between every two methods' classification of individuals compared to nonmatches. Table 2 presents the results. Statistically significant differences were observed between 8 of the 10 pairs. The GLN classifications did not differ from the JT or from the EN classifications.

To quantify the extent of agreement across methods, Cohen's kappa coefficients were calculated. Table 2 shows the results in the right upper quadrant. All 10 coefficients reached statistical significance at the .001 level. The highest agreement was found between the JT and GLN approach ($\kappa = .92$) and the lowest between the EN and the HA and the EN and the HLM approach ($\kappa = .59$).

TABLE 3
Percentages and Frequencies of Clinical Significance Classification for Five Calculation Methods

	CS Deteriorated		*Deteriorated*		*Unchanged*		*Improved*		*Recovered*	
Approach	*%*	*Frequency*	*%*	*Frequency*	*%*	*Frequency*	*%*	*Frequency*	*%*	*Frequency*
Jacobson–Truax[a]	2.3	9	4.1	16	58.5	226	16.1	62	18.9	73
Gulliksen–Lord–Novick	2.1	8	3.1	12	59.3	229	16.6	64	18.9	73
Edwards–Nunnally	3.1	12	8.8	34	42.0	162	24.9	96	21.2	82
Hageman–Arrindell	5.7	22	3.1	12	66.6	257	12.7	49	11.9	46
HLM	1.6	6	3.4	13	65.3	252	13.5	52	16.3	63

Note. $N = 386$. CS = clinical significance; HLM = hierarchical linear modeling.
[a]The Jacobson–Truax categories have been slightly modified. In effectiveness research (rather than clinical trials), a portion of persons (15% to 25%) begin treatment in the functional range and cannot meet the criterion for passing the cutoff into the functional range (achieving CS change), although they can meet reliable change index criterion for improvement and deterioration. Here, "CS deteriorated" indicates patients who began treatment in the functional range and reliably worsened, ending therapy in the dysfunctional range. "Deteriorated" means reliably worsened.

TABLE 4
Paired Comparisons Between Clinically Significant Change Classifications

Approach	*JT*	*GLN*	*EN*	*HA*	*HLM*
JT	—	.93	.75	.55	.80
GLN	*ns* *ns*	—	.74	.52	.82
EN	$p < .001$ $p < .05$	$p < .001$	— *ns*	.41	.64
HA	$p < .001$ $p < .05$	$p < .001$ $p < .001$	$p < .001$ $p < .001$	—	.49
HLM	$p < .001$ *ns*	$p < .001$ ns	$p < .001$ $p < .05$	$p < .001$ *ns*	—

Note. Significance levels for paired comparisons (Wilcoxon test) between the clinically significant change classifications of the five approaches: Results of current study (underlined) and results from McGlinchey, Atkins, and Jacobson (2002). 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; HLM = hierarchical linear modeling.

Table 3 presents the overall rates of classifying patient ($N = 386$) change based on clinical significance criteria. In this comparison, the omnibus test for differences was again highly significant ($W = .124$), $\chi^2(4, N = 386) = 191.01$, $p < .001$. Pairwise comparisons showed statistically significant differences between all classification methods except the JT method and the GLN method (Table 4).

Again we also calculated kappa coefficients to quantify the extent of pairwise agreement across methods. The coefficients are displayed in the right upper quadrant of Table 4. Again, all coefficients reached statistical significance (all $ps < .001$). As for the comparisons of reliable change rates, agreement was highest between the JT and the GLN approach ($\kappa = .93$) and lowest between the EN and the HA approach ($\kappa = .41$).

Table 5 presents the pairwise rates of agreement as well as the average agreement of each method with the remaining four approaches. As Table 5 shows, there was high concordance between the JT and GLN methods (96%). The HA method appeared to be most discrepant from the other methods, providing relatively low rates of improvement, relatively high rates of deterioration and no change, and agreement rates that varied from 61% to 75% (see Tables 3 and 5). The rates shown in Table 3 for clinically significant change parallel those found for reliable change as presented in Table 1. The EN method yields the most liberal estimate of clinically significant change, whereas the HA method was the most conservative in classifying patients as meeting criteria for change.[2] The HLM method, found to be "more sensitive to change" by Speer and Greenbaum (1995), provided relatively low rather than high estimates of individual change and rates of agreement that varied from 73% to 90 %.

DISCUSSION

Examining differences in classification frequency between methods of estimating clinically significant change is an important research task. 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. Such estimates have an impact on practitioners, researchers, and policymakers, as these individuals make policy decisions and recommendations for preferred practices. In addition, the presence or absence of clinically significant change can impact treatment for the individual client, that is, whether to terminate treatment or continue treatment, stepping up (or down) treatment to more (or less) intense interventions, or referral of the client. These applied clinical activities are dependent on a number of practical considerations such as the specific outcome measures that are

[2]In addition to the analyses of the total sample, the calculation of for rates of reliable change was also done with the sample broken into two subgroups (the 258 clients who began treatment in the dysfunctional range, i.e., above the cutoff score 57 and the 128 who began treatment in the functional range, i.e., below the cutoff score of 57). The results were the same for how the methods performed. The subsamples did differ in the extent to which they improved, with less improvement found in the more healthy clients.

TABLE 5
The Percentage of Agreement Between Patient Classification Across Five Methods of Calculating Clinically Significant Change

Approach	*JT*	*GLN*	*EN*	*HA*	*HLM*
JT	—	—	91.8	—	84.9
GLN	95.6	—	—	—	—
EN	73.4	82.7	—	—	89.0
HA	74.6	73.1	60.9	—	—
HLM	88.7	90.0	76.2	73.2	—
Average agreement	83.1	85.4	73.3	70.5	82.0

Note. Results from this study are underlined. Those reported by Speer and Greenbaum (1995), who used only three of the five methods (JT, EN, and HLM), are in the upper right quadrant. McGlinchey, Atkins, and Jacobson (2002) did not report the amount of pairwise agreement. JT = Jacobson–Truax; GLN = Gulliksen–Lord–Novick; EN = Edwards–Nunnally; HA = Hageman–Arrindell; HLM = hierarchical linear modeling.

used in a study (Beckstead et al., 2003), but as this study shows, they also depend on the statistical methods used to estimate clinically significant change.

In this study, we attempted to increase the reseacher and clinician's understanding of the impact of statistical methods that classify clinically significant change by examining five methods of calculating change using a relatively large data set based on routine clinical practice. Two previous studies (McGlinchey et al., 2002; Speer & Greenbaum,1995) on this topic differed from our study in important ways. Speer and Greenbaum's study also studied routine practice but examined only reliable change and did not use all the same methods used here and by McGlinchey et al. Unlike this study, the McGlinchey et al. study was limited to a relatively small number of patients within a single diagnostic category. McGlinchey et al. were forced to collapse data across outcome categories because of limited patient numbers in the deteriorated category. Thus, McGlinchey et al. were unable to fully test the classification categories. Our study used a much larger *N* than either of the prior studies, maximizing the possibility of finding differences if they existed. Nevertheless, examining results across the three studies provides concrete examples of the consequences of using different statistical techniques for classifying patient outcomes.

Results of this study indicate differences in estimates of clinically significant change between all methods except JT and GLN. Both methods produced relatively moderate estimates of clinically significant change (18.9%). Similarity in estimates by the JT and GLN methods was also reported by McGlinchey et al. (2002). Agreement in classification of cases by these methods was above 90% in McGlinchey et al. as well as in this study. This agreement in improvement rates could not be compared with rates in the study by Spear and Greenbaum (1995) who compared the JT and HLL instead of the GLN method for calculating reliable change. It is interesting to note that the JT method produced improvement rates very similar to those of GLN method advocated by Hsu (1996) based on criticisms of the JT method.

The results of findings on the HLM method were somewhat divergent across studies. This study found that the HLM method produced relatively low clinically significant change.[3] McGlinchey et al. (2002) found a similar positioning among methods, although it was only statistically different from the EN method, which produced the highest change rates. Speer and Greenbaum (1995), on the other hand, found the HLM method to differ sharply from other methods by producing the highest reliable change (77%), whereas the other methods estimated reliable change to be from 55% to 63%. The more liberal estimates of reliable change found by Speer and Greenbaum using HLM were not replicated here, nor were they replicated by McGlinchey et al. It is not clear why Speer and Greenbaum's estimates are at odds with our findings (and those reported by McGlinchey et al., 2002), but the findings of this study coupled with those from McGlinchey et al. argue against acceptance of Speer and Greenbaum's recommendation that HLM is the preferred method based on the relatively high improvement rates for HLM's supposed "sensitivity" to treatment effects. As Hsu (1999) pointed out, it is inappropriate to recommend a method based on its production of high improvement rates because the accuracy of rates is the standard that should be used. Speer and Greenbaum assumed that HLM is more accurate only because of its statistical sophistication and use of more than two data points per client. Despite being based on more than two scores, HLM does not appear to provide especially high estimates of clinically significant change or to be particularly sensitive to detecting change.

The EN method provided the most liberal estimates of clinically significant change and/or reliable change in this study (21%). These estimates were dramatically different from HA estimates (12%) based on the work of Hageman and Arrindell (1999b). Such large differences suggest considerable disagreement between these methods in estimating which patients have experienced a clinically significant outcome. Measures of concordance suggested only 61% agreement, the lowest of all pairs. The EN method rated reliable change rates at the second highest level in the investigation by Speer and Greenbaum (1995) and also high in the McGlinchey et al. (2002) analysis. When figures for improved are added to those for clinically significant change in this study, the EN method appears to show an even stronger

[3]Using HLM allows for the investigation of complex models such as curvilinear growth models that include linear and quadratic polynomial components. In this study, only the results of the linear model are described because the linear model was superior to the curvilinear model (i.e., it accounted for more within-subjects variance). In addition, Speer and Greenbaum (1995) recommended the use of the linear model because the results are more comparable to the pretreatment–posttreatment difference score methods.

trend toward producing high rates of improvement. It also classified the greatest number of clients as Deteriorated.

In this study, the HA method was especially unlikely to categorize many patients as improved (but likely to categorize patients as deteriorated in relation to other methods). Similar findings were reported by McGlinchey et al. (2002) who found the HA method produced the lowest estimates of improvement. Most clients who started treatment were estimated to be unchanged at the time of termination.

In our study, three of the five measures (JT, GLN, HLM) provided comparable results, but the other two (EN, HA) deviated in opposite directions. In the Speer and Greenbaum (1995) analysis, three of the four methods studied were comparable, with HLM providing significantly higher estimates of improvement. McGlinchey et al. (2002) reported comparability of estimates of change across four of their five methods, with HA providing the lowest estimates. When combined, the results of these three studies indicate a trend that may hold up across patient samples. Methods of calculating clinical significance are not completely comparable. The EN method appears to produce the most liberal estimates of change, whereas the HA method is the most conservative. The relative standing of measures may be sample dependent.

We argue for the value of using the JT (Jacobson & Truax, 1991) method; it is already the most popular method, it is easy to compute, cutoff estimates are available for a number of widely used instruments (e.g., BDI: Seggar, Lambert, & Hansen, 2002; Symptom Checklist–90–R: Tingey, et al. 1996b; which eliminates the need for calculation), and it provides a moderate point between the extremes of the EN and HA methods. If the JT method is used with measures that have large normative samples, then standard cutoff and RCI values can be applied uniformly by researchers across the country rather than requiring sample- and study-specific recalculations. This means that a standard for clinically significant change can be set and applied regardless of the research or clinical setting in which the study is conducted. However, this is true only on the condition that the distributions of patient and nonpatient data are comparable. Strictly speaking, the description of the situation (with respect to a functional and dysfunctional population) has to be repeated in every study. Practically speaking, in this data, it made little difference whether the JT or GLN methods were employed; they resulted in similar estimates.

Future research is needed on the validity of estimates of clinically significant change that are provided by different statistical methods. Although this study (and past research) does provide important information about differences and similarities between statistical methods applied to clinical samples, it does not clarify which methods are most accurate in representing clinically meaningful change. Ultimately, validity data are needed to evaluate which statistical estimate most accurately reflects meaningful change. Such research presents a challenge because there is no agreed on standard for this kind of therapeutic impact. Beckstead et al. (2003) compared change across four popular outcome measures given to the same patients to see if patients classified by the JT method on one measure were classified the same on a different measure. Lunnen and Ogles (1998), as well as Ankuta and Ables (1993), using the JT method, found some correspondence between patient satisfaction and reliable change, but no research has been published using the other statistical methods that were examined in this study. Research that applies several measures and statistical methods simultaneously with a large sample of clients is needed. In addition, qualitative methods that carefully study a small sample of these clients may also help further our understanding of what the thresholds are for clinically significant change and which statistical method most closely agrees with this type of criterion. However, this type of large-scale research is not very practical if it includes the HLM method because this method requires frequent retesting. If multiple measures were used in such research, the burden on clients would be enormous.

Finally, two critical issues related to the application of statistical methods for estimating clinical significance need to be mentioned. First, in general, it is not possible to solve theory-based questions with an empirical study (e.g., if and to what extent the different methods actually differ). One can only investigate whether the approaches lead to different results in certain situations (as was done in this study). The second issue concerns the usage of reliability scores. From a strict methodological perspective, it is not proper to apply reliability information to single cases. With respect to this argument, one should favor these statistics based on reliability with groups of clients when comparing the results of different studies rather than with the individual client as advocated by Hageman and Arrindell (1999b). As a consequence, the "most correct" approach for single outcomes would be the HLM approach; it is the only method that does not include reliability scores. The HLM analysis uses information from multiwave courses to estimate a kind of individual reliability. However, as was said previously, this supposed advantage of the HLM approach did not lead to strikingly different results. Given that this method is very demanding (and virtually impossible for clinicians to employ), it seems justifiable to advocate the use of the traditional JT approach that would guarantee easy applicability for clinicians as well as researchers. It would also help to insure more comparability across studies, albeit at the expense of losing some methodological accuracy. Continued research on methods of calculating clinical significance is recommended and we caution the reader that until further research is conducted, including validity studies, we cannot be certain of the best method for estimating clinically significant change.

ACKNOWLEDGMENT

This research was supported by cooperation grants from the German American Academic Council and Brigham Young University.

REFERENCES

American Psychiatric Association. (1994). *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.). Washington, DC: Author.

Anderson, E. M., & Lambert, M. J. (2001). A survival analysis of clinically significant change in outpatient psychotherapy. *Journal of Clinical Psychology, 57,* 875–888.

Ankuta, G. Y., & Ables, N. (1993). Client satisfaction, clinical significance, and meaningful change in psychotherapy. *Professional Psychology: Research and Practice, 24,* 70–74.

Beckstead, D. J., Hatch, A. L., Lambert, M. J. Eggett, D. L., Vermeersch, D. A., & Goates, M. K. (2003). Clinical significance of the Outcome Questionnaire(OQ–45.2) *Behavioral Analyst Today, 4,* 74–90.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives for General Psychology, 4,* 53–63.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods.* Newbury Park, CA: Sage.

Cronbach, L. J., & Gleser, G. C. (1959). Interpretation of reliability and validity coefficients: remarks on a paper by Lord. *Journal of Educational Psychology, 50,* 230–237.

Derogatis, L. R. (1977). *The SCL–90 manual: Scoring, administration and procedures for the SCL–90.* Baltimore: Johns Hopkins University School of Medicine, Clinical Psychometrics Unit.

Edwards, D. W., Yarvis, R. M., Mueller, D. P., Zingale, H. C., & Wagman, W. J. (1978). Test-taking and the ability of adjustment scales; Can we assess patient deterioration? *Evaluation Quarterly, 2,* 275–292.

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

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

Hansen, N., Lambert, M. J., & Forman, E. M. (2002). Comparisons of clinically significant change in clinical trials and naturalistic practice settings: The dose-effect relationship and its implication for practice. *Clinical Psychology: Science and Practice,9,* 329–343.

Horowitz, L. M., Rosenberg, S. E., Baer, B. A., Ureno, G., Villasenor, V. S. (1988). Invenotry of interpersonal problems: Psychometric properties and clinical applications. *Journal of Consulting and Clinical Psychology, 56,* 885–892.

Howard, K. I., Lueger, R. J., Maling, M. S., & Martinovich, Z. (1993). A phase model of psychotherapy outcome: Causal mediation of change. *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. (1995). Regression toward the mean associated with measurement error and the identification of improvement and deterioration in psychotherapy. *Journal of Consulting and Clinical Psychology, 63,* 141–144.

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.

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

Jacobson, N. S. (1988). Defining clinically significant change: An introduction. *Behavioral Assessment, 10,* 131–132.

Jacobson, N. S., Follette, W. C., & Revenstorf, D. (1984). Toward a standard definition of clinically significant change. *Behavior Therapy, 17,* 308–311.

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

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.

Kendall, P. C. (1999). Clinical significance. *Journal of Consulting and Clinical Psychology, 67,* 283–284.

Kendall, P. C., Marrs-Garcia, A., Nath, S. R., & Sheldrick, R. C. (1999). Normative comparisons for the evaluation of clinical significance. *Journal of Consulting and Clinical Psychology, 67,* 285–299.

Lambert, M. J., Burlingame, G. M., Umphress, V., Hansen, N. B., Yancher, S. C., Vermeersch, D. A., et al. (1996). The reliability and validity of a new psychotherapy outcome questionnaire. *Clinical Psychology and Psychotherapy, 3,* 249–258.

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 (OQ45.2).* Wilmington, DE: American Professional Credentialing Services.

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

Lambert, M. J., Whipple, J. L., Smart, D. W., Vermeersch, D. A., Nielsen, S. L., & Hawkins, E. J. (2001). The effects of providing therapists with feedback on patient progress during psychotherapy: Are outcomes enhanced? *Psychotherapy Research, 11,* 49–68.

Linn, R. L., & Slinde, J. A. (1977). A determination of the significance of change between pre-and post-testing periods. *Review of Educational Research, 47,* 121–150.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

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

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

McGlinchey, J. B., Atkins, D. C., & Jacobson, N. S. (2002). Clinical significance 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. *Behavior Research and Therapy, 37,* 1211–1217.

Nunnally, J. C. (1967). *Psychometric theory.* New York: McGraw-Hill.

Nunnally, J. C. (1975). The study of change in evaluation research: Principles concerning measurement, experimental design and analysis. In E. L. Streuning & M. Guttentag (Eds.), *Handbook of evaluation research* (pp. 203–236). Beverly Hills, CA: Sage.

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

Rogosa, D., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin, 92,* 726–748.

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

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

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

Spielberger, C. D. (1983). *Manual for the State–Trait Anxiety Inventory STAI (Form Y).* Palo Alto, CA: Consulting Psychologists Press.

Taylor, J. A. (1953). A personality scale of manifest anxiety. *Journal of Abnormal and Social Psychology, 48,* 285–290.

Tingey, R., Lambert, M. J., Burlingame, G., & Hansen, N. (1996a). Assessing clinical significance: Proposed extensions to method. *Psychotherapy Research, 6,* 109–123.

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

Vermeersch, D. A., Lambert, M. J., & Burlingame, G. M. (2000). Outcome Questionnaire 45: Item sensitivity to change. *Journal of Personality Assessment, 74,* 242–261.

Weissman, M. M., & Bothwell, S. (1976). Assessment of social adjustment by patient self-report. *Archives of General Psychiatry, 33,* 1111–1115.

Zung, W. W. K. (1965). A self-rating depression scale. *Archives of General Psychiatry, 12,* 63–70.

APPENDIX

As previously noted, the example of one single patient demonstrates the calculation of the different reliable change indexes (RCIs). The following information is relevant for these calculations.

Information Based on the Sample

Pretreatment mean for sample: $M_{pre} = 68.35$.
Pretreatment standard deviation for sample: $SD_{pre} = 22.33$.
Posttreatment mean for sample: $M_{post} = 57.45$.
Posttreatment standard deviation for sample: $SD_{post} = 22.94$.
Internal consistency: Cronbach's $\alpha = .93$.
Standard error of measurement (S_E): $S_E = SD\sqrt{1 - r_{xx}} = 5.9$.
Correlation between pretreatment and posttreatment scores: $r_{pre*post} = .641$.

Information Based on One Single Patient

Pretreatment score: $x_{pre} = 77$.
Posttreatment score: $x_{post} = 52$.

Jacobson and Truax (1991)

RCI: $$\frac{(x_{post} - x_{pre})}{\sqrt{2S_E^2}} = \frac{(52 - 77)}{\sqrt{2 \times 5.9^2}} = -2.996.$$

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

Gulliksen, Lord, and Novick (Hsu, 1989)

RCI:
$$\frac{(x_{post} - M_{pre}) - r_{xx}(x_{pre} - M_{pre})}{SD_{pre}\sqrt{1 - r_{xx}^2}} = \frac{(52 - 68.35) - .93(77 - 68.35)}{22.33\sqrt{1 - .93^2}} = -2.971.$$

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

Edwards and Nunnally (Speer, 1992)

RCI:
$$[r_{xx}(x_{pre} - M_{pre}) + M_{pre}] \pm 2SD_{pre}\sqrt{1 - r_{xx}}$$
$$= [.93(77 - 68.35) \pm 2 \times 22.33\sqrt{1 - .93} = [76.3945] \pm 3.1262,$$

where the upper boundary = 76.3945 + 3.1262 = 79.52 and the lower boundary = 76.3945 – 3.1262 = 73.27.
The patient is classified as reliably improved because the posttreatment score ($x_{post} = 52$) is below the lower boundary.

Hageman and Arrindell (1999b)

Individual RCI:
$$\frac{(x_{post} - x_{pre})r_{dd} + (M_{post} - M_{pre})(1 - r_{dd})}{\sqrt{r_{dd}}\sqrt{2S_E^2}} =$$
$$\frac{(52 - 77) \times .805 + (57.45 - 68.35)(1 - .805)}{\sqrt{.805}\sqrt{2 \times 5.9^2}} = -2.97.$$

The patient is classified as reliably improved because the score is smaller than –1.65.

$$r_{dd} = \frac{SD_{pre}^2 r_{xx(pre)} + SD_{post}^2 r_{xx(post)} - 2SD_{pre}SD_{post}r_{pre \times post}}{SD_{pre}^2 + SD_{post}^2 - 2SD_{pre}SD_{post}r_{pre \times post}}$$
$$= \frac{22.33^2 \times .93 + 22.94^2 \times .93 - 2 \times 22.33 \times 22.94 \times .641}{22.33^2 - 22.94^2 - 2 \times 22.33 \times 22.94 \times .641} = .805.$$
$$r_{xx(pre)} = \frac{SD_{pre}^2 - S_E^2}{SD_{pre}^2} = \frac{22.33^2 - 5.9^2}{22.33^2} = .93.$$
$$r_{xx(post)} = \frac{SD_{post}^2 - S_E^2}{SD_{post}^2} = \frac{22.94^2 - 5.9^2}{22.94^2} = .93.$$

Individual clinical significance index:
$$\frac{M_{post} + (x_{post} - M_{post})r_{xx(post)} - TRC}{\sqrt{r}\sqrt{r_{xx(post)}}S_E}$$
$$TRC = \frac{SD_{norm}\sqrt{r_{xx(norm)}}M_{pre} + SD_{pre}\sqrt{r_{xx(pre)}}M_{norm}}{SD_{norm}\sqrt{r_{xx(norm)}} + SD_{pre}\sqrt{r_{xx(pre)}}}.$$

Hierarchical Linear Modeling (Bryk and Raudenbush, 1992)

$$\frac{B^*}{\sqrt{V^*}},$$

x_{pre} = individual's raw pretreatment score; x_{post} = individual's raw posttreatment score; SD_{pre} = standard deviation of pretreatment scores; SD_{post} = standard deviation of posttreatment scores; S_E = standard error of measurement; r_{xx} = reliability of measure; M_{pre} = mean of pretreatment scores; M_{post} = mean of posttreatment scores; r_{dd} = reliability of difference scores; $r_{xx(pre)}$ = reliability of pretreatment scores; $r_{xx(post)}$ = reliability of posttreatment scores; TRC = true cutoff score (here on basis of Jacobson's [Jacobson & Truax, 1991] Cutoff C criterion); SD_{norm} = standard deviation of normal or nonpatient population; $r_{xx(norm)}$ = reliability of normal or nonpatient scores; M_{norm} = mean of normal or nonpatient scores; B^* = empirical Bayes estimate of linear slope; $V^{*1/2}$ = standard deviation of the empirical Bayes estimate.

Michael J. Lambert
Department of Psychology
Brigham Young University
272 TLRB
Provo, UT 84602
michael_lambert@byu.edu

Received July 12, 2003
Revised July 30, 2003