Journal of Consulting and Clinical Psychology
2005, Vol. 73, No. 5, 982–989

0022-006X/05/$12.00 DOI: 10.1037/0022-006X.73.5.982

BRIEF REPORT

Assessing Clinical Significance: Does It Matter Which Method We Use?

David C. Atkins and Jamie D. Bedics
Fuller Graduate School of Psychology

Joseph B. McGlinchey
Brown University

Theodore P. Beauchaine
University of Washington

Measures of clinical significance are frequently used to evaluate client change during therapy. Several alternatives to the original method devised by N. S. Jacobson, W. C. Follette, & D. Revenstorf (1984) have been proposed, each purporting to increase accuracy. However, researchers have had little systematic guidance in choosing among alternatives. In this simulation study, the authors systematically explored data parameters (e.g., reliability of measurement, pre–post effect size, and pre–post correlation) that might yield differing results among the most widely considered clinical significance methods. Results indicated that classification across methods was far more similar than different, especially at greater levels of reliability. As such, the existing methods of clinical significance appear highly comparable; future directions for clinical significance use and research are discussed.

Keywords: clinical significance, psychotherapy outcome, psychometrics

One of the difficulties that clinical researchers face is assessing whether their treatments have a "meaningful" impact on their clients. Outcome measures typically focus on symptoms of psychopathology, and increasingly complicated statistical methods are used to determine whether the null hypothesis of no effect should be rejected. However, even when significant change in outcome is detected, we may still know little about whether clients returned to normal functioning or to what extent their lives were positively altered. Such questions address clinical rather than statistical significance (Ogles, Lunnen, & Bonesteel, 2001).

Various methods have been advanced to measure the clinical significance of treatments, including comparisons with normal controls (Kendall, Marrs-Garcia, Nath, & Sheldrick, 1999), assessments of quality of life (Gladis, Gosch, Dishuk, & Crits-Christoph, 1999), and classification of patients into deteriorated, unchanged, improved, or recovered categories (Jacobson, Follette, & Revenstorf, 1984; Jacobson & Truax, 1991). Among these, the Jacobson et al. method remains the most popular. Ogles et al. (2001) reported that 53% of studies reporting some type of clinical significance in the *Journal of Consulting and Clinical Psychology* during a 9-year period used the Jacobson et al. method or a variation thereof. Classification using the Jacobson and Truax (1991) method combines information about an individual's pre- to posttherapy functioning on an outcome measure with normative information about the measure.

However, this method has been criticized on statistical grounds, and several alternative formulae exist for computing the clinical significance classifications, each intending to yield more accurate estimates of meaningful change than the traditional method (Hageman & Arrindell, 1999; Hsu, 1999; Speer, 1992).[1] For the psychotherapy researcher, a critical question is, "Does it matter which method I use to analyze my data?"

Three previous studies have compared various combinations of the original and alternative methods. Speer and Greenbaum (1995) used data from a geriatric outpatient mental health clinic ($N = 73$) to compare five methods of calculating clinical significance.[2] McGlinchey, Atkins, and Jacobson (2002) used data from a trial involving treatments for depression ($N = 129$) to compare five methods and to examine the predictive validity of the classifications of each method. Finally, Bauer, Lambert, and Nielsen (2004) compared the same five methods as McGlinchey et al. using 386 patients from an outpatient clinic. Although these studies offered comparisons of alternative methods, each used a single data set and, because of discrepancies in findings, the results did not provide clear guidance for psychotherapy researchers. Commenting on this state of affairs, several authors have suggested ceasing

David C. Atkins and Jamie D. Bedics, Travis Research Institute, Fuller Graduate School of Psychology; Joseph B. McGlinchey, Department of Psychiatry and Human Behavior, Brown University; Theodore P. Beauchaine, Department of Psychology, University of Washington.

Correspondence concerning this article should be addressed to David C. Atkins, Travis Research Institute, Fuller Graduate School of Psychology, 180 N. Oakland Avenue, Pasadena, CA 91101. E-mail: datkins@fuller.edu

[1] The conceptual and statistical differences between methods are described in the Methods section and the Appendix.

[2] Speer and Greenbaum (1995) only used the Reliable Change Index in their comparisons, which is one of two parts that Jacobson and Truax (1991) recommended in their method.

the development of new methods until more is understood about how current approaches perform (McGlinchey et al., 2002; Speer, 1999).

The current simulation study was designed to assist psychotherapy researchers in making informed choices about which method of clinical significance to use in their research. Using simulated data, we were able to systematically explore the performance of the original and alternative methods across combinations of various data parameters (e.g., effect sizes, reliabilities, and pre–post correlations). It should be noted at the outset that our study did *not* answer the question of which method is "better" or more "accurate," as this task requires an external criterion with which to compare methods. Because of the difficulties in defining such a criterion objectively, the current study focused on the more fundamental question of whether the methods differ at all in their classifications and, if so, when.

Method

The current study compared four methods of clinical significance that use pre- and posttherapy data to render classifications (Jacobson & Truax, 1991; Hageman & Arrindell, 1999; Hsu, 1999; Speer, 1992).[3] The four methods are described below; formulae can be found in the Appendix.

Jacobson-Truax (JT) Method

The JT method was originally proposed by Jacobson et al. (1984) and contains two steps. The first step is to define a cutoff point that separates the "functional" population from the "dysfunctional" population. Jacobson et al. proposed three different cutoffs depending on whether normative information was available for the outcome measure. For the present simulation study, we used Cutoff A, which specifies the "functional" population as those with posttherapy scores that are 2 *SD*s or more from the pretreatment mean. The second step compares an individual's change from pre- to posttherapy to the standard error (*SE*) of measurement of the outcome (i.e., $\pm$ 1.96 *SE*), referred to as the Reliable Change Index (RCI). These two steps are used to classify individuals into one of four categories including: recovered (individual has passed Cutoff A and RCI in the positive direction), improved (has passed RCI in the positive direction but not Cutoff A), unchanged (has passed neither criterion), or deteriorated (has passed RCI in the negative direction).

Gulliksen-Lord-Novick (GLN) Method

Hsu (1989, 1999) criticized the original method devised by Jacobson et al. (1984) for failing to account for regression to the mean and altered the RCI formula to include estimates of the population mean and *SD* toward which scores would be expected to regress. The principal limitation of this method is that population means and *SD*s are rarely known. Hsu (1999) recommended use of the pretreatment mean in the absence of a better estimate; this was the estimate used in the present study.

Edwards-Nunnally (EN) Method

Speer (1992) criticized the original method for reasons similar to those of Hsu (1989), and the EN method addresses regression to the mean by "shrinking" pretherapy scores toward the pretherapy mean by the reliability of the measure. This estimated true score is then placed at the center of a confidence interval so that estimates can be made of the significance of posttherapy change, that is, two *SE*s from the adjusted center.

Hageman-Arrindell (HA) Method

Hageman and Arrindell (1999) offered the most significant revision thus far of the JT method. Based on the work of Cronbach and Gleser (1959), the HA method presented new indices to assess the clinical significance of change (CS_{indiv}) and the reliability of change (RC_{indiv}). The CS_{indiv} index is an attempt to provide a more precise cutoff through a number of modifications including the use of true score mean equivalents and reliability coefficients to account for measurement error. Similar to the EN method, RC_{indiv} also takes into account the reliability of measures at pre- and posttherapy. The HA method also provides formulae for the calculation of group-based analyses, allowing researchers to present cumulative results for the number of cases classified, though the present analyses are restricted to individual classifications to maximize comparability with other methods.

Simulation

Two primary quantities determine clinical significance classifications from these methods: the reliability of the outcome measure and each individual's change from pre- to posttherapy, which is influenced by the average effect size of the pre–post difference. We generated data that systematically altered these two components, with reliabilities ranging from .60 to .95 (in .05 increments, 8 levels total) and pre–post effect sizes (Cohen's *d*) ranging from .10 to 1.00 (in .10 increments, 10 levels total). The HA method also incorporates the pre–post correlation of outcome measures, and thus we explored several correlations (from .25 to .75 in .25 increments, 3 levels total) to ascertain their impact on the HA method. These ranges represent reasonable coverage of psychotherapy outcome data (e.g., Lambert & Ogles, 2004).

Crossing these three factors yielded 240 cells. Within each cell, 500 data sets ($N = 100$ per data set) were generated for a total of 120,000. Pre- and posttherapy *SD*s for the generated data (5.0 and 6.6, respectively) were based on estimates from several recent clinical trials (Christensen, et al., 2004; Jacobson et al., 1996). All simulations and analyses were conducted in R, version 1.9.0 (R Development Core Team, 2004).

Results

Comparison of Methods

In comparing the four methods, we focused on three different statistics: omnibus weighted kappas, pairwise weighted kappas, and raw classification frequencies. Kappa statistics are used to compare the level of agreement in categorical classifications between two raters (or methods, in the present case). Weighted kappas account for the relative distance between categories in assessing disagreements with ordinal data (Agresti, 2002). Omnibus weighted kappas are used to characterize the average agreement across all methods. Pairwise weighted kappas are used to compare agreement for each combination of methods (six total). Kappa values can range from 0 (*no agreement*) to 1 (*perfect agreement*). There are no agreed-on standards for interpreting kappas, though it has been suggested that values of .80 and higher represent "good" agreement (Suen & Ary, 1989). Finally, raw

[3] There is a fifth method based on multilevel modeling that incorporates all available longitudinal data during treatment (see Speer & Greenbaum, 1995). This would require a substantially different data generation process, and the multilevel method does not yield comparable estimates to the pre–post methods when there is curvilinear change (see Footnote 6 in McGlinchey et al., 2002). Thus, we have not included that method here.

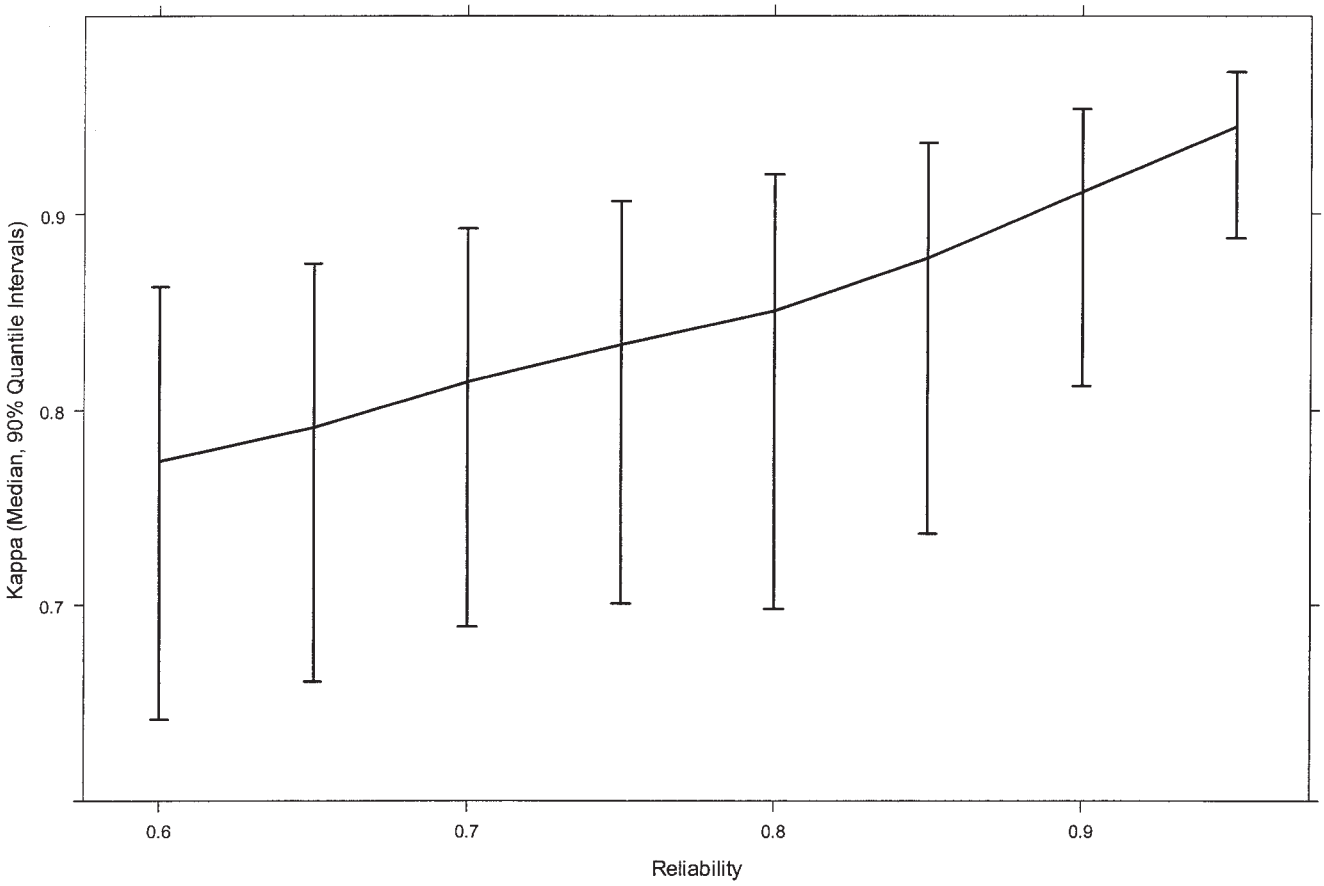

Figure 1. Omnibus agreement among clinical significance methods for various reliabilities.

classification frequencies (deteriorated, unchanged, improved, and recovered) were calculated for each method and data set.

A few details of the HA calculations require further discussion. The HA formulae include the reliability of the difference scores (r_{dd}; Equation 5 in Appendix); r_{dd} is used to weight the individual's reliable change relative to the group's reliable change (see Equation 4). Hageman and Arrindell (1999) recommended that r_{dd} should be limited to values equal to or greater than .40, reflecting "acceptable" r_{dd} reliability. About 18% of data sets yielded values below this level. The r_{dd} values smaller than .40 were strongly related to both low reliability (particularly αs $\leq$ .75) and high pre–post correlations (particularly rs $\geq$ .60). Analyses below are restricted to those data sets that yielded acceptable r_{dd} values ($\geq$.40). Implications of this restriction are described in the Discussion section.

Omnibus Comparisons via Weighted Kappa

As shown in Figure 1, agreement among methods increased as reliability increased. At a reliability of .85, the overall agreement among methods was .88 (90% empirical CI = .73–.93).[4] Agreement among methods was relatively stable and high across effect sizes (*Mdn* = 0.86, 90% empirical CI = .74–.94). There was a slight drop in kappas with increasing pre–post correlations due to changes in the HA method of classification (detailed below in the *Classification Frequencies* section). Kappas ranged from .88 (at r = .25) to .83 (at r = .75).

Pairwise Comparisons via Weighted Kappa

Figure 2 displays pairwise kappas by the reliability of measures. Agreement increased among all methods as reliability increased. Error bars were omitted because they overlapped completely at each level of reliability. Similar to the omnibus comparison, there were few trends in pairwise kappas across the range of effect sizes, with one exception. The agreement between both JT and GLN with the HA method became notably worse at increasing effect sizes (detailed below in the *Classification Frequencies* section).

Classification Frequencies

Figures 3 and 4 present classification frequencies for each patient category for all four methods by reliability and effect size.

[4] All kappa statistics are negatively skewed, and thus all presentations use the median and 90% confidence intervals based on the empirical quantiles (i.e., the end points that encompass 90% of the data).

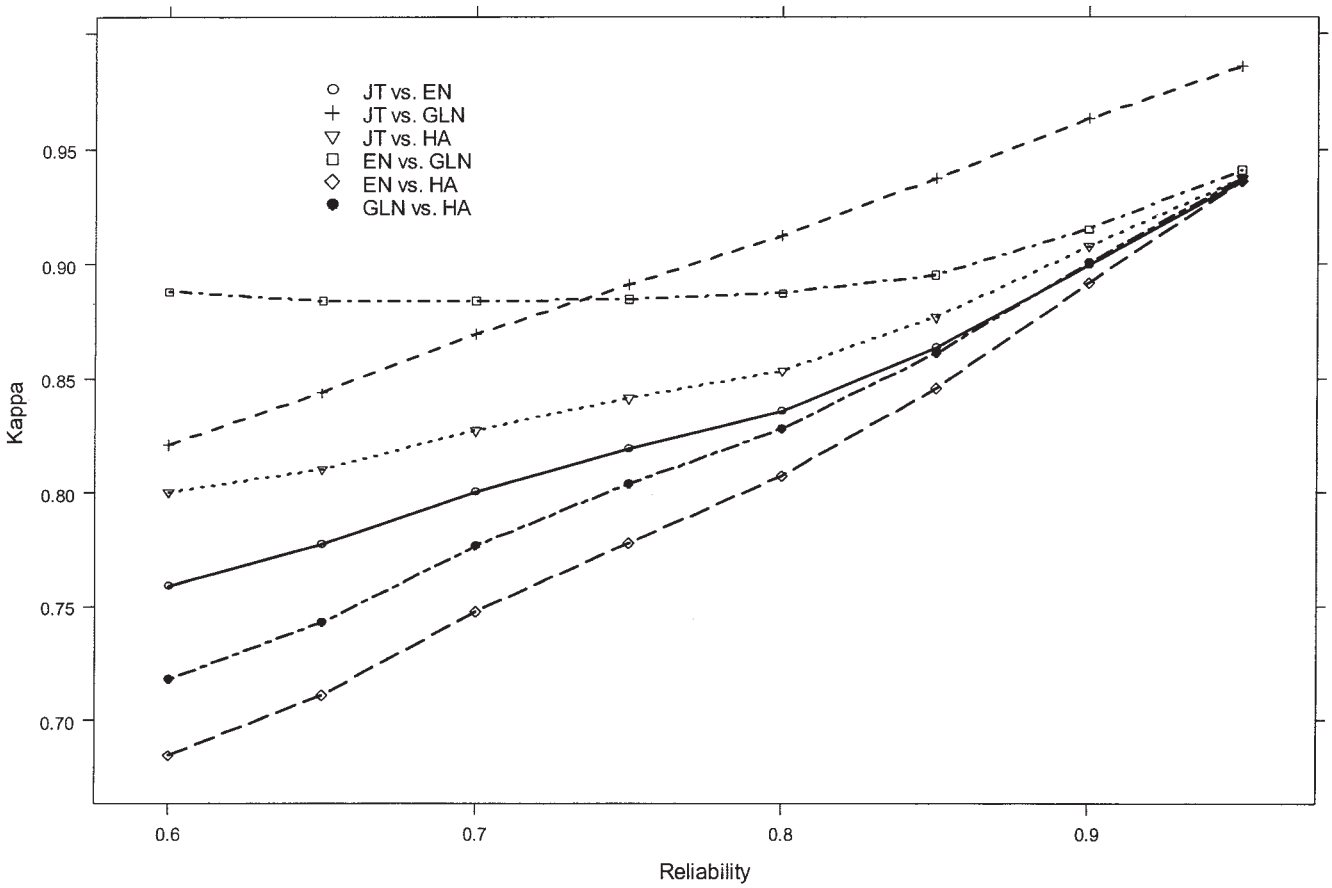

Figure 2. Pairwise agreement among clinical significance methods for various reliabilities. JT = Jacobson-Truax (JT) method; EN = Edwards-Nunnally method; GLN = Gulliksen-Lord-Novick method; HA = Hageman-Arrindell method.

Error bars were again not included because they were completely overlapping. Similarities in categorization are much more striking than discrepancies. The EN method was the most "certain" of the methods in that the greatest number of deteriorated and recovered patients and least number of unchanged patients were routinely classified with this method. Conversely, the HA method was the most conservative with respect to deteriorated and recovered classifications, though greater numbers of patients tended to be classified as improved with this method. In both Figures 3 and 4, categories obtained with the HA method were somewhat different from those obtained with the other methods (albeit to varying degrees). With increasing effect sizes, more patients were classified as improved and fewer were classified as unchanged with the HA method compared with the JT and GLN methods. This explains the decreasing pairwise kappas between these methods reported earlier.

The HA method is the only method that incorporates pre–post correlations. The discrepancy between the HA method and the other methods grew with increasing pre–post correlations, which stems from the strong negative association between the pre–post correlation and r_{dd} ($r = -.58$). As the pre–post correlation increased, r_{dd} decreased and subsequently the individual's change was down-weighted relative to the group change (see Equation 4). The other methods, which do not incorporate the pre–post correlation, were unaffected by this parameter.

Discussion

To our knowledge, the current investigation is the first to systematically vary crucial data parameters via simulation to understand the conditions under which the original and alternative clinical significance methods may differ. Our most important finding is the considerable agreement observed among methods, which was especially great when the reliability of measurement was high. Psychotherapy research routinely uses outcome measures in which αs exceed .90 (Lambert & Ogles, 2004), and there were very few differences between the methods at these reliability levels. Varying effect sizes had less impact on method agreement overall, largely because it affected the classifications obtained with each method in similar ways.

Although similarities exceeded differences, we offer a few comments on the differences that were found. The EN method was the most "certain" in its categorizations; the most deteriorated and recovered cases were classified using this method. This stems

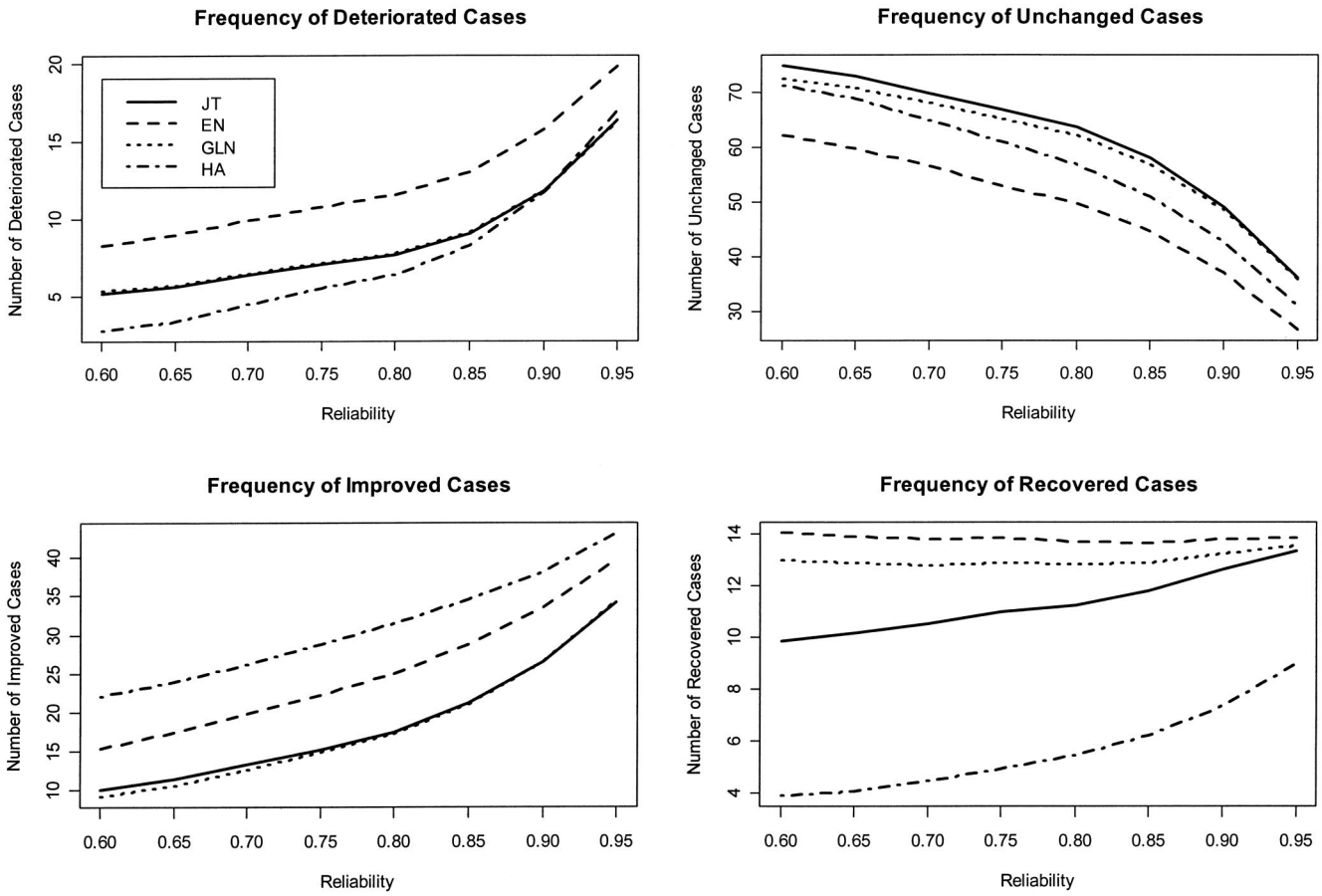

Figure 3. Frequency of classifications among clinical significance methods for various reliabilities. JT = Jacobson-Truax method; EN = Edwards-Nunnally method; GLN = Gulliksen-Lord-Novick method; HA = Hageman-Arrindell method.

directly from the EN method consistently yielding the smallest *SE*s. We found it interesting that although the formulae for the JT and GLN methods are not identical, we obtained virtually identical classifications with them.

Finally, the HA method is both the most complicated and conservative of the methods. With this method, the fewest recovered cases consistently were classified. Moreover, a number of data sets yielded unacceptable r_{dd} values, largely driven by lower reliabilities (αs $\leq$ .75) and higher pre–post correlations (rs $\geq$ .60). These data conditions are neither common nor unheard of in clinical research; thus, it is unclear to what extent these would affect the utility of the HA method.

Recently, Maassen (2000) provided a pointed, statistical critique of the alternative methods considered here. He noted that the newer methods require generally unknown population information to make more precise estimates; in addition, he demonstrated that the JT method provides a large sample approximation of the newer methods, which demand population information. He concluded that "in our view there are strong arguments for preferring the [JT] approach. . . [which] has been undeservedly regarded [as] inferior" (p. 631). Our simulation results would concur with Maassen's critique that in the absence of population-based information, the methods performed similarly, and thus, no one method can be preferred over any other for statistical reasons (i.e., their classifications are virtually indistinguishable). Because of its wide use and ease of computation, the JT method may be preferred by researchers, and if population information is known, alternative methods could certainly be warranted.

What about the future study of clinical significance? The current simulation did not vary every possible data parameter (e.g., pretreatment *SD*) and "real-world" data do not always conform to simulated data; thus, further comparisons among methods might focus on classification or accuracy via an external criterion. However, we are not optimistic that future comparisons would be very elucidating, especially because differential *accuracy* of methods requires differential *classification*, which by and large was not shown in the present study. At the same time, the broader issue of the external validity of clinical significance methods in general is quite important, and further research exploring the relationship of clinical significance to other markers of client improvement would be worthwhile.

As we noted earlier, the methods considered here are only some of the approaches to measuring clinical significance. Sheldrick, Kendall, and Heimberg (2001) demonstrated the utility of a multimethod approach in their study examining the effectiveness of

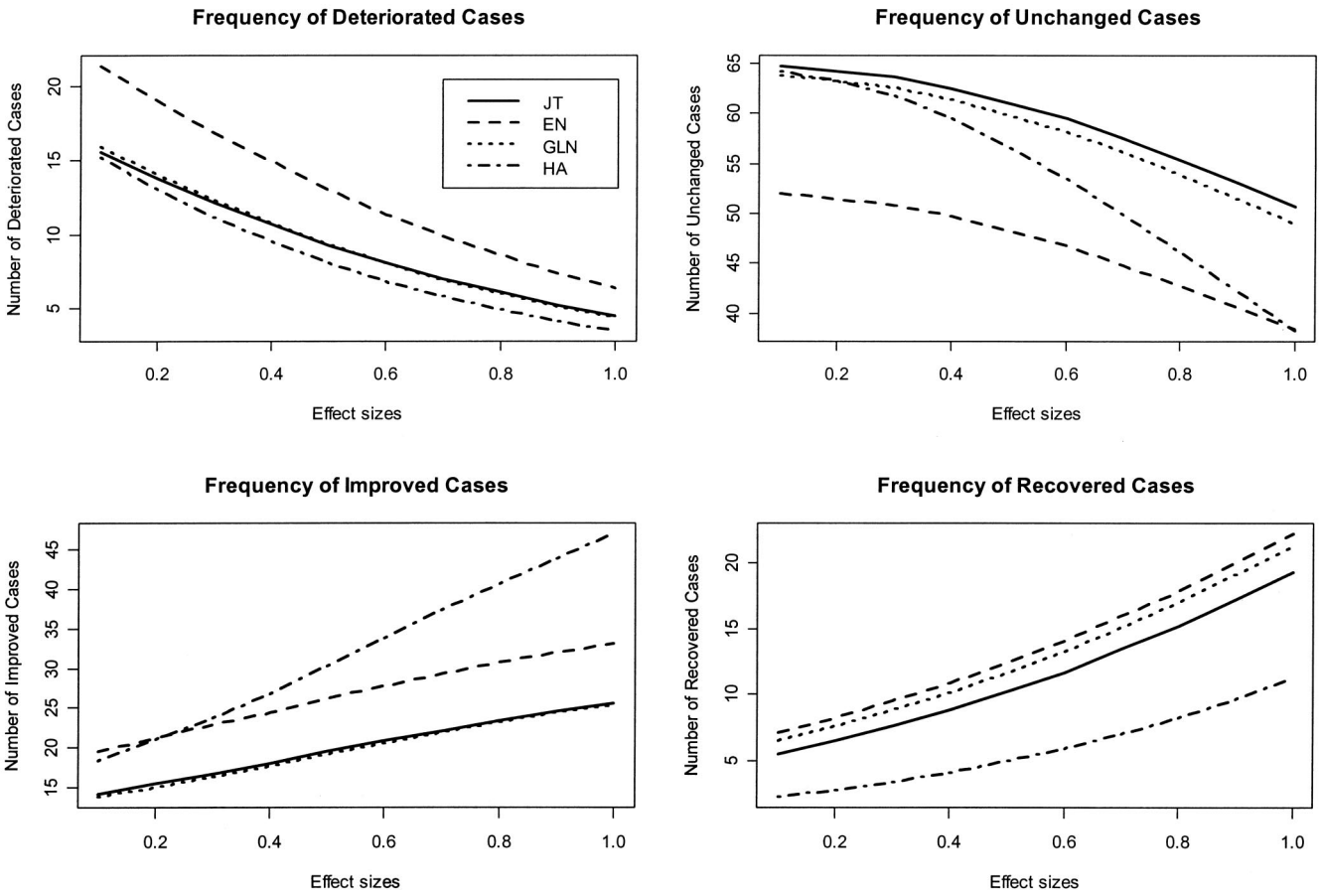

Figure 4. Frequency of classifications among clinical significance methods for various effect sizes. JT = Jacobson-Truax method; EN = Edwards-Nunnally method; GLN = Gulliksen-Lord-Novick method; HA = Hageman-Arrindell method.

three treatments for children diagnosed as having conduct disorders. In this study, a joint RCI and equivalence testing method was used, with results supporting equivalence testing as the only method capable of discriminating among the three treatments. Beutler and Moleiro (2001), cited this finding as supportive of the complementary nature of these methods, in which RCI acts as a measurement of the amount of change and equivalence testing acts as a measure of the clinical meaning of change.

Future research may consider the strengths and weaknesses of various approaches to clinical significance and whether there is a certain package of methods (e.g., Jacobson et al.'s method plus normative comparisons or quality of life measures) that yields the greatest clinical utility. Moreover, the issue of measure correspondence can complicate the interpretation of clinical significance and has yet to be addressed; studies often use multiple measures, which may yield differing results in terms of clinical significance (Ogles, Lambert, & Sawyer, 1995). In addition, there needs to be further conceptual clarity about what is precisely meant by clinical significance vis-à-vis the impact of therapy on clients' lives. Thus, a profitable area of future research may focus on what clients believe are the most important effects of therapy on their lives.

References

Agresti, A. (2002). *Categorical data analysis* (2nd ed.) . New York: Wiley.

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

Beutler, L. E., & Moleiro, C. (2001). Clinical versus reliable and significant change. *Clinical Psychology: Science and Practice, 8,* 441–445.

Christensen, A., Atkins, D. C., Berns, S. B., Wheeler, J., Baucom, D. H., & Simpson, L. (2004). Integrative versus traditional behavioral couple therapy for moderately and severely distressed couples. *Journal of Consulting and Clinical Psychology, 72,* 176–191.

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.

Gladis, M. M., Gosch, E. A., Dishuk, N. M., & Crits-Christoph, P. (1999). Quality of life: Expanding the scope of clinical significance. *Journal of Consulting and Clinical Significance, 67,* 320–331.

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

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

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., Dobson, K. S., Truax, P. A., Addis, M. E., Koerner, K., Gollan, J. K., et al. (1996). A component analysis of cognitive-behavioral treatment for depression. *Journal of Consulting and Clinical Psychology, 64,* 295–304.

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

Lambert, M. J., & Ogles, B. M. (2004). The efficacy and effectiveness of psychotherapy. In M. J. Lambert (Ed.), *Bergin and Garfield's Handbook of Psychotherapy and Behavior Change* (pp. 139–193). New York: Wiley.

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

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.

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.

R Development Core Team (2004). *R: A language and environment for statistical computing.* Vienna: R Foundation for Statistical Computing.

Sheldrick, R. C., Kendall, P. C., & Heimberg, R. G. (2001). The clinical significance of treatments: A comparison of three treatments for conduct disordered children. *Clinical Psychology: Science and Practice, 8,* 418–429.

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.

Suen, H. K., & Ary, D. (1989). *Analyzing quantitative behavioral observation data.* Hillsdale, NJ: Erlbaum.

Appendix

Computational Formula: Reliable Change Indices of Clinical Significance Methods

Jacobson & Truax (JT) Method (1991)

$$\frac{(X_{post} - X_{pre})}{\sqrt{2[S_{pre}\sqrt{1 - r_{xx}}]^2}} \tag{1}$$

Note. X_{post} = individual's raw score at posttreatment; X_{pre} = individual's raw score at pretreatment; S_{pre} = standard deviation of sample at pretreatment; r_{xx} = reliability of measure.

Gulliksen-Lord-Novick (GLN) Method (Hsu, 1989, 1999)

$$\frac{[X_{post} - M_{pop}] - r_{xx}[X_{pre} - M_{pop}]}{S_{pop}\sqrt{(1 - r_{xx}^2)}} \tag{2}$$

Note. X_{post} = individual's raw score at posttreatment; X_{pre} = individual's raw score at pretreatment; M_{pop} = hypothesized population mean toward which scores would be regressing; S_{pop} = standard deviation of population toward which scores would be regressing. The pretreatment *M* and *SD* were used for M_{pop} and S_{pop}; r_{xx} = reliability of measure.

Edwards-Nunnally (EN) Method (Speer, 1992)

$$[r_{xx}(X_{pre} - M_{pre}) + M_{pre}] \pm 2S_{pre}\sqrt{1 - r_{xx}} \tag{3}$$

Note. r_{xx} = reliability of measure; X_{pre} = individual's raw score at pretreatment; M_{pre} = mean of sample at pretreatment; S_{pre} = standard deviation of sample at pretreatment.

Hageman & Arrindell (HA) Method (1999)

A. Reliability of Change (RC_{INDIV}) Index

$$\frac{(X_{post} - X_{pre})r_{dd} + (M_{post} - M_{pre})(1 - r_{dd})}{(\sqrt{r_{dd}})(\sqrt{2S_E^2})} \tag{4}$$

Note. X_{post} = individual's raw score at posttreatment, X_{pre} = individual's raw score at pretreatment, r_{dd} = reliability of difference scores, M_{post} = mean of sample at posttreatment, M_{pre} = mean of sample at pretreatment, S_E = standard error of the estimate.

Determining the answer to Equation 4 requires additional computations. The reliability of difference scores (r_{dd}) is obtained by the following formula:

$$r_{dd} = \frac{s_{pre}^2 r_{xx(pre)} + s_{post}^2 r_{xx(post)} - 2s_{pre}s_{post}r_{pre*post}}{s_{pre}^2 + s_{post}^2 - 2s_{pre}s_{post}r_{pre*post}} \tag{5}$$

The reliabilities for both pretreatment scores, $r_{xx(pre)}$, and posttreatment scores, $r_{xx(post)}$, are determined by the following formulae:

$$r_{xx(pre)} = \frac{(S_{pre}^2 - S_E^2)}{S_{pre}^2} \tag{6}$$

$$r_{xx(post)} = \frac{(S_{post}^2 - S_E^2)}{S_{post}^2} \tag{7}$$

The standard error of the estimate (S_E) is obtained by the following formula:

$$S_E = S_{sample}\sqrt{1 - r_{sample}} \quad (8)$$

Note. S_{sample} = pretherapy SD; r_{sample} = reliability of measure.

B. Alternative cutoff: Clinical Significance of Change (CS_{INDIV}) Index

$$\frac{M_{post} + (X_{post} - M_{post})r_{xx(post)} - TRC}{(\sqrt{r_{xx(post)}})S_E} \quad (9)$$

Note. X_{post} = individual's raw score at pretreatment, $r_{xx(post)}$ = reliability of posttreatment scores, M_{post} = mean of sample at posttreatment, S_E = standard error of the estimate, TRC = true cutoff score.

Using Cutoff A criterion, the value for *TRC* would be computed as follows:

$$A_{true} = M_{pre} - 2S_{pre}\sqrt{r_{xx(pre)}} \quad (10)$$

Note. S_{pre} = standard deviation of sample at pretreatment, $r_{xx(pre)}$ = reliability of pretherapy scores, M_{pre} = mean of sample at pretreatment.

Received September 10, 2004
Revision received February 4, 2005
Accepted February 8, 2005 ■