Psychotherapy Research
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/tpsr20

Using clinical significance in psychotherapy outcome research: The need for a common procedure and validity data

Michael J. Lambert [a b] & Benjamin M. Ogles [a b]
[a] Department of Psychology , Brigham Young University , Provo, Utah
[b] Department of Psychology , Ohio University , Athens, Ohio, USA
Published online: 22 Sep 2009.

To cite this article: Michael J. Lambert & Benjamin M. Ogles (2009) Using clinical significance in psychotherapy outcome research: The need for a common procedure and validity data, Psychotherapy Research, 19:4-5, 493-501, DOI: 10.1080/10503300902849483

To link to this article: http://dx.doi.org/10.1080/10503300902849483

Psychotherapy Research, July–September 2009; 19(4–5): 493–501

PSYCHOTHERAPY RESEARCH METHODS

Using clinical significance in psychotherapy outcome research: The need for a common procedure and validity data

MICHAEL J. LAMBERT & BENJAMIN M. OGLES

Department of Psychology, Brigham Young University, Provo, Utah & Department of Psychology, Ohio University, Athens, Ohio, USA

(Received 19 August 2008; revised 24 February 2009; accepted 25 February 2009)

Abstract
It is recommended that an estimate of clinical significance be included in all psychotherapy outcome studies and that this estimate be based on the work of Jacobson and Truax (1991). The concept of clinical significance is defined and put in the context of broadly accepted statistical methods along with its advantages and a rationale for using the Jacobson methods. One implication of this recommendation is that the use of the term will have a standard meaning. Examples of loss of meaning are provided and suggest that conclusions about best practices will be affected unless such a voluntary step is taken. Some problems with the concept of clinical significance are noted and a call for validity studies is made.

Keywords: clinical significance; statistical methodology; outcome research; mental health services research

In general, three groups of statistics have been used to express the consequences of psychotherapy for patients: statistical significance of within- and between-group differences, effect size (ES), and clinical significance. The data produced by research projects designed to evaluate the efficacy or effectiveness of therapy is typically submitted to statistical tests of significance. An outcome achieves *statistical* significance if the magnitude of the mean difference is beyond what could have resulted by chance alone. Statistical analyses and statistical significance are essential for therapy evaluation because they inform us that the degree of difference was unlikely due to chance. Sole reliance on statistical significance, however, can lead to perceiving differences (i.e., treatment gains) as potent when, in fact, they may be clinically unremarkable.

ES statistics provide additional information by providing an index of the degree or magnitude of the average change. These statistics have proven especially valuable for meta-analytic reviews of studies by allowing researchers to sum and average the size of treatment effects across different measures and studies, thereby estimating the percentage of patients who benefit from treatments (assuming normal distribution of scores in the samples studied). A small effect indicates a smaller difference in group means than a large effect. Even ES indices, however, do not give us information regarding within-group variation or the clinical relevance of group or individual change.

Although statistical significance tests and ES indices provide important information regarding mean differences, they do not give information regarding the variety of responses to treatment within the treated group. Some individuals who received treatment may have substantial improvement, whereas others show no change or even deteriorate. The range or variability of those individual responses is accounted for in the statistical test, but individual change is not adequately described or considered as a method of evaluating effective treatment.

The strength of clinical significance methodology is that it considers change on the individual patient level. This provides important additional information that extends beyond the statistical test or ES estimate. In addition, clinical significance estimates

Correspondence concerning this article should be addressed to Michael J. Lambert, Department of Psychology, Brigham Young University, 272 TLRB, Provo, UT 84602, USA. E-mail:michael_lambert@byu.edu

ISSN 1050-3307 print/ISSN 1468-4381 online
DOI: 10.1080/10503300902849483

are useful for small-sample studies in which power to detect statistical differences may be lacking and individual changes may be masked by group variance. Such information can be used to help manage the treatment of ongoing cases as well as gauge the impact of care on groups of individuals.

Brief History

As early as the 1970s (Bergin, 1971; Kazdin, 1977; Lick, 1973), the examination of individual changes occurring during formal psychotherapy research projects was seen as important and necessary to supplement statistical significance. Several methods were developed such as goal attainment scaling, in which individualized goals were rated at the end of treatment (Kiresuk, Smith, & Cardillo, 1994). However, psychotherapy studies that included clinical significance were relatively rare outside of behavior therapy, where tracking individual change has been a hallmark activity.

From these beginnings, two rather independent bodies of literature evolved for designating the clinical significance of psychotherapy: social validity and clinical significance. Social validity (Kazdin, 1977; Wolf, 1978) emerged as a method of assessing the perspective of individuals outside the therapeutic relationship regarding the importance of psychosocial interventions and outcomes. Social validity provides a cohesive rationale and two specific methodologies (subjective evaluation and social comparison) for evaluating the relevance of client change. Subjective evaluation refers to gathering data about clients by individuals who are "likely to have contact with the client or are in a position of expertise" (Kazdin, 1998, p. 387). This allows the researcher to understand whether the client has made qualitative changes that are, in turn, observable by others. The underlying premise is that socially valid changes resulting from the intervention in question will result in the client's postintervention behavior being indistinguishable from a normal reference group. Although social validity emphasizes an examination of practical change from the perspective of societal members, clinical significance takes a slightly narrower view of meaningful change by identifying methods defined by clinician-researchers (Ogles, Lunnen, & Bonesteel, 2001).

Early in the 1980s, Jacobson, Follette, and Revenstorf (1984) presented a statistical solution for systematically estimating the clinical significance of change based on self-report scales completed by distressed couples undergoing couple therapy by reanalyzing data from a clinical trial and estimating the percentage who attained "normal" functioning. Since this work was published, there has been a steady refinement of methods and statistical formulas as well as increased applications in clinical trials. Importantly, these methods form a core condition for the emerging line of psychotherapy research aimed at monitoring and improving individual patient treatment response: patient-focused research (Lambert, 2001; Lambert et al., 2003).

The two most prominent requirements of clinically significant change are that (a) treated clients make statistically reliable improvements as a result of treatment (Jacobson, Roberts, Berns, & McGlinchey, 1999) and (b) treated clients are empirically indistinguishable from "normal" or nondeviant peers after treatment (Kendall, Marrs-Garcia, Nath, & Sheldrick, 1999). Several statistical methods can be used to evaluate these propositions. Before turning to these, it is important to note that the use of such methods broadens and modifies our views of the effects of psychotherapy. In general, the consequence of using clinical significance methods softens our claims for the effects of psychotherapy and makes it clear that a portion of patients who undergo treatment do not respond to the degree that might be hoped for and that a small group of patients (perhaps 5–10%) actually worsen (Lambert & Ogles, 2004).

For example, Hansen, Lambert, and Forman (2002) analyzed 28 clinical trials, including 2,109 patients and 89 treatment conditions that studied a variety of disorders and treatment methods, and found that, on average, 58% of patients met the criteria for clinically significant change following an average of 13 treatment sessions. In contrast, recovery rates in routine care (averaging about three sessions) resulted in about 6.5%, with deterioration close to 8%.

Examining a specific study can also be informative. Ogles, Lambert, and Sawyer (1995) examined the clinical significance of the Treatment of Depression Collaborative Research Program (TDCRP) data. When combining estimates across three empirically supported treatments (interpersonal psychotherapy, cognitive–behavioral therapy, and imipramine plus clinical management) and three measures, they found that 69 (55%) of the 125 clients who participated in a minimum of 12 sessions and 15 weeks of treatment met the criteria for clinically meaningful change on all three measures, with 0 to 5% deteriorating. We might summarize this clinical significance data by saying that if we give our "best" treatments, in high doses (relative to routine care), a patient has better than a 50% chance of reliably changing and reentering the ranks of normal functioning. This kind of information provides a meaningful supplement to inferential and ES statistics.

The Jacobson–Truax Method

As mentioned, Jacobson et al. (1984) introduced a first statistical method to assess clinically significant change. The variety of statistical calculation methods now used suggests a two-step criterion for clinically significant change based on this original proposal and its minor modifications formulated by Jacobson and Truax (1991). The first step is to determine whether the observed change from pretest to posttest is statistically reliable (i.e., whether the observed difference in scores can be attributed to real change rather than the measurement error of the outcome instrument). This reliable change index (RCI) is calculated by dividing the difference between the observed posttest (x_{post}) and pretest (x_{pre}) scores by the standard error of differences (Jacobson & Truax, 1991):

$$\text{RCI} = \frac{(x_{post} - x_{pre})}{\sqrt{2S_E^2}}$$

The size of the standard error of measurement (S_E) depends on the reliability (r) of the outcome measure:

$$S_E = SD\sqrt{1 - r}$$

This means that the more reliable the instrument, the smaller the resulting standard error and thus the smaller the observed change between pre- and posttest scores that is required to achieve a statistically reliable change. In the discussion of which reliability score is most accurate for the calculation of the RCI, Martinovich, Saunders, and Howard (1996) as well as Tingey, Lambert, Burlingame, and Hansen (1996) recommended the use of internal consistency rather than of test–retest reliability scores. If the RCI is smaller than –1.96, a person scored reliably lower on the posttest compared with the pretest assessment ($p < .05$) and thus has shown reliable improvement on the outcome measure. Similarly, a score beyond 1.96 in the opposite direction would indicate that the individual reliably deteriorated. An important advantage of such calculations is that they can be standardized on large samples for commonly used measures, thereby producing a standard cutoff score used by all researchers who utilize the measure rather than a variable score that is sample and study specific.

The second step in the calculation of clinical significance is the estimation of a cutoff point between a patient (dysfunctional) and a nonpatient (functional) population. Jacobson et al. (1984) proposed "that a change in therapy is clinically significant when the client moves from the dysfunctional to the functional range during the course of therapy" (p. 340). They suggested three different cutoff scores that could be calculated depending on the available patient or nonpatient data sets and the distribution characteristics (Jacobson et al., 1999).

The use of their Cutoff C is recommended when adequate normative data sets (patient and nonpatient) are available and there is overlap between the two distributions.

Cutoff C is a weighted midpoint between the means of a patient and a nonpatient population:

$$\text{Cutoff } C = \frac{(SD_{patient}M_{nonpatient}) + (SD_{nonpatient}M_{patient})}{(SD_{patient} + SD_{nonpatient})}$$

Using these two steps, each individual can be classified as recovered (passed both criteria), improved (passed only the RCI criterion in the positive direction), unchanged (did not pass the RCI criterion), or deteriorated (passed the RCI criterion in the negative direction). To date, the Jacobson–Truax (JT) method has been the most frequently reported method for assessing clinically meaningful change in psychotherapy outcome studies (Ogles et al., 2001). Several authors have criticized the JT method on statistical grounds and suggested alternative methods that they believed would yield more accurate estimates of meaningful change than the original conceptualization. Six alternative methods and their calculations have been summarized by Lambert, Hansen, and Bauer (2008) for the calculation of the RCI that have been used in comparative studies investigating convergences and divergences in classification rates of different methods. The main point of criticism addressed by newer approaches was that the JT approach did not take into account regression to the mean. Regression to the mean implies that in repeated assessments with the same (not perfectly reliable) outcome measure, more extreme scores naturally become less extreme over time.

Criticizing the JT method for not accounting for this phenomenon, Hsu (1989, 1999) suggested the Gulliksen–Lord–Novick (GLN) method. The GLN formula for calculating the RCI includes estimates of the population mean and standard deviation toward which scores are assumed to regress. It is problematic, however, because these means and standard deviations are rarely known (Maassen, 2000). Other authors have suggested alternative solutions for dealing with the problem of regression. The Edwards–Nunnally (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 the formulation of confidence intervals for calculating pre- 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, similar to 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. Furthermore, the ease of presentation offered by confidence intervals is an additional benefit of this method.

The Hageman and Arrindell (HA; 1999a) method, drawing on Cronbach and Gleser's (1959) use of the phi coefficient as a measure of discrimination, 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 pre- 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}.

Speer (1992) advocated a multiwave data approach using growth curve modeling (e.g., hierarchical linear modeling [HLM]; Bryk & Raudenbush, 1992). One of the advantages of a multiwave approach is that it uses more than two data points per individual and, by doing so, reflects the change that occurs between pre- 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 and the capability of handle missing data.

Diversity in Methods and Their Consequences

Although addressing methodological and statistical controversies on how to calculate clinical significance most accurately is important, of more pressing concern to many psychotherapy researchers is the question of what are the practical consequences of using one method instead of another. Three empirical studies have addressed this question and investigated the convergences and divergences in classifications between different methods. Speer and Greenbaum (1995) compared the classification rates of 73 patients, with the results showing relatively high rates of agreement (78–81%) between four of the methods. Speer and Greenbaum (1995) recommend the use of HLM when longitudinal data exist and identifying deteriorated patients is not of interest (because HLM classified none of the 73 patients in their sample as reliably deteriorated). Otherwise, they recommend the JT method.

McGlinchey, Atkins, and Jacobson (2002) compared five methods of estimating clinical significance rates in a sample of 128 patients with depression. Three of the methods were identical to those used by Speer and Greenbaum (1995), but there was notable variability.

Bauer, Lambert, and Nielsen (2004) compared the classification rates of the same five methods as McGlinchey et al. (2002) in a sample of 386 patients who were treated in routine care at a university-based outpatient clinic. The average agreement of one method with the other four ranged from 71 to 85%. Instead of using empirical data, Atkins, Bedics, McGlinchey, and Beauchaine (2005) conducted a simulation study to systematically explore the performance of four methods by varying several relevant parameters (ESs, reliabilities, pre–post correlations). This allowed them to evaluate not only whether the methods differ but also under which conditions they differ. Overall, the results showed considerable agreement among methods, especially in the case of high reliability in the outcome measure.

In contrast to comparing classification rates of different methods, several studies have investigated the extent to which the use of different outcome measures, and different perspectives (e.g., therapist, client, and spouse ratings), produces comparable estimations of clinically significant change. For example, one study evaluated comparable estimations of multiple measures of outcome using the TDCRP data (Ogles et al., 1996). When comparing the number of individuals who could be classified as clinically significant changers using the JT method and three different measures (Beck Depression Inventory [BDI], Hamilton Depression Rating Scale, and Hopkins Symptom Checklist), a sizable degree of correspondence existed among the measures (75% of the clients were classified by all three measures consensually). Still, 25% of the clients made reliable improvement or deterioration on one measure but not on others. Later studies indicate that different measures produce varying rates of clinically significant change (e.g., Beckstead et al., 2003). Consequently, one must keep in mind that the rates of clinically significant change are dependent on both the specific outcome measures that are used and the statistical methods used to estimate clinically significant change.

It is important to differentiate between the suitable ways to calculate rates of clinically significant change and the practical relevance of differences between methods for psychotherapy research. The results of the three comparative studies mentioned

previously indicate that convergences and divergences between methods vary from study to study. The resulting confusion that arises when each researcher defines clinically significant change in his or her own way quickly becomes apparent by perusing the results and methods sections of outcome studies. For example, in the 2008 special issue of *Psychotherapy Theory, Research Training and Practice* devoted to "New Treatments in Psychotherapy," very little consistency was found for operational definitions of clinical significance (if any occurred at all). The first article in the series of studies (Constantino et al., 2008) reported on an integrative cognitive therapy (ICT) for depression compared with standard cognitive therapy (CT). These authors used the BDI and claimed to apply the JT criteria. However, instead of basing the JT cutoff scores on normative data, they applied the more informal suggestion of Beck, Ward, Mendelson, Mock, and Erbaugh (1961): 0–9, nondistressed; 10–15, minimally distressed; 16–19, moderately distressed; 20–29, moderately to severely distressed; 30–63, severely distressed. In this breakdown, a score of 15 or less indicated return to normal functioning. They used the RCI as recommended by Jacobson and Truax but did not specify the calculated cut score, which was apparently sample specific. Given this circumstance, the 82% of patients who were judged to be recovered in the ICT treatment surpassed the recovery rate in CT (55% recovery).

As this single study illustrates, the JT method can be implemented in different ways. In their review of 74 studies published in the *Journal of Consulting and Clinical Psychology*, Ogles et al. (2001) found that "there was considerable variation" even among studies that purported to use the JT methodology. They went on to indicate that it was sometimes difficult to ascertain how the JT method was implemented, and in other instances unique variations on the method were tailored to the study. Clearly, the variability in methods makes it difficult to compare treatment effects across studies.

Other Limitations to Clinical Significance

Despite the potential advantages to using clinical significance methods in reporting outcome, there are numerous limitations to these methods. In addition to the problems associated with the diversity of statistical procedures without an agreed-upon method that were mentioned previously, there are several other limitations to the clinical significance methods, including (a) the lack of suitable norms for many potentially useful outcome measures and for many relevant clinical populations; (b) difficulty making applications work with special populations, such as the well and chronically ill; (c) the fact that clinical significance relies on categorizations that are unreliable at the category edges; and (d) the limited evidence for the validity of clinical significance methods.

Lack of normative data and changing parameters. In addition to variability in methods, studies also use or cannot find appropriate parameters to use for the calculation of clinical significance. Many studies do not report the means, standard deviations, or reliability parameters that were used to calculate the RCI or clinical cutoff (Ogles et al., 2001). The most common difference among published studies is that some researchers use sample-specific parameters, whereas others use normative samples. This depends mostly on the measure being used: Some measures have a rich foundation of normative data in multiple types of studies with various populations, whereas others have limited or nonexistent normative data. Lack of normative data will continue to force researchers to use sample-specific data unless continuing efforts produce expanded data bases for their use. Regardless of the reasons for this discrepancy, the variation in parameters produces heterogeneous results across studies and samples.

Patients entering treatment in the functional range and chronic illness. Another limitation of the clinical significance methods is their inability to adequately categorize change in patients entering treatment already in the functional range. This fact is captured in the interchange between Gray (2003) and Hansen, Lambert, and Forman (2003). Depending on the clinical setting, a large number of patients may initiate treatment with low to mild levels of psychiatric symptoms and have little room for improvement. If these patients score at or below the established cutoff between the functional and dysfunctional normative distributions, it will be impossible for them to meet the criteria for recovery using clinical significance methodology.

A related problem involves the assessment of individuals with chronic conditions. In these instances, it may be improbable or even impossible for the individual to return to the functional distribution. For example, one of the authors recently treated an individual who had treatable mental health issues that were concurrent with a chronic medical condition. Although the individual improved in treatment, the symptom-based outcome measure included many items that were unlikely to change as a result of the lingering symptoms of the medical condition (e.g., fatigue, low energy, aches and pains, tingling). Similarly, individuals with a chronic disorder such as diabetes or bipolar disorder may not be expected to have exactly the same level of functioning as they had before the onset of the

disorder and on outcome measures may not return to the range of the functional group. It seems evident that these individuals can make clinically meaningful change; however, the current methods are not well suited to assess that change.

Tingey et al. (1996) attempted to partially address these issues when they recommended viewing patient functioning as a continuum rather than using a dichotomous functional/dysfunctional operationalization of patient functioning. Based on a continuum, appropriate cutoffs can be created between any two relevant populations, assuming they form distinct distributions by meeting specified criteria for nonoverlap. Tingey et al. (1996) demonstrated this procedure using the Symptom Checklist 90- Revised (SCL-90-R; Derogatis, 1983) by defining cutoffs between severe and moderate symptom ranges, moderate and mild symptom ranges, and mild and asymptomatic symptom ranges. Thus, for patients who enter treatment in a mildly distressed range, an appropriate asymptomatic range of functioning can be defined and relevant normative data collected to establish a cutoff point and an RCI. As demonstrated by Tingey et al. (1996) and others (Grundy, Lambert, & Grundy, 1996; Seggar, Lambert, & Hansen, 2002), identifying, recruiting, and assessing asymptomatic participants from the community for making comparisons with individuals who come to treatment only mildly symptomatic is a labor- and resource-intensive process.

Similarly, one with a chronic illness might move from a severe to a moderate range and thus meet the definition of clinically meaningful change. This, too, requires significant amounts of normative data in order to identify the appropriate categories of asymptomatic, mild, moderate, severe, and so on, for each outcome measure.

Unreliable categorizations on the margins. Determining whether a client moves past a cutoff point raises concerns about the nature of assessing change, the precision of the cutoffs, and the accuracy of estimates that fall around the margins of the categories. For example, if a client must have a score on the BDI that is 12 or below to be considered part of the functional distribution, what of a client who moves from a score of 45 at pretreatment to a score of 13? The fact that the individual made such a large change raises interesting issues about our definitions, but even more relevant for the current discussion, how precise is the cutoff in this circumstance? Should not the cutoff between two distributions have a range of scores that acknowledges the potential error inherent in determining the cutoff in the first place (e.g., scores in the range of 10–14 cannot be discriminated from one another and so all are considered equally likely to be below 12)? The problems with reliable categorizations for individuals on the cusp also present a problem for the JT definitions. When has one made a change that moves him or her back in to the functional range? Precision in measurement requires high-quality normative and psychometric data that, in many instances, is not available. The next generation of clinical significance research will benefit from heightened attention to these issues.

Validity. In some ways it is ironic that a statistical method was developed to examine the clinical significance of treatment effects for individuals. Indeed, one might suppose that a method for determining the clinical significance of treatment for any given client should be inherently idiographic. In fact, there are studies that have examined the clinical significance of treatment outcomes using more individualized methods. For example, Åsenlöf, Denison, and Lindberg (2006) developed individualized criteria for assessing clinically meaningful change that were based in the JT method and incorporated each patient's goal priorities. For the vast majority of studies, however, the algorithms for determining clinical significance are rooted in group-level statistics. When using these group-based statistical methods to identify individual changes, the validity of the measurement categories becomes crucial. For example, the SCL-90-R may be an adequate measure of the number and intensity of symptoms, yet a set level of decrease in reported symptoms may or may not correspond to actual behavioral or functional improvements. Having a change score that is reliable or a posttreatment score that falls within the range of the functional distribution is evidence of meaningful improvement. Yet the validity of reliable change or movement beyond a cutoff as criteria for meaningful change has not been substantiated.

Some studies have attempted to verify that reliable improvements have external validity. Ankuta and Abeles (1993) provided the first bits of evidence that reliable changes were valid. They compared the satisfaction of clients who did or did not demonstrate clinically significant improvement. They operationalized "satisfaction" as the extent of self-reported change resulting from therapy and found that clients who made clinically significant improvement reported greater satisfaction than those who did not make clinically significant change.

Lunnen and Ogles (1998) conducted a multiperspective, multivariable validity analysis of the RCI component of the JT methodology. Clients were separated into three groups based on their change scores on the Outcome Questionnaire (OQ-45.1; Lambert, Lunnen, Umphress, Hansen, & Burlingame, 1994): improvers, no-changers, and

deteriorators. Improvers had greater amounts of change and better therapeutic alliances (from both therapist and client points of view) than no-changers and deteriorators. The groups did not differ in terms of satisfaction with services. Clients demonstrating reliable deterioration were not significantly different from nonchangers on any of the outcome variables reported by any of the perspectives. From this study, one might conclude that the JT method is a valid method for assessing improvement, but deterioration was not substantiated through a retrospective approach. Another interpretation, however, might be that the JT method identified true deteriorators, but the therapists and clients using retrospective ratings were less sensitive to changes identified through the prospective session by session ratings of outcome. Schulte (2008), in studying the constructs of client-perceived treatment suitability and expectancy of a positive outcome as they related to psychotherapy, noted the tendency of expectancy ratings to correlate with final status (posttest functioning) and retrospective evaluations of treatment rather than amount of improvement, thus suggesting the importance of examining both outcome variables (reliable change and entering the ranks of normal functioning).

Several relatively recent articles have continued the investigation of the JT constructs as a secondary component of studies designed to assess the meaningfulness of client changes in treatment (e.g., Åsenlöf et al., 2006; Newnham, Harwood, & Page, 2007; Openshaw, Waller, & Sperlinger, 2004). In each case, researchers investigated the concurrent or convergent validity of the outcomes classified as clinically significant on the main measure of outcome using the JT method with assessment of outcome on other important areas of change. For example, Newnham et al. (2007) found that clients classified as recovered, improved, no change, or deteriorated using the Short Form-36 (SF-36) had corresponding levels of self-rated quality of life and clinician-rated overall distress. This suggests that the Jacobson categories are valid indicators of real differences in behavior on other important areas of functioning from multiple perspectives.

This collection of studies begins the process of demonstrating that the JT method may be a valid method for identifying meaningful change, but clearly more research in this area is needed to verify that these statistical procedures are consistent with other indices of meaningful functional or behavioral improvement.

Recommendations

In the face of these findings and the growing tendency of authors of specific psychotherapy outcome studies to use an even wider variety of methods and procedures, it is recommended that the JT method in the form presented here be used in all outcome research when possible. Because of uncertainty about the best methods and limitations inherent any of the methods, other methods can also be used in addition to the JT methods based on researcher needs and preferences. However, a single standard would make interpretation of the results of a wide variety of studies much simpler and more straightforward; otherwise, there will be more confusion about the classification of recovered and deteriorated patients.

Several points lead to this recommendation: First and most important, it has yet to be demonstrated that any other approach is superior in terms of more accurate estimations of clinically significant change (McGlinchey et al., 2002). As noted by Hsu (1999), it is inappropriate to recommend a particular method based on its production of higher improvement rates. Because of the lack of validation studies (with the exception of McGlinchey et al., 2002, who did not find differences in the performance of the different methods), it remains an open question as to which method is most sensitive to detecting meaningful changes in patients undergoing psychotherapeutic treatment. Second, as noted by Maassen (2000), the newer methods require generally unknown population information to make more precise estimates than the JT method. Maassen (2000) as well as Atkins et al. (2005) conclude that the JT method has been undeservedly regarded as inferior and also recommend its use. Finally, the JT method is relatively easy to compute and already the most popular approach (Ogles et al., 2001), which allows for comparisons of change rates across studies unless it is modified.

In addition to accepting the JT method as the standard way of determining the clinical significance of the findings, we recommend that standard methods for selecting parameters also be advanced. Our recommendation is that the internal consistency estimate for the measure should be used in calculations of the standard error of the difference score and that normative data should be used to establish cutoffs that are used consistently across studies. A common and comparable data base of studies with rates of clinically meaningful change determined using the same parameters and methods will bring added consistency and interpretive meaning for clinicians, researchers, and clients who consume the psychotherapy research literature.

Finally, additional data regarding validity are needed. It is clear that the evolving body of literature about clinically significant change is gradually beginning to explore the factors that are associated with

change that is classified as meaningful or not. Meeting a rationally determined statistical definition of clinical significance is a solid first step, but we need better data about what improved scores on outcome measures mean when translated into functioning, quality of life, interpersonal relationships, and so on, on a person by person basis. Importantly, the validity of clinically meaningful change must incorporate change that moves in the wrong direction (deterioration). Even fairly recent studies (e.g., Fisher & Wells, 2004) focus on improvement versus no change and neglect the fact that a small portion of clients in treatment deteriorate. It should be a standard portion of any assessment of clinical significance to include data regarding those individuals who worsened during treatment.

Conclusion

Reporting the effects of interventions on clients will be advanced by including the use of classification of individual change using a standard definition of clinically meaningful change. This is an important aspect of efficacy trials and effectiveness studies but essential in methods that involve improving outcomes in patient-focused research, where methods have been developed to predict treatment failure, deliver alarm messages to therapists and clients, and provide problem-solving strategies to clinicians before clients leave treatments (Harmon et al., 2007; Lambert et al., 2003; Lutz et al., 2006; Slade, Lambert, Harmon, Smart, & Bailey, 2008). Psychotherapy research has been historically undervalued by clinicians (Morrow-Bradley & Elliott, 1986) but can have far greater impact on client well-being and clinical practice through the use clinical significance methods and increased attention to changes that occur in the life functioning of individual clients.

References

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

Åsenlöf, P., Denison, E., & Lindberg, P. (2005). Idiographic outcome analyses of the clinical significance of two interventions for patients with musculoskeletal pain. *Behavior Research and Therapy*, *44*, 947–965.

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

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

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry*, *4*, 561–571.

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.

Bergin, A. E. (1971). The evaluation of therapeutic outcomes. In S. L. Garfield & A. E. Bergin (Eds), *The handbook of psychotherapy and behavior change* (pp. 217–270). New York: Wiley.

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

Constantino, M. J., Marnell, M. E., Haile, A. J., Kanther-Sista, S. N., Wolman, K., Zappert, L., & Arnow, B. A. (2008). Integrative cognitive therapy for depression: A randomized pilot comparison. *Psychotherapy: Theory, Research, Practice, Training*, *45*, 122–134.

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

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.

Fisher, P. L., & Wells, A. (2004). How effective are cognitive and behavioral treatments for obsessive-compulsive disorder? A clinical significance analysis. *Behaviour Research and Therapy*, *43*, 1543–1558.

Gray, G. V. (2003). Psychotherapy outcomes in naturalistic settings: A reply to Hansen, Lambert, and Forman. *Clinical Psychology: Science and Practice*, *10*, 505–506.

Grundy, C. T., Lambert, M. J., & Grundy, E. M. (1996). Assessing clinical significance: Application to the Hamilton Rating Scale for Depression. *Journal of Mental Health*, *5*, 25–33.

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

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

Hansen, N. B., Lambert, M. J., & Forman, E. M. (2003). The psychotherapy dose-effect in naturalistic settings revisited: Response to Gray. *Clinical Psychology: Science and Practice*, *10*, 507–508.

Harmon, S. C., Lambert, M. J., Smart, D. M., Hawkins, E., Nielsen, S. L., Slade, K., & Lutz, W. (2007). Enhancing outcome for potential treatment failures: Therapist-client feedback and clinical support tools. *Psychotherapy Research*, *17*, 379–392.

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

Jacobson, N. S., Roberts, L. J., Berns, S. B., & McGlinchey, J. B. (1999). Method 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.

Kazdin, A. E. (1977). Assessing the clinical or applied importance of behavior change through social validation. *Behavior Modification*, *1*, 427–452.

Kazdin, A. E. (1998). *Research design in clinical psychology* (3rd ed). Boston: Allyn & Bacon.

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.

Kiresuk, T. J., Smith, A., & Cardillo, J. E. (Eds). (1994). *Goal attainment scaling: Applications, theory, and measurement.* Hillsdale, NJ: Erlbaum.

Lambert, M. J. (2001). Psychotherapy outcome and quality improvement: Introduction to the special section on client-focused research. *Journal of Consulting and Clinical Psychology*, *69*, 147–149.

Lambert, M. J., Hansen, N. B., & Bauer, S. (2008). Assessing the clinical significance of outcome results. In A. Nezu & C. M. Nezu (Eds), *Evidence-based outcome research: A practical guide to conducting randomized controlled trials for psychosocial interventions* (pp. 359–378). New York: Oxford University Press.

Lambert, M. J., Lunnen, K., Umphress, V., Hansen, N., & Burlingame, G. M. (1994). *Administration and scoring manual for the Outcome Questionnaire (OQ-45.1)*. Salt Lake City, UT: IHC Center for Behavioral Healthcare Efficacy.

Lambert, M. J., & Ogles, B. M. (2004). The efficacy and effectiveness of psychotherapy. In M. J. Lambert (Ed.), *Bergin & Garfield's handbook of psychotherapy and behavior change* (5th ed). New York: Wiley.

Lambert, M. J., Whipple, J. L., Hawkins, E. J., Vermeersch, D. A., Nielsen, S. L., & Smart, D. W. (2003). Is it time for clinicians to routinely track patient outcome? A meta-analysis. *Clinical Psychology: Science and Practice*, *10*, 288–301.

Lick, J. (1973). Statistical vs. clinical significance in research on the outcome of psychotherapy. *International Journal of Mental Health*, *2*, 26–37.

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

Lutz, W., Lambert, M. J., Harmon, S. C., Tschitsaz, A., Schurch, E., & Stulz, N. (2006). The probability of treatment success, failure and duration: What can be learned from empirical data to support decision making in clinical practice? *Clinical Psychology & Psychotherapy*, *13*, 223–232.

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

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.

Morrow-Bradley, C., & Elliott, R. (1986). Utilization of psychotherapy research by practicing psychotherapists. *American Psychologist*, *41*, 188–197.

Newnham, E. A., Harwood, K. E., & Page, A. C. (2006). Evaluating the clinical significance of responses by psychiatric inpatients to the mental health subscales of the SF-36. *Journal of Affective Disorders*, *98*, 91–97.

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. Guttenag (Eds), *Handbook of evaluation research*. Beverly Hills, CA: Sage.

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.

Openshaw, C., Waller, G., & Sperlinger, D. (2004). Group cognitive-behavior therapy for bulimia nervosa: Statistical versus clinical significance of changes in symptoms across treatment. *International Journal of Eating Disorders*, *36*, 363–375.

Schulte, D. (2008). Patients' outcome expectancies and their impression of suitability as predictors of treatment outcome. *Psychotherapy Research*, *18*, 481–494.

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

Slade, K., Lambert, M. J., Harmon, S. C., Smart, D. W., & Bailey, R. (2008). Improving psychotherapy outcome: The use of immediate electronic feedback and revised clinical support tools. *Clinical Psychology & Psychotherapy*, *15*, 287–303.

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.

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

Wolf, M. M. (1978). Social validity: The case for subjective measurement or how applied behavior analysis is finding its heart. *Journal of Applied Behavior Analysis*, *11*, 203–214.