Clinical significance: History, application, and current practice

Article *in* Clinical Psychology Review · May 2001

DOI: 10.1016/S0272-7358(99)00058-6 · Source: PubMed

3 authors, including:

Benjamin M. Ogles
Brigham Young University - Pro…
74 PUBLICATIONS **2,464** CITATIONS
SEE PROFILE

Krb Bonesteel
Ohio University
2 PUBLICATIONS **150** CITATIONS
SEE PROFILE

Pergamon

Clinical Psychology Review, Vol. 21, No. 3, pp. 421–446, 2001

0272-7358/01/$–see front matter

PII S0272-7358(99)00058-6

CLINICAL SIGNIFICANCE: HISTORY, APPLICATION, AND CURRENT PRACTICE

Benjamin M. Ogles, Kirk M. Lunnen, and Kyle Bonesteel

Ohio University

ABSTRACT. *The meaningfulness of psychotherapy outcome as measured in therapy research is a persistent and important issue. Following a period of emphasis on statistically significant findings for treated versus control groups, many researchers are renewing efforts to investigate the meaningfulness of individual change. Several statistical methods are available to evaluate the meaningfulness of clients' changes occurring as a result of treatment. This article reviews the history of the clinical significance concept; describes the various methods for defining improvement, recovery, and clinically significant change; examines current criticisms of the methods; and describes the current use of the methods in practice.* © *2001 Elsevier Science Ltd.4*

KEY WORDS. Clinical significance, Social validity, Outcome assessment.

BEGINNING IN THE 1970s (Bergin, 1971; Kazdin, 1977; Lick, 1973; Wolf, 1978), a subtle shift in psychotherapy research began to emerge. Examination of individual change occurring during psychotherapy became increasingly important. While the primary research methodology (randomized clinical trials) continued to be the preferred empirical route for the study of therapy efficacy, several therapy researchers began considering the clinical or practical meaning of change (both for the group and individual) in addition to considering statistical differences between treated groups of clients. In one respect, this movement represented a return to the original roots of psychotherapy research. In another sense, however, psychotherapy research had embarked on a path determined to demonstrate scientifically that therapy can and does help clients to observably improve. In this review, a brief history of the clinical significance movement is presented followed by a description of several current methods for examining clinical significance. Finally, a review of the use of current methods is presented.

Correspondence should be addressed to Dr. Benjamin M. Ogles, Ohio University, Department of Psychology, 241 Porter Hall, Athens, OH 45701. E-mail: Ogles@oak.cats.ohiou.edu.

A BRIEF HISTORY OF MEANINGFUL CHANGE

The earliest studies of psychotherapy relied primarily on therapist ratings of client improvement. For example, Bergin (1971) and later Bergin and Lambert (1978) review the data collected at the Berlin Psychoanalytic Institute in the 1920s. Therapists, in this case psychoanalysts, classified clients as either uncured, improved, much improved, or cured. It should be noted that only clients who had completed treatment were classified—premature terminations and ongoing cases were not included. The clinical meaningfulness of therapist ratings in this report was implicit in the categories. It was clearly assumed that the analysts had sufficient information to make classifications based on clinically meaningful changes observed in the clients. The following quote translated from the original report, cited in Bergin and Lambert (1978) is quite revealing:

> We were most particular in what was to be understood as "cured." Included were only such cases where success meant not merely the disappearance of symptoms but also the manifestation of analytically acceptable personality changes and, wherever possible, confirmative follow-up. (p. 141)

Similar definitions were also applied to the improved, much improved, and uncured groups. The degree to which therapists made reliable or valid ratings of these cases is not at issue here. Nor are we interested in the methodological problems involved with lack of a control group, dropout, and other issues. We wish only to suggest that the earliest studies of psychotherapy implicitly focused on demonstrating that clients made *clinically meaningful* change.

As the battle lines in the argument concerning the benefits of psychotherapy became clearly drawn, psychotherapy researchers began to focus on developing scientifically rigorous methods of demonstrating the efficacy of their interventions. Research designs were changed to include wait-list or no-treatment control groups. Measures were developed to assess client, therapist, and judge rated points of view (Lambert, Masters, & Ogles, 1991). These measures were also scaled such that averages could be calculated and more sophisticated statistical techniques could be used to test mean differences between groups rather than relying on descriptions of the number of people who improved. These methodologies ultimately resulted in numerous studies which, in an aggregate fashion, demonstrated the efficacy of psychotherapy (Smith, Glass, & Miller, 1980; Lambert & Bergin, 1994). Inferential statistical methods had shown that the average person who receives psychotherapy is better off at the end of treatment than 80% of people with similar problems who do not receive therapy. Rather than breathe a collective sigh of relief, however, many researchers questioned what this statistical significance really meant.

Challenging the Reliance on Statistical Tests

To be scientifically rigorous, psychotherapy researchers have utilized inferential procedures to compare group means and to examine both within- and between-group variability. If these tests of mean and variance differences are found to be beyond the range of chance (usually at the .05 level) and reliable, the effects are deemed "significant." However, this type of analysis is hampered in at least two fundamental ways (Grundy, 1994). First, because information is based on group means and variances, it is impossible to winnow out information regarding a specific client (Barlow, 1981; Hugdahl & Ost, 1981; Kaz-

din, 1977). Second, results do not easily lend themselves to interpretation—put simply, what do they really mean (Barlow, 1981; Lick, 1973)? Consequently, many researchers have decided that tests of statistical significance should have a less immutable place in outcome research (e.g., Barlow, 1981; Bergin & Strupp, 1972). Indeed, some have gone so far as to condemn this over-reliance on statistical inference in psychotherapy research as one of the "worst things that ever happened in the history of psychology" (Meehl, 1978, p. 817). Another researcher asked ironically, "can no one recognize a decisive result without a significance test?" (Stevens, 1968). Nevertheless, even more moderate critics of the established methodologies emphasized the dangers in taking statistically significant results too far (Carver, 1978). In this regard, Bergin and Strupp (1972) concluded:

> Among researchers as well as statisticians there is a growing disaffection from traditional experimental design and statistical procedures . . .With respect to inquiry in the area of psychotherapy, the kinds of effects we need to demonstrate at this point in time should be significant enough so that they are readily observable by inspection or descriptive statistics. If this cannot be done, no fixation upon statistical and mathematical niceties will generate fruitful insights, which obviously can come only from the researcher's understanding of the subject matter and descriptive data under scrutiny. (p. 440)

Nearly 20 years later, Jacobson and Truax (1991) came to similar conclusions. They suggest that while a statistical test between the means of treated and control groups gives evidence that treatment efficacy is unlikely to be the result of a chance finding, the test provides no information regarding the "size, importance, or clinical significance" of the results (p. 12). In addition, statistical tests do not provide information about within-group variation.

Magnitude of Effects

One method of evaluating the *size* of change involves the calculation of effect sizes or other similar statistics that report the degree or magnitude of the relationship between variables. A small effect would be indicative of a less meaningful result than a moderate or large effect. As psychotherapy research evolved, several researchers advocated for an increasing focus on effect sizes rather than relying on statistical tests (e.g., Cohen, 1992). Advocates of effect size point out the weaknesses of relying on the statistical significance (e.g., *p* value) to make decisions about treatments. For example, a treatment may produce statistically significant improvement when compared to a control group yet the magnitude of the change is minimal (Kendall, Marrs-Garcia, Nath, & Sheldrick, 1999). In this case, the small effect size is detected by a powerful research design (e.g., large *N*, tightly controlled methodology, etc.). In contrast, the effect size gives information as to the magnitude of the relationship between variables or the size of the difference.

Despite advantages of effect size and other measures of magnitude of relationship, effect sizes do not provide information about the clinical meaning of the findings. Effect sizes do not give us information regarding within-group variation or the clinical relevance of the results.

Within-Group Variation

In some of the earliest reviews of psychotherapy research (Bergin, 1966; Bergin, 1971; Bergin & Lambert, 1978), Bergin noticed that treatment tended to increase the vari-

ability of outcomes. Not only did more people improve while in treatment, but a larger percentage of treated individuals also deteriorated when compared to untreated individuals. Bergin suggested that this increase in variability tempered the "reported effectiveness of psychological treatments" (Bergin & Lambert, 1978, p. 152) and advocated for additional research to examine the reasons for this increased variability. Furthermore, research directed at creating more potent treatments might decrease the variability of responses to treatment, which would push the average in the direction of improvement and thus prevent the diluted effectiveness of treatment caused by increased variability.

Since Bergin noticed this difference in variability, several lines of research evolved in an attempt to better understand within-group variation in therapy studies. A small group of therapy researchers has studied deterioration that occurs during treatment (e.g., Mohr, 1995). A larger group tackled the therapy process and potential variables that might account for differential outcome among clients receiving similar treatments (e.g., Greenberg & Pinsof, 1986). Another group of researchers focused on developing more potent interventions that might reduce variability and increase the effect sizes of treatment as compared to controls (e.g., Imber et al., 1990). A final group of researchers focused on developing methods for identifying practical or clinically meaningful within-group change for individual clients without relying solely on mean differences between groups (e.g., Jacobson & Truax, 1991). It is this collection of methodologies developed to determine "clinical significance" that is the focus of this review.

Clinical Significance

One of the continuing complaints of clinicians who attempt to make practical application of psychotherapy research is the lack of information regarding the clinical significance or practical importance of research findings (Barlow, 1981; Barlow, Hayes, & Nelson, 1984; Jacobson, Follette, & Revenstorf, 1984; Kazdin, 1977; Persons, 1991). When investigators rely on statistics to inform them whether two groups are significantly different following treatment, they produce evidence that the treatment of interest is more effective than the comparison treatment or control. However, statistically significant differences between groups do not necessarily indicate practical, meaningful, or clinically significant differences between groups, nor for individuals within the groups. For example, let us consider a weight loss treatment that is compared with a control group.

Forty individuals who are at extreme risk for detrimental physical consequences related to their obesity are selected to participate in the study. The subjects are randomly assigned to either the treatment or control group. After 2 months, those receiving treatment have lost an average of 16 pounds each. The comparison group, however, has lost on the average no weight during the elapsed time. The statistical test reveals a significant finding for the treatment group as compared to the control group. These statistical effects suggest that differences between the groups are real as opposed to differences that are "illusory, questionable, or unreliable" (Jacobson & Truax, 1991, p. 12). However, the statistical test does not give information regarding the variety of responses to treatment within the treated group. With an average of 16 pounds weight loss, some individuals who received treatment may have lost 30 pounds or more while others who received treatment lost no weight or even gained weight! Similarly, one must question the practical significance of a 16-pound weight loss for a

morbidly obese individual. Does this amount of change reduce risk of mortality or improve the quality of life of the individual? These questions are not answered by a statistical test or an effect size.

To ameliorate this situation, researchers have attempted to establish methods to measure clinically significant rather than statistically significant changes (Jacobson, 1988; Jacobson, Follette, & Revenstorf, 1984; Jacobson, Dobson, Fruzzetti, Schmaling, & Salusky, 1991; Jacobson, Roberts, Berns, & McGlinchey, 1999; Kazdin, 1977; Wolf, 1978). Two parallel and mutually beneficial lines of research evolved to more clearly demonstrate the utility of psychological interventions. In the mid to late 1970s Social Validity (Kazdin, 1977; Wolf, 1978) emerged as a method of acknowledging the importance of including the perspective of individuals outside the therapeutic relationship to help determine the importance of psychosocial interventions and outcomes. It provided a cohesive rationale and two specific methodological tactics (subjective evaluation and social comparison) to evaluate the relevance of change. Later, and as a natural progression of social validity, methods for determining the clinical significance of interventions were developed. While social validity emphasizes a broader examination of practical change from the perspective of participants and societal members, clinical significance takes a slightly narrower view of meaningful change by identifying methods defined by clinician-researchers.

SOCIAL VALIDITY

The roots of the social validity movement were found in applied behavioral analysis. Wolf (1978) describes the subtle trend of increased awareness of the "social importance" of interventions among applied behavioral analysts in this way:

> The message we seemed to be getting was that "social importance" was a subjective value judgment that only society was qualified to make. If our objective was, as described in *JABA (Journal of Applied Behavior Analysis)*, to do something of social importance, then we better develop systems and measures for asking society whether we were accomplishing this objective. (pp. 206–207)

This societal input is viewed as important on three levels (Kazdin, 1977). First from the standpoint of the focus of the intervention itself—are the *goals* of the treatment/behavior modification in harmony with societal goals? Second, are the mandate *procedures* of the given intervention acceptable with regard to societal appropriateness (e.g., do the ends justify the means)? Finally (and most importantly within the context of this particular article), are the *effects* of the intervention deemed as important by society? Central to this aspect of social validity is determining if the effects of treatment are indeed clinically or practically important. Kazdin (1977) suggests two primary ways to evaluate this question: (1) the subjective evaluation method, and (2) the social comparison method.

Subjective Evaluation

In this method the client's behavior is evaluated by individuals who are "likely to have contact with the client or in a position of expertise" (Kazdin, 1998, p. 387). This allows the researcher to tap whether the client has made qualitative changes that are in turn observable by others. A study of predelinquent boys conducted by Werner et al.

(1975) provides an example of the use of this method. In this study a list of important behaviors related to "positive" police-suspect interactions was compiled. Predelinquent participants were in turn trained to incorporate these behaviors. Trained boys consistently outperformed nontrained boys in a simulated role-play situation with an actual police officer. This was determined by observation of target and nontarget behaviors, respectively. Videotapes of the sessions were then judged by police officers, citizens, and college students. Judges consistently rated the trained boys as lower in "suspiciousness" and higher in "cooperativeness," "politeness," and general better behavior. The fact that these findings were consistent across all subjects and raters implies that the socially important changes did in fact occur as a result of the training program.

Addition of the societal perspective to help subjectively evaluate the meaningfulness of change supplies additional information that is unique from that of either the professional or the client themselves. For example, while both a therapist and her/his client may feel that significant changes have occurred as a result of therapy, the client's spouse may not agree. This disagreement may provide insight on aspects of the client's condition missed by the other two parties. In addition, with the ever-increasing involvement of third-party payments and government-sponsored mental health care programs, society has "invested" in the therapeutic process and consequently has some claim to an accounting of what has been done with its money.

> As long as Jones paid me for his psychotherapy or friendship, or however he wanted to use the time I sold him, it was none of Smith's business. But when Smith's taxes or insurance premiums began to contribute to my fee, Smith's interest in what I was doing with Jones increased. In other words, Smith now expects me to be accountable—and in terms that he can understand. (Aldrich, 1975, p. 509)

The increasing emphasis on consumerism also supports the focus on gathering data from individuals outside the treatment to help make judgments about effective treatment. As a result, researchers began to consider the social relevance of behavioral treatments through the collection of subjective evaluations of treatment. Therapy studies have included a range of outside observers (e.g., nurses, parents, spouses, bosses, teachers, etc.) who rate the "validity" of the outcome.

Subjective evaluations, however, are also limited by a number of important considerations. While subjective information provided by individuals not involved directly in the treatment of the client may provide additional insight and help corroborate improvement, there is a danger that this information will be stretched beyond its applicability and used as a type of prescriptive guideline (Kazdin, 1977). For example, if hospital workers were used for subjective evaluations of patients in their respective wards, the information they provided would invariably be colored by the context in which they are dealing with the given client (e.g., they would be likely to focus on behaviors that relate to "manageability" and hospital appropriateness (Kazdin, 1977). As a result, the researcher must proceed with caution and remind him/herself that this subjective information is based on imperfect, external judgments that may or may not hit upon the most relevant aspects of the client's behavior.

The major thrust of subjective evaluation within the social validity movement was to incorporate the societal viewpoint into the process of determining the meaningfulness of change. Strupp and Hadley (1977) extended subjective methods of social validity with regard to psychotherapy itself by providing specific guidelines regarding who should provide information. They determined that to rely on traditional thera-

pist outcome reports alone was insufficient, and that relying on societal factors alone was also insufficient. They concluded that "only by considering multiple perspectives will it be possible to derive a truly comprehensive definition of mental health and meaningful psychotherapy outcomes" (p. 187). They suggest that there are three such perspectives that are of particular interest: (1) the client themselves, (2) the mental health professional, and (3) society (including significant persons in the individual's life). Although this extends beyond the traditional view of social validity, a brief description of the two additional participants and their "subjective evaluations" of therapy deserve mention here.

Individual perspective. The individual defines successful psychotherapy as that which makes him/her feel "better." This judgment is independent of the therapist and society, who consequently may not necessarily agree with them. However, while this may be the case, the individual him/herself is the *only* judge with direct, internal, and intimate knowledge of what real changes have occurred as a result of therapy. To disregard this perspective seems to fly in the face of the underlying circularity of the entire psychotherapeutic process itself. In other words, the entire therapy process is reliant on the client's own subjective feelings and imperfect communications with the therapist; as such, it seems ludicrous to rely on this information throughout the treatment experience only to throw out similar subjective information when outcome is considered.

However, accepting such subjective information does have drawbacks. Wolf (1978) points out:

> When we are asking for a verbal description of a private event, such as satisfaction with our treatment program, we must be very cautious because we have no adequate way of checking the reliability of the verbal report in an independent way. (p. 212)

In addition, social desirability, demand characteristics, and response sets can all impact the reliability of self-report information (Nichols, Greene, & Schmolck, 1989). Results of dozens of studies indicate that some people have a tendency to adjust their responses in an attempt to portray themselves in what they perceive is a socially desirable way (Meyers, 1990). For example, Zanna and Olson (1982) found that, when in the presence of others, subjects were more likely to endorse test items that espouse opinions contrary to their own but in agreement with the "majority view."

Despite the potential for socially desirable responding, most methods for establishing the benefit of treatment and for determining clinically significant change are based on the client's report of change in symptoms or functioning. The majority of the methods described below base the definition of clinical significance on changes in symptoms reported by clients. In addition to these symptom-based definitions, a major movement in outcome research involves the assessment of quality of life as a way of demonstrating treatment relevance.

In the past two decades, changes in daily life functioning have become increasingly important indicators of "real" change. Particularly in health care and service to people with chronic mental illness, symptomatic change is only part of the expected result of treatment. Unless changes occur in the client's ability to function at work, play, home, and in social situations, symptomatic change is not "meaningful." In fact, Kaplan (1990) suggests that the only important dependent variables in health psychology research are quality of life and mortality. He argues convincingly that other measures

such as cholesterol, blood pressure, stress, coping, and so forth have little utility unless the quality of life for the person is improved or the risk of mortality decreases. Similarly, quality of life is an important indicator of current mental health or outcome of therapy. If the client's quality of life is improved, then the treatment outcome can be assumed to be reasonable.

Quality of life assessment can be conducted within any of the perspectives mentioned above (e.g., client, therapist, society). Historically, quality of life was measured by objective observable standards (e.g., in economic research by monetary indicators, or in healthcare by patient mobility, etc.). Researchers are, however, increasingly relying on subjective evaluations of well-being, life functioning, or quality of life (see, for example, Bigelow, McFarland, & Olson, 1991; Frisch, 1994). A myriad of specialized quality-of-life instruments are also available in health psychology. Instruments have been developed for geriatric patients, cancer patients, and several other specific populations.

Assessing client's quality of life is potentially an important contributor to the evaluation of clinically significant treatment affects. Relatively little research has been conducted to assess the complete spectrum of quality of life in psychotherapy outcome research. However, many of the dimensions of quality of life or areas of functioning have been assessed. For example, social relationships have long been considered an important focus of psychological treatments and an indicator of positive outcome in therapy. Similarly, symptom-based measures like the Beck Depression Inventory (BDI; Beck, Steer, & Garbin, 1988) partially assess the level of functioning related to the emotional well-being area of quality-of-life. Some outcome measures used in psychotherapy research match other areas of life functioning identified in global quality-of-life measures. Overall quality-of-life ratings generated from instruments created specifically to assess the level and quality of functioning in theoretically important areas of living provide an additional outcome perspective that may enhance the meaningfulness of outcome data (Gladis, Gosch, Dishuk, & Crits-Christoph, 1999).

Professional perspective. Traditionally, the perspective of the mental health professional has served as the primary source of outcome information (Barlow, 1981; Strupp & Hadley, 1977). Overreliance on this single information source largely arises from the notion that only the professional has the "expertise" or statistical sophistication to accurately evaluate the effects of treatment. Perhaps the professional is the most likely to rely on more theoretical, quantified information to determine if treatment is helpful. Yet, Strupp and Hadley (1977) argue that using the therapist as the *only* informational perspective may not be the best scenario. Certainly there is much to be said about the value of his/her input into the overall outcome picture. The professional's perspective, in contrast to the other two, is based on extensive knowledge of personality structure and exhaustive training in behavioral observation and assessment. Considering treatment effects from the vantage of this expertise provides essential additional information.

The professional perspective is not always limited to the therapist who is conducting the intervention. In many therapy studies, a trained rater (often a professional or student in training) conducts interviews or other ratings based on observations of the client. Independent raters have the benefit of the professional perspective without the potential bias that may come with extensive involvement and investment in the treatment. At the same time, the independent rater's judgment about the meaning of change is likely to be similar to the therapist's perspective.

These three perspectives (client, therapist, society) form the bedrock of subjective outcome evaluation. The individual can provide unique internal and intimate knowl-

edge of treatment effects. The professional is able to view treatment effects in light of his/her expertise and training. Finally, the societal vantage provides objective information regarding the day-to-day impact of services on the individual's behavior on those around him/her. Only when all three viewpoints are considered will a truly comprehensive picture of the effects of treatment emerge.

Social Comparison

Subjective evaluation by individuals with expertise or contact with clients is one way to approach social validity. A second social validity methodology recommended by Kazdin (1977) is social comparison. In this method treatment effectiveness is evaluated based on pre- and postevaluations of the client's behavior with a reference group of "nondeviant" peers. The underlying premise is that socially valid, or clinically significant, changes due to the intervention in question will result in the client's postintervention behavior being indistinguishable from a normal reference group. Patterson (1974) used this method in evaluating the impact of behavioral interventions with misbehaving schoolboys. The pretreatment numbers of disruptive acts (such as teasing, yelling, whining, and fighting or hitting) performed by these boys were compared with the number of similar acts performed by a matched "nondeviant" group of boys and were found to be significantly higher. The comparison was repeated at posttreatment. Number of disruptive acts performed by the experimental group decreased to a level that was well within the range of the control group; consequently, magnitude of these changes represent socially and clinically important improvement.

There are some important considerations to this method of social validation. Foremost among these is whether or not there is in fact a "normative" level and consequent appropriate control for the phenomena or behavior(s) in question. The normal level of functioning may not be an effective standard by which outcomes for certain behaviors should be evaluated. Kazdin (1977) uses the example of behavioral techniques to increase recycling of materials and other environmentally relevant behaviors which seek to increase *all* levels (including the "normal" level) of behaviors in this area. If the goal were to increase recycling, meaningful change would not necessarily be represented by the current behaviors of the majority.

Another consideration is how the researchers are to identify the normative group itself. With profoundly autistic children, for example, would normative comparison be more appropriate with less-pronounced autistic children or with "normal" children as a reference? Comparison with the former may not necessarily represent a socially valid change, while comparison with the latter may be so stringent that successful outcome is impossible. For example, in one study, verbalization of a psychiatric patient increased after training (Stahl et al., 1974). The increase was very discrepant (about 30%) from the level of intelligent, normally functioning individuals, but was close (about 9%) to the level of other hospitalized psychiatric patients of similar education who were not considered verbally deficient (Kazdin, 1977, p. 441). Consequently, the decision as to which "normative" sample will serve as the reference group can tremendously effect conclusions of clinical importance or lack thereof.

The social validity methodology (Foster & Mash, 1999; Kazdin, 1977; Wolf, 1978) is important not only from an informational standpoint but also because it serves to provide some cohesive professional direction. "It seems to us that by giving the same status to social validity that we now give to objective measurement and its reliability, we will bring the consumer, that is society, into our science, soften our image, and make more

sure our pursuit of social relevance" (Wolf, 1978, p. 207). While the social validity methodology does not completely satisfy the question of how to determine clinically significant improvement, it provides some promising methods for evaluating meaningfulness of change and lays the foundation for newer approaches.

METHODS FOR EXAMINING CLINICAL SIGNIFICANCE

From the foundation of social validity, several specific methodologies have been developed to examine the clinical relevance of changes occurring during psychological treatment. The methods vary depending on the individual's definition of meaningful or clinically significant change, but all methods are more narrowly focused on researcher-defined or "clinical" definitions of significant change. The three most prominent definitions of clinically significant change include: (1) treated clients make statistically reliable improvements as a result of treatment (improvement), (2) treated clients are empirically indistinguishable from "normal" or nondeviant peers following treatment (recovery), or (3) a combination of return to normal functioning plus reliable improvement.

Improvement

Improvement is defined as making statistically reliable change. This method for defining clinical significance is based on the social validity idea that clients make meaningful change when that change is sufficiently large that it is easily noticed by others (Wolf, 1978). In one way, this is a return to some of the original psychological research of the late 19th century. Although Weber's (1978) exploration of the Just Noticeable Difference (JND) phenomenon was focused primarily on physiological perceptions, the principles involved apply equally in this circumstance. Do clients make changes that are perceptually noticeable by the therapist, the spouse, the boss, or themselves? Beginning with this basic idea, several researchers developed statistical methods for calculating the reliability of change scores. Although not a direct application of the JND concept, a statistical calculation is similar since it identifies a point at which the change is sufficiently large as to be confidently considered reliable and not the product of error. One might still ask, however, if the change is sufficiently large to be noticeable. Both issues, large enough to be reliable and large enough to be noticeable, are important. Let us consider each in turn.

Reliable change. Several methods for evaluating the reliability of change scores are available in the literature. The most frequently used in practice is the Jacobson-Truax method (Jacobson & Truax, 1991). Jacobson and Truax calculate a Reliable Change Index (RCI) for each individual based on the pretreatment score (X_{pre}), the posttreatment score (X_{post}) and the standard error of the difference between two test scores (S_{diff}):

$$RCI = \frac{X_{post} - X_{pre}}{S_{diff}}$$

The change is considered reliable, or unlikely to be the product of measurement error, if the change index (RCI) is greater than 1.96. When the individual has a change score greater than 1.96, one can reasonably assume that the individual has improved.

Other less frequently used methods for statistically determining reliable change are available. For example, Speer and Greenbaum (1995) reviewed and evaluated four methods of calculating the significance of individual client change using pre- and posttreatment scores: Edwards-Nunnally (Edwards, Yarvis, Mueller, Zingale, & Wagman, 1978), Jacobson-Truax (Jacobson & Truax, 1991), Hsu-Linn-Lord (Hsu, 1989), and Nunnally-Kotsch (Nunnally & Kotsch, 1983). Each of these methods examines the posttreatment score in relationship to the pretreatment score while considering the reliability and distribution of test scores (e.g., mean and standard deviation for either the treated or a normative sample). In addition, Speer and Greenbaum compared a hierarchical linear modeling (HLM) method (which allows for the inclusion of multiple data points) with the other methods. The HLM method is particularly useful for clinical data that has missing data points. The HLM modeling process estimates slopes or rates of change using available data. As a result, rates of improvement for a given sample can be projected based on the data available. In this way, individuals that might otherwise be classified as nonchangers due to missing data could be identified as improvers using HLM estimates.

Speer and Greenbaum (1995) first described, then compared the five methods for identifying improvers. They suggest that the Edwards-Nunnally and HLM methods identified significantly more clients as improvers. It is difficult to ascertain what that might mean, however, since one method may be more or less *conservative* rather than more or less *valid*. In addition, Speer and Greenbaum (1995) did not use the same basic assumptions for each formula (i.e., different population parameters were used in the different calculations). As a result, the results may reflect differences in parameters rather than differences in the formulae.

Several other methods for calculating improvement are available. For example, when examining the effectiveness of psychological interventions for headaches, investigators calculate the percent improvement based on the frequency and severity of headaches per week (obtained from headache diaries) at pretreatment and posttreatment (Blanchard & Schwarz, 1988).

$$\text{Percent Improvement} = 100 \times \frac{\text{Headache Index Pretreatment} - \text{Headache Index Posttreatment}}{\text{Headache Index Pretreatment}}$$

Using the percent improvement, a 50% reduction in headache activity, in the absence of increased medication, is defined as clinically significant improvement (Blanchard & Schwarz, 1988).

Similarly, agoraphobia researchers have developed criteria for identifying "improvement" based on the combined changes in ratings on several outcome measures (Michelson, Mavissakalian, & Marchione, 1985). A score of 1 is assigned to a change ≥ 2 on each of the instruments used for assessing outcome. Clients are then classified as low, medium, or high improvement following treatment based on their total score (e.g., high improvement 3 or 4, medium improvement 1 or 2, and low improvement 0). In this way, the investigators have evidence of improvement based on a priori cutoff points from a combination of self-report, judge-rated, therapist-rated, and behavioral approach measures of outcome.

Noticeable change. Although changes from pre- to posttreatment may be large enough to be statistically reliable for a given individual, this does not guarantee that the

change will be noticeable or meaningful to the client or others. Returning to the obesity treatment example, one individual may make a reliable change in weight following treatment. In fact, given the test-retest reliability of weight measurement, very little change would be necessary to produce a change that was statistically reliable and unlikely to be the product of measurement error. However, a statistically reliable change may not denote a noticeable change.

Ankuta and Abeles (1993) were the first to address this question more directly. They compared clients who demonstrated clinically significant improvement according to Jacobson and Truax's methodology (as measured on the SCL-90-R; Derogatis, 1983) with the client's own perceived satisfaction with therapy. They operationalized "satisfaction" as extent of self-reported change resulting from therapy (as measured on Strupp, Fox, and Lessler's [1969] Patient Questionnaire). They found that clients designated as having experienced "clinically significant" improvement did indeed report higher levels of satisfaction than those experiencing "non-clinically significant" change. This provides important initial evidence of the validity of Jacobson and Truax's methodology and suggests that changes were at least noticeable to the clients.

Lunnen and Ogles (1998) expanded on Ankuta and Abeles' evaluation by performing a multiperspective, multivariable analysis of the RCI component of Jacobson and Truax's methodology. They divided clients who were receiving outpatient therapy into one of three groups based on their change scores on the Outcome Questionnaire (OQ-45.1): improvers, no-changers, and deteriorators. When clients demonstrated reliable change—either improvement or deterioration on the OQ-45—they were matched with clients who were unchanged. Clients in all three groups then rated their perceived change, satisfaction with treatment, and the helping alliance. Their spouses/significant others also rated perceived change and satisfaction with the treatment, and the therapist rated perceived change and the helping alliance.

Lunnen and Ogles (1998) found that both perceived change and therapeutic alliance were significantly higher for individuals who reliably improved than for nonchangers and deteriorators from both client and therapist perspectives. Satisfaction with services, however, did not differ among the groups. Clients demonstrating reliable deterioration were not significantly different from nonchangers on any of the outcome variables reported by any of the three perspectives. They concluded that the RCI is an effective way of evaluating symptomatic improvement, but that it is less effective as an indicator of deterioration.

These two studies begin the process of demonstrating that the statistical methods for identifying improvement may also be valid indicators of noticeable or meaningful change to the client, the therapist, and others. Certainly, more research in this area is needed to examine the correspondence of reliable change and noticeable or meaningful change.

Recovery

Kendall and Grove (1988) suggest taking the perspective of the "skeptical potential consumer" of psychological treatments to better understand the concept of clinical meaningfulness. In order to convince the skeptic, an intervention must be of practical value and should lead to "changes that materially improve the client's functioning" (p. 148). They go on to suggest that the most convincing demonstrations of treatment efficacy provide evidence that once troubled clients are now "not distinguishable from a . . . representative nondisturbed reference group" (p. 148). In other words, if we can

demonstrate that clients are easily distinguished from a group of peers before treatment while after treatment their behavior is indistinguishable from peers, we have demonstrated a clinically meaningful change. This approach is slightly different from the social validity methods discussed above. In the social validity context, a social comparison was performed through direct observational ratings. In Kendall and Grove's approach the comparison is empirical and based on normative distributions of outcome measures. In this case, a return to the empirical norm represents a significant and meaningful change (Kendall et al., 1999).

Several empirical methods for determining the return to normal are available. An obvious method would be to conduct a diagnostic interview prior to and following treatment. Individuals would be expected to meet diagnostic criteria for a disorder prior to treatment. Following treatment, however, one might expect that their symptoms and signs would be diminished or alleviated such that they no longer meet the criteria for the diagnosis. For example, Barrett, Dadds, and Rapee (1996) conducted a study to intervene with childhood anxiety. They reported between 57% and 71% (depending on the treatment group) of the children no longer met criteria for an anxiety disorder following treatment.

An even more statistically oriented approach involves the use of existing normative data for a given measure of pathology. Treated clients' scores on the measure are then compared with the normative distribution to determine if they have recovered or not. A variety of methods can be used to determine cutoff scores or percentile levels at which recovery is defined (Kendall & Grove, 1988; Kendall et al., 1999). For example, Elkin et al. (1989) considered the clinical significance of the NIMH treatment for depression collaborative research program (TDCRP) by identifying the number of clients who "met a predefined level of clinical recovery" (p. 974). Recovery was defined as a score of 6 or less on the Hamilton Rating Scale for Depression (HRSD; Hamilton, 1967) or 9 or less on the Beck Depression Inventory (BDI; Beck et al., 1988). These cutoffs were determined by referring to previous research, which indicated that few remaining symptoms of depression occurred at scores of this level.

A slightly different method involves considering the functioning of clients following treatment without comparison to a normative sample. For example, agoraphobia researchers developed criteria for identifying "endstate functioning" based on the combined posttreatment ratings of several outcome measures (Michelson, Mavissakalian, & Marchione, 1985). Clients were given one point each for specified ratings on several outcome measures: (a) <3 on the Global Assessment of Severity; (b) <3 on the Self-Rating of Severity; (c) <4 on the Phobic Anxiety and Avoidance Scales; and (d) 20 on a Behavioral Avoidance course with <4 on the Subjective Units of Discomfort during the approach test. High endstate functioning was defined as a score of 3 or 4, medium endstate functioning was defined as a score of 1 or 2, and low endstate functioning was defined as a score of zero.

A variety of other methods, cutoffs, or comparisons can be conducted to evaluate whether clients recover following treatment (see Kendall et al., 1999). In addition, some methods attempt to combine the notions of improvement and recovery. Perhaps the most widely known and used is the method developed by Jacobson and colleagues.

Improvement Plus Recovery—Clinical Significance a la Jacobson

Jacobson and colleagues (Jacobson, 1988; Jacobson, Follete, & Revenstorf, 1984; Jacobson & Revenstorf, 1988; Jacobson et al., 1999; Jacobson & Truax, 1991) combined the notions of improvement and recovery to determine the clinical significance of in-

dividual change. Jacobson and Truax (1991) propose two criteria for assessing clinical significance.

First, clients receiving psychological interventions should move from a theoretical dysfunctional population to a functional population as a result of treatment (recovery). In other words, if the distributions of individuals in need of treatment and "healthy individuals" are represented graphically, the *treated* client should be more likely to be identified as a member of the healthy distribution. For example, a depressed client receiving cognitive therapy must have a BDI score after treatment that is more similar to the scores for the general population than to the results of untreated depressed individuals. This follows the work of Kendall and Grove (1988), who developed statistical methods for comparing treated clients with normative groups.

Second, the change for a client must be reliable—the pre- to posttreatment change must be large enough while considering the reliability of the instrument and the variability of the normative group that differences can be attributed to "real" change and not to measurement error (improvement). To determine the reliability, Jacobson and Truax (1991) use the reliable change score formula described earlier.

If the client meets both criteria, movement from one distribution to the other and an RCI greater than 1.96, then the change is considered "clinically significant." Interested readers can refer to Jacobson and Truax (1991) for an example of this method using the Dyadic Adjustment Scale as a measure of outcome for marital therapy, or Ogles, Lambert, and Sawyer (1995) for an example of clinical significance using the National Institute for Mental Health treatment for depression collaborative research program data. Similarly, Ogles, Lambert, and Masters (1996) present data useful to calculate clinical significance for several commonly used measures of psychotherapy outcome. While the ideas are fairly straightforward, several additional issues and difficulties with this method must be addressed.

When the functional and dysfunctional distributions are overlapping, several different cutoff points (or clinical cutoff indices) may be used to determine criterion 1. Jacobson and Truax (1991) suggest three possible cutoff points—the posttreatment score is considered part of the functional distribution when it falls within 2 standard deviations of the mean of the functional group, at least 2 standard deviations away from the mean of the dysfunctional group, or at least halfway between these two points. This is not a complicated task if one knows the distributions for both the dysfunctional and functional groups. However, adequate normative distributions may not be available for many psychological instruments used to assess outcome.

An increasingly large number of studies are using the Jacobson method to investigate the clinical significance of individual change within outcome studies. One way of utilizing the Jacobson method involves the graphic depiction of pre- to posttreatment change. In Figure 1, a graph with the pretreatment score on the Ohio Scales Parent Rated problem severity scale on the x-axis and the posttreatment score on the y-axis is presented. The horizontal line represents the posttreatment cutoff score necessary to be considered part of the functional distribution. The center diagonal line running from corner to corner is the line of no change. Clients who have the same pretreatment and posttreatment scores will be plotted on this line (Client A). The dashed diagonal lines on either side of the "line of no change" represent the change scores necessary to result in an RCI greater than 1.96. Clients between the dashed diagonal lines (Client B) did not improve sufficiently to rule out random fluctuations or test unreliability as the source of the change. Clients plotted outside (above the top diagonal

line or below the bottom diagonal line) the dashed lines can be considered to have made reliable changes for the better (Client C) or for the worse (Client D). Individuals who made reliable changes for the better and had end-of-treatment scores similar to the functional population are plotted below the diagonal and the cutoff score (Client E). Overall, the Jacobson and Truax (1991) method presents a useful approach for determining the clinical significance of individual change occurring during treatment.

Some Criticisms of Current Methods

While the methods presented here provide novel and practical approaches to demonstrating improvement, recovery, or both, problems also exist in terms of the validity of the instruments, multiple measures, potential rater bias, regression to the mean, base rates of change, and the limits of a functional distribution.

The first problem has to do with the validity of the instruments used to assess the clinical change. For example, the SCL-90R may be an adequate indicator of the number and intensity of symptoms endorsed by a person, yet a decrease in reported symptoms may or may not correspond to behavioral changes. In addition, clients entering treatment do not always appear dysfunctional on outcome measures either because of lack of sensitivity of the measures, measurement error, or perhaps temporary fluctuations in the symptoms (Saunders, Howard, & Newman, 1988). Having a change score that is reliable or a posttreatment score that falls within the normal distribution is one piece of evidence of meaningful improvement. Yet the validity of these statistically derived indicators have not been definitively substantiated.

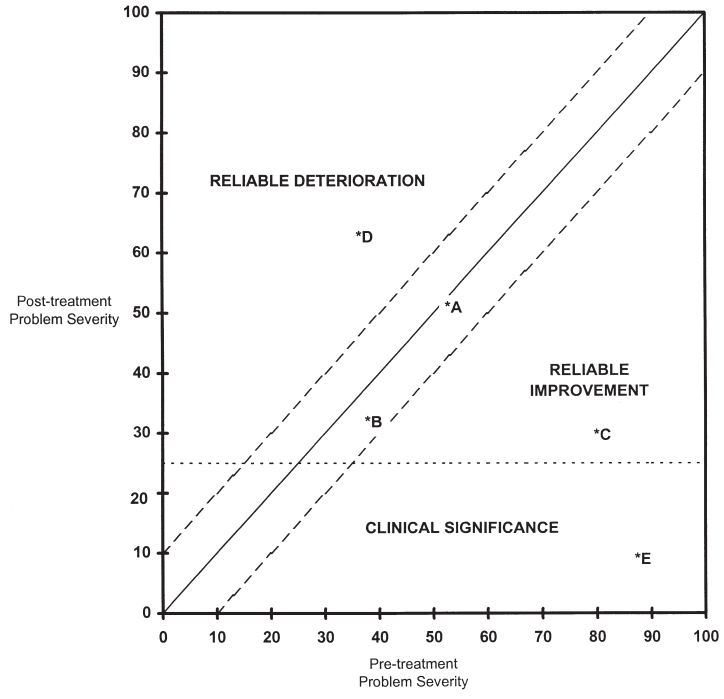

FIGURE 1. Clinical Significance Using the Ohio Scales Problem Severity Scale.

Similarly, most instruments are unidimensional, while people in treatment present with multidimensional clinical problems. Should we then require that a client show clinically significant change on several measures of the problem to be considered meaningfully improved? And what then do we do in cases of desynchrony (e.g., agoraphobia, where we might observe behavioral change with no accompanying physiological change)? One study evaluated the correspondence of multiple measures of outcome using Jacobson's clinical significance method and the Treatment for Depression Collaborative Research Program Data (Ogles, Lambert, & Sawyer, 1995). When comparing the number of individuals who could be classified as clinically significant changers using three different measures (Beck Depression Inventory, Hamilton Depression Rating Scale, and Hopkins Symptom Checklist), a reasonable degree of correspondence existed among the measures. Over 75% of the clients were classified by all three measures consensually. Nevertheless, 25% of the clients made changes on one measure, but not on others. This potential disagreement among measures will require further study. Importantly, definitions of outcome must be considered and reconsidered.

Another problem with some methods for assessing clinically meaningful change involves the problem of rater bias. Perhaps self-report instruments are too "reactive" (c.f., Smith, Glass, & Miller, 1980) to be used for judging clinical meaningfulness, particularly when social validity implies that someone other than the client can observe the utility of the change that has occurred. Blanchard and Schwarz (1988) suggest that clinical meaningfulness involves objective observable criteria. Perhaps the original investigations of social validity in which clients were observed with peers, or videotaped doing role-played scenarios would be preferred methods of determining clinical utility. At the same time, many self-report measures are more objective (e.g., self-monitoring, diaries, etc.) when they involve frequency ratings. A more exacting and detailed look at the variability among self-report measures may reveal important differences that influence the assessment of clinical change.

A third problem with clinical significance when represented by reliable change scores involves regression to the mean. Speer (1992) argues that the methods used to calculate reliable change could be biased by regression to the mean. That is, those individuals who have high pretreatment scores on the given outcome measure may be the most likely to make large improvements. Speer (1992) recommends an alternative method for calculating clinical significance when regression to the mean is identified empirically. This method may be particularly useful to administrators who are summarizing evaluation data for a clinic or center. In these circumstances, no comparison group is used, and data collected by the clinic indicating positive change for many clients may be a function of regression to the mean.

The process of adjusting for regression to the mean basically involves two steps: (1) checking to see if regression to the mean is operating within the sample data, and (2) if so, adjusting scores prior to calculating improvement rates that are based on size of change. Speer (1992) presents the statistical methods for conducting both portions of the adjustment. The more recent use of the RCI methods presented in the literature typically incorporate Speer's recommendations (i.e., Ogles, Lambert, & Sawyer, 1995).

Some researchers also argue that methods for classifying clients who move from the dysfunctional to functional distributions do not consider the base rate of movement between the two distributions (e.g., Hsu, 1996). As a result, they propose modifications to the formulas that strengthen conclusions made when categorizing clients into groups based on their posttreatment scores.

Finally, Tingey, Lambert, Burlingame, and Hansen (1996) argue that Jacobson's method is too conservative, since a client who was severely decompensated and then improved to the level of a mild disturbance would not be considered clinically improved. Although it may be accurate to conclude that the person is not part of the functional distribution, a person with chronic illness may be "meaningfully" improved at a mild level of dysfunction, even from a social validity point of view. Similar problems occur in medical treatment where return to normal functioning is impossible (e.g., lost limbs, chronic illness, etc.). No one would argue that a prosthesis functions exactly like a lost limb. Yet, a prosthesis may facilitate numerous tasks in a meaningful way. The ultimate question then becomes, how many capabilities or functions must be added by the prosthesis before it can be considered to create a clinically relevant and meaningful change? Similarly, if a person with a psychological dysfunction reliably changes yet may never fall within the boundaries of a functional distribution, should we rule out clinically meaningful change? And if not, how much change should they make before we call the change clinically meaningful?

Tingey et al. (1996) suggest identifying multiple distributions which can then be used to describe a continuum of dysfunction. Specifically, they identified four populations using the Symptom Checklist-90 Revised: an asymptomatic group, a normal population, a mild disturbance group, and a severely disturbed group. In this case clinical significance does not require movement into the "functional" distribution, but rather movement into the next or adjacent distribution regardless of where the client starts. This extension of the method seems particularly relevant for the treatment of people with chronic disturbance, where clinical significance may be measured in terms of rehospitalization rates rather than by comparisons to a "normal" reference group. However, few instruments have identifiable distributions along a continuum of severity. Certainly much more work needs to be done before we can easily identify clinically meaningful change in this way. Nevertheless, interested readers can attempt to develop multiple distributions as needed (see, for example, Grundy, 1994; Grundy, Lambert, & Grundy, 1996).

Despite these varied criticisms, methods for examining improvement, recovery, or both are becoming more widely accepted and used. To further examine current use of these methods in practice we conducted a review of outcome studies published in one journal.

CLINICAL SIGNIFICANCE IN PRACTICE

To get a sense of how clinical significance has been reported in practice, we reviewed articles in the *Journal of Consulting and Clinical Psychology* over a 9-year period (1990–1998). During this period of time, 74 studies were identified that conducted analyses to further examine clinically significant change. Overall, a wide spectrum of interventions, populations, and designs were represented within this 9-year period. Furthermore, some authors relied heavily on these techniques in interpreting their data, while others incorporated this type of analysis more as an afterthought.

Although clearly articulated methods of calculating clinical significance, recovery, or improvement do exist (Blanchard & Schwarz, 1988; Jacobson et al., 1984; Kendall & Grove, 1988), there was considerable variation in authors' application of these methods. In terms of the treatment outcome literature within this journal, there seem to be five primary ways in which percentage/proportion values are obtained.

Jacobson's Clinical Significance: Reliable Change and Clinical Cutoff Criterion

Approximately 35% (26 of 74) of the studies sampled used the Jacobson and Truax (1991) method for determining clinical significance. To be considered in this category, authors had to clearly articulate that both a reliable change index (RCI) and a clinical cutoff were calculated. This appeared to be the most difficult for many investigators to implement, with a number opting to calculate only a clinical cutoff and others calculating only the RCI. It appeared that those who calculated both indices were able to do so primarily because of the psychometric data available on the measures used in their studies.

Even within the Jacobson methodology, however, there was considerable variation. Recall that there are three ways in which the clinical cutoff may be determined. Recovery can be defined as placing the treated individual within 2 standard deviations of the normative or functional sample (criterion A), 2 standard deviations away from the pathological or dysfunctional sample (criterion B), or at the midpoint between the 2 samples (criterion C). Among the studies sampled, it was sometimes difficult to ascertain exactly which cutoff point was being used. On a number of occasions, Jacobson's methodology would be cited without a detailed explanation of how the clinical cutoff criterion requirement was met. In general, all possible ways of calculating this index were represented. Some authors were concerned with movement toward the normative sample (e.g., Barkham et al., 1996; Jamison & Scogin, 1995; McLean & Hakstian, 1990), while others required movement beyond the midway point (e.g., Barkley, Gueyremont, Anastopoulus, & Fletcher, 1992; Davis, Olmsted, & Rockert, 1990; Gallagher-Thompson & Steffen, 1994; Shefler, Dasberg, & Ben-Shakhar, 1995), or away from dysfunctional group means (e.g., Muran et al., 1995; Propst, Ostrom, Watkins, Dean, & Mashburn, 1992). Still others considered clinical cutoffs related to their own samples' pretreatment means (e.g., Foa, Ruthbaum, Riggs, & Murdock, 1991; Pelham et al., 1993).

Normative Comparisons

Jacobson's recovery. Nearly a third of the studies, 28% (21 of 74), relied on some kind of normative comparison approach for assessing clinical significance. Of the 21 normative comparison studies, a majority (13) used a variation of Jacobson's method to arrive at percent values for improvement. These approaches were similar in that they choose to calculate only a clinical cutoff score, without also calculating the RCI for each client (e.g., Jones, Ghannam, Nigg, & Dyer, 1993; Jones & Pulos, 1993; Pelham et al., 1993). Furthermore, studies were also placed in this category if it was not clear that both indices were used. A unique variation of the Jacobson cutoff was offered by Foa et al. (1991) in a study that used measures that lacked extensive normative data. These authors considered clinically significant change to have occurred when follow up scores exceeded 2 standard deviations below the mean of the pretreatment sample (rather than mean values based on norms). Such an approach to clinical significance represents how Jacobson's method can be altered while still providing valuable information.

Kendall and Grove's method. Nineteen of the 74 total studies (26%) used some variation of the Kendall and Grove (1988) method of determining clinically meaningful change through normative comparisons. Clinical significance has also been addressed by categorizing treated individuals as either "responders" or "improvers." In this approach, treatment effects are assessed in terms of the means and standard deviations of a norma-

tive comparison group much like with the clinical criterion cutoff used by Jacobson and his colleagues. Some authors (Wilson, Becker, & Tinker, 1995) have compared pre- and posttreatment scores in terms of a *z*-distribution, while others use preexisting and arbitrarily chosen percentile cutoffs. An example of this can be seen in Webster-Stratton's (1994) study on the effects of parent training on a variety of family and child symptom measures. To be classified as a responder to treatment, the parent had to report a score on a symptom measure (e.g., BDI) within the normal range (below the 90th percentile). To be classified as a responder on the parent-child interaction and marital observational measures (neither of which had established normative data), the parent or child had to show a 30% improvement above baseline. Others studies which purport to measure clinically significant change based on normative comparisons used similar calculations (e.g., Kendall, 1994; Morin, Kowatch, Barry, & Walton, 1993; Telch, Schmidt, Jaimez, Jacquin, & Harrington, 1995).

Although not explicitly citing Kendall and Grove (1988), some authors used a similar procedure to examine the clinical significance of client's change. These authors used predetermined scores from several dependent measures to define various levels of end-state functioning (e.g., Jaycox, Foa, & Morral, 1998) or recovery. Clients whose scores fell in the predetermined range were then classified as recovered (e.g., Schulberg, Pilkonis, & Houck, 1998) or as having "high end-state functioning" (e.g., McLean, Woody, Taylor, & Koch, 1998).

Symptom Improvement

RCI only. Two of the 74 studies used only the RCI to calculate clinical significance without considering movement into the functional distribution (Lunnen & Ogles, 1998). In these studies, clients who made reliable improvement were considered meaningfully changed regardless of the severity of the posttreatment score.

Blanchard's method. A final way in which clinical significance has been addressed, particularly in the area of health psychology, involves investigating the degree of symptom reduction or improvement. Six of the 74 total studies sampled (8% overall total) used the method proposed by Blanchard and Schwarz (1988) to determine clinical significance described above. Blanchard et al. (1990) defined clinically significant change as a 50% reduction on the headache index score. A similar procedure has been followed for determining clinically meaningful change in irritable bowel syndrome (Greene & Blanchard, 1994) and recurrent abdominal pain (Sanders, Shepherd, Cleghorn, & Woolford, 1994).

Summary

As is clear from this brief review of published studies, many researchers studying the effects of psychological interventions are including the analysis of clinically meaningful change within their publications. Most studies simply report the percentage or proportion of clinically significant changers in a descriptive manner, with no formal statistical analysis (Paivio & Greenberg, 1995; Piper, Azim, McCallum, & Joyce, 1990). Some studies also include a reference to the number of individuals who deteriorate while enrolled in treatment (e.g., Lunnen & Ogles, 1998). Some authors (Blanchard et al., 1990; Greene & Blanchard, 1994; Ogles, Lambert, & Sawyer, 1995) have taken

this a step farther, however, by conducting statistical tests to ascertain differences in the proportion of clinically improved individuals across treatment groups.

Inclusion of clinical significance data provides useful information for the researcher and the clinician. The clinical significance data gives rich information about the individuals who are involved in treatment. Similarly, the clinical significance data builds upon statistical tests to provide information about within-group variation. Using methods for examining clinical significance also provides data regarding the meaningfulness of treatment outcome. The techniques described in this paper have unique features that add to many treatment intervention studies.

Several problems are also evident when reviewing studies that apply the clinical significance methodologies. Many studies do not report the parameters used to calculate the RCI or clinical cutoff. For example, the size of change needed to produce improvement is in part related to the reliability coefficient of the dependent measure. Some studies use an internal consistency estimate of reliability while others use a test-retest estimate when calculating the RCI. Similarly, some studies use sample specific parameters while others use normative samples. This variation in parameters produces heterogeneous results across studies and samples. Perhaps agreed-upon definitions of cutoff scores could be used to help improve the comparability of studies.

Of course, standardized definitions of cutoff scores for recovery or improvement depend upon sufficient data for the dependent measures. The examination of clinical significance in many studies is hampered by the use of instruments that have limited psychometric or normative data. When attempting to compare treated individuals with a "functional" group, the lack of quality normative data presents a significant challenge. Continuing efforts to further expand the data available for measures used as dependent variables is a must.

Similarly, studies with certain populations appear to be more amenable to the clinical significance methodology. For example, studies of chronic mental health disorders (e.g., schizophrenia, bipolar affective disorder) or studies of chronic health conditions (e.g., diabetes, asthma, arthritis) raise potential problems with the clinical significance methodology. Determining cutoff points and other criteria that indicate whether or not meaningful change has occurred is more difficult when addressing chronic conditions. Unnecessarily conservative definitions of improvement should not dictate the success of a treatment when the condition is long-term.

The issue of deterioration that occurs during treatment is often neglected. Although deterioration is becoming a more prominent issue (see Mohr, 1995 for a comprehensive review), many studies do not investigate this issue. With regard to studies sampled in our review, a negative change exceeding 1.96 *SED* seemed to be the most common way of operationalizing deterioration (e.g., Barkley et al., 1992; Baucom, Sayers, & Sher, 1990; Goldman & Greenberg, 1992; Ogles, Lambert, & Sawyer, 1995). However, most studies did not consider or report data concerning deterioration. One interesting approach for examining deterioration was presented by Jacobson et al. (1991), in which deterioration was defined as the absence of identifiable recovery or improvement. Recovery was indicated by a 1.96 RCI and a BDI score below 10, while improvement was indicated by 1.96 RCI with a BDI greater or equal to 10. Deterioration was described as those clients who neither recovered nor improved. In the area of health psychology, deterioration was defined as the absence of a reduction in symptom frequency (Blanchard et al., 1990) or a drop below 50% improvement on a symptom index when improvement had initially been made (Greene & Blanchard, 1994). Clearly, more can be done to identify and study the reasons for deterioration in treatment.

Overall, a large number of treatment outcome studies reviewed in a 9-year period have reported some kind of data regarding clinical significance. Several different methods are used and there is a fair amount of variability in the way in which the results are reported (at least in terms of the nonexhaustive review provided here). That is, what constitutes clinically meaningful change may be slightly different depending on the psychometric properties of the outcome measures and the methods used by the investigators. Nevertheless, in any given year, between 3 and 14 studies (in the *Journal of Consulting and Clinical Psychology*) have cited and used the methods of Jacobson and Truax (1991), Kendall and Grove (1988), Blanchard and Schwarz (1988), or others. Thus, for many authors, sole reliance on statistical significance continues to be an unsatisfactory way of assessing the meaningfulness of change.

CONCLUSIONS

As investigators return to the roots of psychotherapy research, methods for investigating the clinical meaning of changes are becoming a standard addition to the typical therapy outcome study. Within this review, a history of the concept of clinical significance was presented along with a description of current methods for examining clinical significance. Finally, the application of these methods in the current treatment outcome literature was examined.

With this background several conclusions are offered regarding the state of clinical significance methods.

1. Several methods for examining clinical significance are already used in psychotherapy outcome studies and investigators appear to be increasingly interested in using these methods. Indeed, the need for the examination of clinical significance was advocated for studies submitted to the *Journal of Consulting and Clinical Psychology* (Kendall, 1997). Current studies of therapy effectiveness may need to examine both statistical and clinical significance in order to be sufficiently thorough in presentation.
2. The most common application of clinical significance assessment methods follows the standard examination of statistical significance. Investigators use the methods as a way to demonstrate that statistically meaningful findings are also clinically relevant.
3. Three methods for examining clinical significance appear to be the most commonly cited and used: (1) some variation of the Jacobson and Truax (1991) approach, (2) normative comparison using the Kendall and Grove (1988) approach, and (3) the symptom reduction index proposed by Blanchard and Schwarz (1988). Each method produces a final count of the number of individuals who can be identified as changing in a clinically significant manner.
4. Great variation exists in the application of the methods for examining clinical significance (Jacobson et al., 1999). Even when investigators use the same method or similar populations of clients, the parameters selected for use in formulae differ significantly. This lack of uniformity diminishes the utility of the process and raises serious concerns regarding the potential misinterpretation of results, especially when comparing studies.
5. Several thorny statistical issues have been raised regarding methods for calculating clinical significance (e.g., regression to the mean, false positives as a result of base rates, variations in formulae, identification and selection of cutoff points,

selection of normative samples, selection of appropriate parameters). Continued research must sort out the details of these interesting and complex issues. Recommendations for future directions in this journey are described below.

FUTURE DIRECTIONS

Given the current status of the clinical significance literature, what additions and expansions of the literature are warranted? Several potential avenues of study are suggested here.

1. Continued study will help to clarify the validity of clinical significance definitions and help to standardize the more promising methods for examining clinical significance. For example, a programmatic focus on reliable change indices may help to identify a standard definition for the size of change necessary to be identified as improved. Perhaps a 1-standard-deviation change on a symptom measure may be both reliable and noticeable. A series of studies will help to identify systematic factors that influence differences among measures and methods.
2. Additional data (e.g., normative and other comparison samples, reliability estimates across samples, etc.) regarding many of the instruments used as dependent variables will be especially useful for investigators who examine clinical significance. Similarly, agreement upon the parameters to use when investigating clinical significance will help to make more informed cross study comparisons.
3. Investigating the validity of the various methodologies for identifying "recovered," "improved," or "changed" individuals will also be a continuing challenge. Clearly, the entire body of literature regarding the identification and use of cutoff scores becomes relevant to this issue. Both the statistical issues relevant to classification and the theoretical issues relevant to the construct validity of the classifications need further study.
4. Within disorders that are more chronic the search for less conservative definitions of clinical significance may be especially fruitful. Especially as psychosocial interventions become more prevalent within health care settings, the identification of clinically meaningful change must address the issues involved with chronic conditions. Similarly, definitions of clinically meaningful change when studying psychosocial interventions with individuals who have chronic and severe mental disorders must be developed. The recent emphasis on quality of life measure within these areas may be an especially appealing line of continued work.
5. Expanding the notions of improvement and recovery to consider deterioration will also be important. Many studies ignore the fact that some individuals receiving treatment get worse. In one study (Lunnen & Ogles, 1998), therapists and clients could not differentiate between individuals who did not change and those who deteriorated. Further study of the factors that lead to deterioration in treatment and avenues for successful identification warrant further study.

Overall, the study of clinical significance will remain a promising field for investigation in the years to come. Especially in this age of accountability, the ability of behavioral health professionals to demonstrate that their interventions are not just statistically satisfactory will increase. The establishment of clinical relevance will help to verify that psychosocial interventions are meaningful to clients, therapists, and society.

Acknowledgment—The authors would like to thank Frank Keefe for comments on an earlier draft of this paper.

REFERENCES

Aldrich, C. K. (1975). The long and short of psychotherapy. *Psychiatric Annals, 5*, 52–58.

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

Barkham, M., Rees, A., Stiles, W. B., Shapiro, D. A., Hardy, G. E., & Reynolds, S. (1996). Dose effect relations in time limited psychotherapy for depression. *Journal of Consulting and Clinical Psychology, 64*, 927–935.

Barkley, R. A., Gueyremont, D. C., Anastopoulus, A. D., & Fletcher, K. E. (1992). A comparison of three family therapy programs for treating family conflicts in adolescents with attention deficit hyperactivity disorder. *Journal of Consulting and Clinical Psychology, 60*, 450–462.

Barlow, D. H. (1981). On the relation of clinical research to clinical practice: Current issues, new directions. *Journal of Consulting and Clinical Psychology, 49*, 147–155.

Barlow, D. H., Hayes, S. C., & Nelson, R. O. (1984). *The scientist-practioner: Research and accountability in clinical and educational settings*. New York: Pergamon.

Barrett, P. M., Dadds, M. R., & Rapee, R. M. (1996). Family treatment of childhood anxiety: A controlled trial. *Journal of Consulting and Clinical Psychology, 64*, 333–342.

Baucom, D. H, Sayers, S. L., & Sher, T. G. (1990). Supplementing behavioral marital therapy with cognitive restructuring and emotional expressiveness training: An outcome investigation. *Journal of Consulting and Clinical Psychology, 58*, 636–645.

Beck, A. T., Steer, R. A., & Garbin, M. G. (1988). Psychometric properties of the Beck depression inventory: Twenty-five years of evaluation. *Clinical Psychology Review, 8*, 77–100.

Beck, J. G., Stanley, M. A., Baldwin, L. E., Deagle, E. A., & Averill, P. M. (1994). Comparison of cognitive therapy and relaxation training for panic disorder. *Journal of Consulting and Clinical Psychology, 62*, 818–826.

Bergin, A. E. (1966). Some implications of psychotherapy research for therapeutic practice. *Journal of Abnormal Psychology, 71*, 235–246.

Bergin, A. E., & Lambert, M. J. (1978). The evaluation of therapeutic outcomes. In: S. L. Garfield & A. E. Bergin, (Eds.). *The handbook of psychotherapy and behavior change (2nd ed.)*. New York: John Wiley.

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

Bergin, A. E., & Strupp, H. H. (1972). *Changing frontiers in the science of psychotherapy*. Chicago: Aldine-Atherton.

Bigelow, D. A., McFarland, B. H., & Olson, M. (1991). Quality of life of community mental health program clients: Validating a measure. *Community Mental Health Journal, 27*, 43–55.

Blanchard, E. B., Appelbaum, K. A., Radnitz, C. L., Michultka, D., Morrill, B., Kirsch, C., Hillhouse, J., Evans, D. D., Guarnieri, P., Attanasio, V., Andrasik, F., Jaccard, J., & Dentinger, M. P. (1990). Placebo controlled evaluation of abbreviated progressive muscle relaxation and of relaxation combined with cognitive therapy in the treatment of tension headache. *Journal of Consulting and Clinical Psychology, 58*, 210–215.

Blanchard, E. B., & Schwarz, S. P. (1988). Clinically significant changes in behavioral medicine. *Behavioral Assessment, 10*, 171–188.

Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review, 48*, 378–399.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155–159.

Davis, R., Olmsted, M., & Rockert, K. (1990). Brief group psychoeducation for bulimia nervosa: Assessing the clinical significance of change. *Journal of Consulting and Clinical Psychology, 58*, 882–885.

Derogatis, L. R. (1983). *SCL-90: Administration, Scoring, and Procedures Manual for the Revised Version*. Baltimore: Clinical Psychometric Research.

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

Elkin, I., Shea, M. T., Watkins, J. T., Imber, S. D., Sotsky, S. M., Collins, J. F., Glass, D. R., Pilkonis, P. A., Leber, W. R., Docherty, J. P., Fiester, S. J., Parloff, M. B. (1989). National Institute of Mental Health treatment of depression collaborative research program: General effectiveness of treatments. *Archives of General Psychiatry, 46*, 971–982.

Foa, E. B., Ruthbaum, B., Riggs, D., & Murdock, T. (1991). Treatment of posttraumatic stress disorder in rape victims: A comparison between cognitive behavioral procedures and counseling. *Journal of Consulting and Clinical Psychology, 59*, 715–723.

Foster, S. L., & Mash, E. (1999). Assessing social validity in clinical treatment research: Issues and procedures. *Journal of Consulting & Clinical Psychology, 67*, 308–319.

Frisch, M. B. (1994). *Manual and treatment guide for the Quality of Life Inventory*. Minneapolis, MN: NCS Assessments.

Gallagher-Thompson, D., & Steffen, A. M. (1994). Comparative effects of cognitive behavioral and brief psychodynamic psychotherapies for depressed family caregivers. *Journal of Consulting and Clinical Psychology, 62*, 543–549.

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 Psychology, 67*, 320–331.

Goldman, A., & Greenberg, L. (1992). Comparison of integrated systemic and emotionally focused approaches to couples therapy. *Journal of Consulting and Clinical Psychology, 60*, 962–969.

Greenberg, L. S., & Pinsof, W. M. (Eds.). (1986). *The psychotherapeutic process: A research handbook*. New York: Guilford.

Greene, B., & Blanchard, E. B. (1994). Cognitive therapy for irritable bowel syndrome. *Journal of Consulting and Clinical Psychology, 62*, 576–582.

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.

Grundy, E. M. (1994). Assessing clinical significance: Application to the Child Behavior Checklist (Doctoral dissertation, Brigham Young University, 1994). *Dissertation Abstracts International, 55*, 593.

Hamilton, M. (1967). Development of a rating scale for primary depressive illness. *British Journal of Social and Clinical Psychology, 6*, 278–296.

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

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

Hugdahl, K., & Ost, L. (1981). On the difference between statistical and clinical significance. *Behavioral Assessment, 3*, 289–295.

Imber, S. D., Pilkonis, P. A., Sotsky, S. M., Elkin, I., Watkins, J. T., Collins, J. F., Shea, M. T., Leber, W. R., & Glass, D. R. (1990). Mode-specific effects among three treaments for depression. *Journal of Consulting and Clinical Psychology, 58*, 352–359.

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

Jacobson, N. S., Dobson, K., Fruzzetti, A. E., Schmaling, K. B., & Salusky, S. (1991). Marital therapy as a treatment for depression. *Journal of Consulting and Clinical Psychology, 59*, 547–557.

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

Jacobson, N. S., Follette, W. C., & Revenstorf, D. (1986). Toward a standard definition of clinically significant change. *Behavior Therapy, 17*, 308–311.

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

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

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

Jamison, C., & Scogin, F. (1995). The outcome of cognitive bibliotherapy with depressed adults. *Journal of Consulting and Clinical Psychology, 63*, 644–650.

Jaycox, L. H., Foa, E. B., & Morral, A. R. (1998). Influence of emotional engagement and habituation on exposure therapy for PTSD. *Journal of Consulting & Clinical Psychology, 66*, 185–192.

Jones, E. E., Ghannam, J., Nigg, J. T., & Dyer, J. P. (1993). A paradigm for single case research: The time series study of a long term psychotherapy for depression. *Journal of Consulting and Clinical Psychology, 61*, 381–394.

Jones, E. E., & Pulos, S. M. (1993). Comparing the process in psychodynamic and cognitive behavioral therapies. *Journal of Consulting and Clinical Psychology, 61*, 306–316.

Kaplan, R. M. (1990). Behavior as the central outcome in health care. *American Psychologist, 45*, 1211–1220.

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. (1994). Treating anxiety disorders in children: Results of a randomized clinical trial. *Journal of Consulting and Clinical Psychology, 62*, 100–110.

Kendall, P. C. (1997). Editorial. *Journal of Consulting and Clinical Psychology, 65*, 3–5.

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

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., & Bergin, A. E. (1994). The effectiveness of psychotherapy. In: A. E. Bergin and S. L. Garfield (Eds.). *The handbook of psychotherapy and behavior change (4th ed.)*. New York: John Wiley.

Lambert, M. J., Masters, K. S., & Ogles, B. M. (1991). Outcome research in counseling. In C. E. Watkins & L. J. Schneider (Eds.), *Research in counseling* (pp. 51–83). Hillsdale, NJ: Lawrence Erlbaum.

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. M. (1998). A multi-perspective, multi-variable evaluation of reliable change. *Journal of Consulting and Clinical Psychology, 66*, 400–410.

McLean, P. D., & Hakstian, A. R. (1990). Relative endurance of unipolar depression treatment effects: Longitudinal follow-up. *Journal of Consulting and Clinical Psychology, 58*, 482–488.

McLean, P. D., Woody, S., Taylor, S., & Koch, W. J. (1998). Comorbid panic disorder and major depression: Implications for cognitive-behavioral therapy. *Journal of Consulting & Clinical Psychology, 66*, 240–247.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting & Clinical Psychology, 46*, 806–835.

Meyers, D. G. (1990). *Social Psychology (3rd ed.)*. New York: McGraw-Hill.

Michelson, L., Mavissakalian, M., & Marchione, K. (1985). Cognitive and behavioral treatments for agoraphobia: Clinical, behavioral, and psychophysiological outcomes. *Journal of Consulting and Clinical Psychology, 53*, 913–925.

Mohr, D. C. (1995). Negative outcome in psychotherapy: A critical review. *Clinical Psychology-Science & Practice, 2*, 1–27.

Morin, C. M., Kowatch, R. A., Barry, T., & Walton, E. (1993). Cognitive behavior therapy for late life insomnia. *Journal of Consulting and Clinical Psychology, 61*, 137–146.

Muran, J. C., Gorman, B. S., Safran, J. D., Twining, L., Samstang, L. W., & Winston, A. (1995). Linking in session change to overall outcome in short term cognitive therapy. *Journal of Consulting and Clinical Psychology, 63*, 651–657.

Nichols, D. S., Greene, R. L., Schmolck, P. (1989). Criteria for assessing inconsistent patterns of item endorsement on the MMPI: Rationale, development, and empirical trials. *Journal of Clinical Psychology, 45*, 239–250.

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

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

Ogles, B. M., Lambert, M. J., & Sawyer, J. D. (1995). The clinical significance of the National Institute of Mental Health collaborative depression study data. *Journal of Consulting and Clinical Psychology, 63*, 321–326.

Paivio, S., & Greenberg, L. (1995). Resolving "unfinished business": Efficacy of experiential therapy using empty chair dialogue. *Journal of Consulting and Clinical Psychology, 63*, 419–425.

Patterson, G. R. (1974). Interventions for boys with conduct problems: multiple settings, treatments, and criteria. *Journal of Consulting & Clinical Psychology, 42*, 471–481.

Pelham, W. E., Carlson, C., Sams, S. E., Vallano, G., Dixon, J., & Hoza, B. (1993). Seperate and combined effects of methylphenidate and behavior modification on boys with attention deficit disorder in the classroom. *Journal of Consulting and Clinical Psychology, 61*, 506–515.

Persons, J. B. (1991). Psychotherapy outcome studies do not accurately represent current models of psychotherapy: A proposed remedy. *American Psychologist, 46*, 99–106.

Piper, W. E., Azim, H. F., McCallum, M., & Joyce, A. S. (1990). Patient suitability and outcome in short term individual psychotherapy. *Journal of Consulting and Clinical Psychology, 58*, 475–481.

Propst, L. R., Ostrom, R., Watkins, P., Dean, T., & Mashburn, D. (1992). Comparative efficacy of religious and nonreligious cognitive behavioral therapy for the treatment of clinical depression in religious individuals. *Journal of Consulting and Clinical Psychology, 60*, 94–103.

Sanders, M. R., Shepherd, R. W., Cleghorn, G., & Woolford, H. (1994). The treatment of recurrent abdominal pain in children: A controlled comparison of cognitive behavioral family intervention and standard pediatric care. *Journal of Consulting and Clinical Psychology, 62*, 306–314.

Saunders, S. M., Howard, K. I., & Newman, F. L. (1988). Evaluating the clinical significance of treatment effects: Norms and normality. *Behavioral Assessment, 10*, 207–218.

Schulberg, H. C., Pilkonis, P. A., & Houck, P. (1998). The severity of major depression and choice of treatment in primary care practice. *Journal of Consulting & Clinical Psychology, 66*, 932–938.

Shadish, W. R., Montgomery, L. M., Wilson, P., Bright, I., & Okwumabua, T. (1993). Effects of family and marital psychotherapies. *Journal of Consulting and Clinical Psychology, 61*, 992–1002.

Shefler, G., Dasberg, H., & Ben-Shakhar, G. (1995). A randomized controlled outcome and follow-up study of Mann's time limited psychotherapy. *Journal of Consulting and Clinical Psychology, 63*, 585–593.

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

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

Speer, D. C., & 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 & Clinical Psychology, 63*, 1044–1048.

Stahl, J. R., Thompson, L. E., Leitenberg, H., & Hasazi, J. E. (1974). Establishment of praise as a conditioned reinforcer in socially unresponsive psychiatric patients. *Journal of Abnormal Psychology, 83*, 488–496.

Stevens, S. S. (1968). Measurement, statistics, and the schemapiric view. *Science, 161*, 849–856.

Strupp, H. H., & Hadley, S. W. (1977). A tripartite model of mental health and therapeutic outcome: With special reference to negative effects in psychotherapy. *American Psychologist, 32*, 187–196.

Telch, M. J., Schmidt, N. B., Jaimez, T. L., Jacquin, K. M., & Harrington, P. J. (1995). Impact of cognitive behavioral therapy on quality of life in panic disorder patients. *Journal of Consulting and Clinical Psychology, 63*, 823–830.

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

Weber, E. H. (1978). *The sense of touch*. New York: Academic Press.

Webster-Stratton, C. (1994). Advancing videotape parent training: A comparison study. *Journal of Consulting and Clinical Psychology, 62*, 583–593.

Werner, J. S., Minkin, N., Minkin, B. L., Fixsen, D. L., Phillips, E. L., & Wolf, M. M. (1975). "Intervention package": An analysis to prepare juvenile delinquents for encounters with police officers. *Criminal Justice & Behavior, 2*, 55–84.

Wilson, S., Becker, L., & Tinker, R. (1995). Eye movement desensitization and reprocessing (EMDR) treatment for psychologically traumatized individuals. *Journal of Consulting and Clinical Psychology, 63*, 928–937.

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.

Zanna, M. P., & Olson, J. M. (1982). Individual differences in attitudinal relations. In: M. P. Zanna, E. T. Higgens, and C. P. Herman (Eds.), *Consistency in social behavior: The Ontario symposium, Vol. 2*. Hillsdale, NJ: Erlbaum.