Journal of Consulting and Clinical Psychology
1991, Vol 59, No 1, 12-19

Clinical Significance: A Statistical Approach to Defining Meaningful Change in Psychotherapy Research

Neil S. Jacobson and Paula Truax
University of Washington

In 1984, Jacobson, Follette, and Revenstorf defined clinically significant change as the extent to which therapy moves someone outside the range of the dysfunctional population or within the range of the functional population. In the present article, ways of operationalizing this definition are described, and examples are used to show how clients can be categorized on the basis of this definition. A reliable change index (RC) is also proposed to determine whether the magnitude of change for a given client is statistically reliable. The inclusion of the RC leads to a twofold criterion for clinically significant change.

There has been growing recognition that traditional methods used to evaluate treatment efficacy are problematic (Barlow, 1981; Garfield, 1981; Jacobson, Follette, & Revenstorf, 1984; Kazdin, 1977; Kendall & Norton-Ford, 1982; Smith, Glass, & Miller, 1980; Yeaton & Sechrest, 1981). Treatment effects are typically inferred on the basis of statistical comparisons between mean changes resulting from the treatments under study. This use of statistical significance tests to evaluate treatment efficacy is limited in at least two respects. First, the tests provide no information on the variability of response to treatment within the sample; yet information regarding within-treatment variability of outcome is of the utmost importance to clinicians.

Second, whether a treatment effect exists in the statistical sense has little to do with the clinical significance of the effect. Statistical effects refer to real differences as opposed to ones that are illusory, questionable, or unreliable. To the extent that a treatment effect exists, we can be confident that the obtained differences in the performance of the treatments are not simply chance findings. However, the existence of a treatment effect has no bearing on its size, importance, or clinical significance. Questions regarding the *efficacy* of psychotherapy refer to the benefits derived from it, its potency, its impact on clients, or its ability to make a difference in peoples' lives. Conventional statistical comparisons between groups tell us very little about the efficacy of psychotherapy.

The effect size statistic used in meta-analysis seems at first glance to be an improvement over standard inferential statistics, inasmuch as, unlike standard significance tests, the effect size statistic does reflect the size of the effect. Unfortunately, the effect size statistic is subject to the same limitations as those outlined above and has been even more widely misinterpreted than standard statistical significance tests. The size of an effect is relatively independent of its clinical significance. For example, if a treatment for obesity results in a mean weight loss of 2 lb and if subjects in a control group average zero weight loss, the effect size could be quite large if variability within the groups were low. Yet the large effect size would not render the results any less trivial from a clinical standpoint. Although large effect sizes are more likely to be clinically significant than small ones, even large effect sizes are not necessarily clinically significant.

The confusion between statistical effect or effect size and efficacy is reflected in the conclusions drawn by Smith et al., (1980) on the basis of their meta-analysis of the psychotherapy outcome literature. In their meta-analysis, they found moderate effect sizes when comparing psychotherapy with no or minimal treatment; moreover, the direction of their effect sizes clearly indicated that psychotherapy outperformed minimal or no treatment. On the basis of the moderate effect sizes, the authors concluded that "Psychotherapy is *beneficial,* [italics added] consistently so and in many different ways. . . . The evidence overwhelmingly supports the *efficacy* [italics added] of psychotherapy" (p. 184).

Such conclusions are simply not warranted on the basis of either the existence or the size of statistical effects. In contrast to criteria based on statistical significance, judgments regarding clinical significance are based on external standards provided by interested parties in the community. Consumers, clinicians, and researchers all expect psychotherapy to accomplish particular goals, and it is the extent to which psychotherapy succeeds in accomplishing these goals that determines whether or not it is effective or beneficial. The clinical significance of a treatment refers to its ability to meet standards of efficacy set by consumers, clinicians, and researchers. While there is little consensus in the field regarding what these standards should be, various criteria have been suggested: a high percentage of clients improving; a level of change that is recognizable by peers and significant others (Kazdin, 1977; Wolf, 1978); an elimination of the presenting problem (Kazdin & Wilson, 1978); normative levels of functioning by the end of therapy (Kendall & Norton-Ford, 1982; Nietzel & Trull, 1988); high end-state functioning by the end of therapy (Mavissakalian, 1986); or changes that significantly reduce one's risk for various health problems.

Elsewhere we have proposed some methods for defining clin-

Preparation of this article was supported by Grants MH 33838-10 and MH-44063 from the National Institute of Mental Health, awarded to Neil S. Jacobson.

Correspondence concerning this article should be addressed to Neil S. Jacobson, Department of Psychology NI-25, University of Washington, Seattle, Washington 98195.

ically significant change in psychotherapy research (Jacobson, Follette, & Revenstorf, 1984, 1986; Jacobson & Revenstorf, 1988). These methods had three purposes: to establish a convention for defining clinically significant change that could be applied, at least in theory, to any clinical disorder; to define clinical significance in a way that was consistent with both lay and professional expectations regarding psychotherapy outcome; and to provide a precise method for classifying clients as "changed" or "unchanged" on the basis of clinical significance criteria. The remainder of this article describes the classification procedures, illustrates their use with a sample of data from a previous clinical trial (Jacobson et al., 1989), discusses and provides tentative resolutions to some dilemmas inherent in the use of these procedures, and concludes by placing our method within a broader context.

A Statistical Approach to Clinical Significance

Explanation of the Approach

Jacobson, Follette, and Revenstorf (1984) began with the assumption that clinically significant change had something to do with the return to normal functioning. That is, consumers, clinicians, and researchers often expect psychotherapy to do away with the problem that clients bring into therapy. One way of conceptualizing this process is to view clients entering therapy as part of a dysfunctional population and those departing from therapy as no longer belonging to that population. There are three ways that this process might be operationalized:

(*a*) The level of functioning subsequent to therapy should fall outside the range of the dysfunctional population, where range is defined as extending to two standard deviations beyond (in the direction of functionality) the mean for that population.

(*b*) The level of functioning subsequent to therapy should fall within the range of the functional or normal population, where range is defined as within two standard deviations of the mean of that population.

(*c*) The level of functioning subsequent to therapy places that client closer to the mean of the functional population than it does to the mean of the dysfunctional population.

This third definition of clinically significant change is the least arbitrary. It is based on the relative likelihood of a particular score ending up in dysfunctional versus functional population distributions. Clinically significant change would be inferred in the event that a posttreatment score falls within (closer to the mean of) the functional population on the variable of interest. When the score satisfies this criterion, it is statistically more likely to be drawn from the functional than from the dysfunctional population.

Let us first consider some hypothetical data to illustrate the use of these definitions. Table 1 presents means and standard deviations for hypothetical functional and dysfunctional populations. The variances of the two populations are equal in this data set. Assuming normal distributions, the point that lies half-way between the two means would simply be

$$c = (60 + 40)/2 = 50$$

where c is the cutoff point for clinically significant change. The cutoff point is the point that the subject has to cross at the time of the posttreatment assessment in order to be classified as changed to a clinically significant degree. The relationship between cutoff point c and the two distributions is depicted in Figure 1. If the variances of the functional and dysfunctional populations are unequal, it is possible to solve for c, because

$$(c - M_1)/s_1 = (M_0 - c)/s_0;$$

or

$$c = \frac{s_0 M_1 + s_1 M_0}{s_0 + s_1}.$$

Because the cutoff point is based on information from both functional and dysfunctional populations and because it allows precise determination of which population a subject's score belongs in, it is often preferable to compute a cutoff point based only on one distribution or the other.

Unfortunately, in order to solve for c, data from a normative sample are required on the variable of interest, and such norms are lacking for many measures used in psychotherapy research. When normative data on the variable of interest are unavailable, the cutoff point can be estimated using the two standard deviation solution (*a*) suggested above as an alternative option. But because the two standard deviation solution does not take well-functioning people into account, it will not provide as accurate an estimate of how close subjects are to their well-functioning peers as would a cutoff point that takes into account both distributions. When the two distributions are overlapping as in the hypothetical data set, the two standard deviation solution will be quite conservative. As Figure 1 indicates, the cutoff point established by the two standard deviation solution is more stringent than c:

$$a = M_1 + 2s_1 = 40 + 15 = 55.$$

When functional and dysfunctional solutions are nonoverlapping, a will not be conservative enough. Not only are norms on functional populations desirable, but ideally norms would also be available for the dysfunctional population. As others have noted (Hollon & Flick, 1988; Wampold & Jensen, 1986), if each study uses its own dysfunctional sample to calculate a or c, then each study will have different cutoff points. The results would then not be comparable across studies. For example, the more severely dysfunctional the sample relative to the dysfunctional population as a whole, the easier it will be to "recover" when the cutoff point is study specific.

A third possible method for calculating the cutoff point is to adopt the second method mentioned above, and use cutoff point b, which indicates two standard deviations from the mean of the functional population. As Figure 1 shows, with our hypothetical data set the cutoff point would then be

$$b = M_0 - 2s_1 = 60 - 15 = 45.$$

When functional and dysfunctional distributions are highly overlapping, as in our hypothetical data set, b is a relatively lenient cutoff point relative to a and c (see Figure 1). On the other hand, if distributions are nonoverlapping, b could turn out to be quite stringent. Indeed, in the case of nonoverlapping distributions, only b would ensure that crossing the cutoff point could be translated as "entering the functional population." Another potential virtue of b is that the cutoff point would not vary depending on the nature of a particular dysfunctional sample:

Table 1
Hypothetical Data From an Imaginary Measure Used To Assess Change in a Psychotherapy Outcome Study

Symbol	Definition	Value
M_1	Mean of pretest experimental and pretest control groups	40
M_2	Mean of experimental treatment group at posttest	50
M_0	Mean of well functioning normal population	60
s_1, s_0	Standard deviation of control group, normal population, and pretreatment experimental group	7.5
s_2	Standard deviation of experimental group at posttest	10
r_{xx}	Test–retest reliability of this measure	.80
x_1	Pretest score of hypothetical subject	32.5
x_2	Posttest score of hypothetical subject	47.5

Once norms were available, they could be applied to any and all clinical trials, thus ensuring standard criteria for clinically significant change.

Which criteria are the best? That depends on one's standards. On the basis of our current experience using these methods, we have come to some tentative conclusions. First, when norms are available, either *b* or *c* is often preferable to *a* as a cutoff point: In choosing between *b* and *c*, when functional and dysfunctional populations overlap, *c* is preferable to *b*; but when the distributions are nonoverlapping, *b* is the cutoff point of choice. When norms are not available, *a* is the only cutoff point available: To avoid the problem of different cutoff points from study to study, *a* should be standardized by aggregating samples from study to study so that dysfunctional norms can be established. An example is provided by Jacobson, Wilson, and Tupper (1988), who reanalyzed outcome data from agoraphobia clinical trials and aggregated data across studies using the Fear Questionnaire to arrive at a common cutoff point that could be applied to any study using this questionnaire.

A Reliable Change Index

Thus far we have confined our discussion of clinically significant change to the question of where the subject ends up following a regimen of therapy. In addition to defining clinically significant change according to the status of the subject subsequent to therapy, it is important to know *how much* change has occurred during the course of therapy. When functional and dysfunctional distributions are nonoverlapping, this additional information is superfluous, because by definition anyone who has crossed the cutoff point would have changed a great deal during the course of therapy. But when distributions do overlap, it is possible for posttest scores to cross the cutoff point yet not be statistically reliable. To guard against these possibilities, Jacobson et al. (1984) proposed a reliable change index (RC), which was later amended by Christensen and Mendoza (1986):

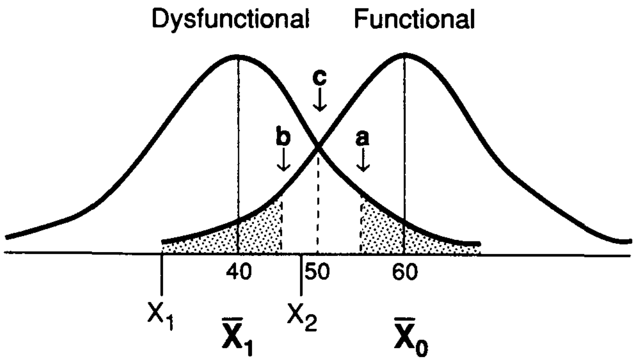

Figure 1. Pretest and posttest scores for a hypothetical subject (x) with reference to three suggested cutoff points for clinically significant change (a, b, c).

$$RC = \frac{x_2 - x_1}{S_{diff}}$$

where x_1 represents a subject's pretest score, x_2 represents that same subject's posttest score, and S_{diff} is the standard error of difference between the two test scores. S_{diff} can be computed directly from the standard error of measurement S_E according to this:

$$S_{diff} = \sqrt{2(S_E)^2}\,.$$

S_{diff} describes the spread of the distribution of change scores that would be expected if no actual change had occurred. An RC larger than 1.96 would be unlikely to occur ($p < .05$) without actual change. On the basis of data from Table 1,

$$S_E = s_1\sqrt{1 - r_{xx}} = 7.5\sqrt{1 - .80} = 3.35$$

$$S_{diff} = \sqrt{2(3.35)^2} = 4.74$$

$$RC = 47.5 - 32.5/4.74 = 3.16.$$

Thus, our hypothetical subject has changed. RC has a clearcut criterion for improvement that is psychometrically sound. When RC is greater than 1.96, it is unlikely that the posttest score is not reflecting real change. RC tells us whether change reflects more than the fluctuations of an imprecise measuring instrument.

An Example Using a Real Data Set

To illustrate the use of our methods with an actual data set, we have chosen a study in which two versions of behavioral marital therapy were compared: a research-based structured version and a clinically flexible version (Jacobson et al., 1989). The purpose of this study was to examine the generalizability of the marital therapy treatment used in our research to a situation that better approximated an actual clinical setting. However, for illustrative purposes, we have combined that data from the two treatment conditions into one data set. Table 2 shows the pretest and posttest scores of all couples on two primary outcome measures, the Dyadic Adjustment Scale (DAS; Spanier, 1976) and the global distress scale of the Marital Satisfaction Inventory (GDS; Snyder, 1979), and a composite measure, which will be explained below. Data from the DAS only are also depicted in Figure 2. Points falling above the diagonal represent improvement, points right on the diagonal indicate no change, and points below the line indicate deterioration. Points falling outside the shaded area around the diagonal represent changes that are statistically reliable on the basis of RC ($> 1.96S_{diff}$); above the shaded area is "improvement" and below is "deterioration." One can see those subjects, falling within the shaded area, who showed improvement that was not reliable and could have constituted false positives or false negatives were it not for RC. Finally, the broken line shows the cutoff point separating distressed (D) from nondistressed (ND) couples. Points above the dotted line represent couples who were within the functional range of marital satisfaction subsequent to therapy. Subjects whose scores fall above the dotted line and outside the shaded area represent those who recovered during the course of therapy.

To understand how individual couples were classified, let us first consider Figure 3. Figure 3 depicts approximations of the distributions of dysfunctional (on the basis of this sample) and functional (on the basis of Spanier's norms) populations for the DAS. Using cutoff point criteria *c*, the point halfway between dysfunctional and functional means is 96.5. This is almost exactly the cutoff point that is found using Spanier's norms for functional (married) and dysfunctional (divorced) populations (cf. Jacobson, Follette, Revenstorf, Baucom, Hahlweg, & Margolin, 1984). If norms had not been available and we had to calculate a cutoff point based on the dysfunctional sample alone using the two standard deviation solution, the cutoff point would be 105.2. Finally, *b*, the cutoff point that signifies entry into the functional population, is equal to 79.4.

Given that the dysfunctional and functional distributions overlap, we have already argued that *c* is the preferred criteria. Indeed, a convention has developed within the marital therapy field to use 97 as a cutoff point, which is virtually equivalent to *c*. However, there is a complication with this particular measure, which has led us to rethink our recommendations. The norms on the DAS consist of a representative sample of married people, without regard to level of marital satisfaction. This means that a certain percentage of the sample is clinically distressed. The inclusion of such subjects in the normative sample shifts the distribution in the direction of dysfunctionality and creates an insufficiently stringent *c*. If all dysfunctional people had been removed from this married sample, the distribution would have been harder to enter, and a smaller percentage of couples would be classified as recovered. An ideal normative sample would exclude members of a clinical population. Such subjects are more properly viewed as members of the dysfunctional population and therefore distort the nature of the normative sample. Given the problems with this normative sample, it seemed to us that *a* was the best cutoff point for clinically significant change. At least when *a* is crossed we can be confident that subjects are no longer part of the maritally distressed population, whereas the same cannot be said of *c*, given the failure to exclude dysfunctional couples in the normative sample.

Table 2 also shows how subjects were classified on the basis of RC. Some couples showed improvement but not enough to be classified as recovered, whereas others met criteria for both improvement and recovery. In point of contrast, Table 2 depicts pretest and posttest data for a second measure of marital satisfaction, the Global Distress Scale (GDS) of the Marital Satisfaction Inventory (Snyder, 1979). Subjects were also classified as improved (on the basis of RC) or recovered (on the basis of a cutoff point) on this measure. Figure 4 shows approximations of the dysfunctional and functional populations. If we consider the three possible cutoff points for clinically significant change, criterion *c* seems preferable given the rationale stated earlier for choosing among the three. The distributions do overlap, and if *c* is crossed, a subject is more likely to be a member of the functional than the dysfunctional distribution of couples. The criteria for recovery on the GDS listed in Table 2 are based on the use of *c* as a cutoff point.

Table 3 summarizes the data from both the DAS and the GDS, indicating the percentage of couples who improved and recovered according to each measure. Not surprisingly, there was less than perfect correspondence between the two measures. It is unclear how to assimilate these discrepancies. Moreover, some subjects were recovered on one measure but not on the other, thus creating interpretive problems regarding the status of individual subjects.

Given that both the DAS and GDS measure the same construct, one solution to integrating the findings would be to derive a composite score. These two measures of global marital satisfaction can each be theoretically divided into components of true score and error variance. However, it is unlikely that either duplicates the true score component of the construct "marital satisfaction." To preserve the true score component of each measure, a composite could be constructed that retained the true score component. Jacobson and Revenstorf (1988) have suggested estimating the true score for any given subject (j), using test theory, by adopting the formula

$$T_j = Rel(X_j) + (1 - Rel)M$$

where T represents true score, *Rel* equals reliability (e.g., test–retest), and X is the observed score (Lord & Novick, 1968). The standardized true score estimates can then be averaged to derive a multivariate composite. Cutoff points can then be established.

Tables 2 and 3 depict results derived from this composite. Because no norms are available on the composite, the cutoff point was established using the two standard deviation solution.[1]

Finally, let us use this data set to illustrate one additional

[1] The proportion of recovered couples is greater in the composite than it is for the component measures for several reasons. First, there are four couples for whom GDS data are missing. In all four instances, the couples failed to recover. Composites could be computed only on the 26 cases for whom we had complete data. Second, in several instances couples were subthreshold on one or both component measures but reached criteria for recovery on the composite measure. It is of interest that in this important sense the composite measure was more sensitive to treatment effects than either component was.

Table 2
Individual Couple Scores and Change Status on Dyadic Adjustment Scale, Global Distress Scale, and Composite Measures

Subject	Pretest	Posttest	Improved but not recovered	Recovered
Dyadic Adjustment Scale				
1	90.5	97.0	N	N
2	74.0	124.0	N	Y
3	97.0	97.5	N	N
4	73.5	88.0	Y	N
5	61.0	96.5	Y	N
6	66.5	62.5	N	N
7	68.5	112.5	N	Y
8	86.5	103.5	Y	N
9	88.5	90.0	N	N
10	68.5	82.5	Y	N
11	98.0	105.0	N	N
12	80.5	99.5	Y	N
13	89.5	112.5	N	Y
14	91.5	101.0	N	N
15	83.5	99.5	Y	N
16	60.5	79.5	Y	N
17	83.0	88.0	N	N
18	88.0	100.5	Y	N
19	98.5	119.0	N	Y
20	78.5	116.0	N	Y
21	99.5	116.0	N	Y
22	79.5	129.0	N	Y
23	84.5	113.0	N	Y
24	92.5	118.0	N	Y
25	93.0	92.0	N	N
26	85.0	114.0	N	Y
27	64.0	68.0	N	N
28	61.0	52.0	N	N
29	80.0	60.5	N	N
30	82.5	104.5	Y	N
Global Distress Scale				
1	68.0	62.5	N	N
2	74.5	56.0	N	Y
3	58.5	58.0	N	N
4	73.5	71.0	N	N
5	78.5	60.5	Y	N
6	76.0	77.0	N	N
7	76.5	58.5	N	Y
8	63.0	52.0	N	Y
9	70.0	65.5	N	N
10	75.0	73.0	N	N
11	63.5	64.0	N	N
12	73.5	55.5	N	Y
13	71.5	53.0	N	Y
14	63.5	55.0	N	Y
15	57.0	50.0	N	N
Global Distress Scale (*continued*)				
16	75.0	78.0	N	N
17	63.0	65.5	N	N
18	75.0	62.0	Y	N
19	71.5	60.5	Y	N
20	68.0	51.0	N	Y
21	75.5	50.0	N	Y
22	67.5	44.0	N	Y
23	62.5	55.5	N	N
24	69.5	56.0	N	Y
25	61.0	60.5	N	N
26	67.0	47.5	N	Y
27	75.5	—	—	—
28	75.5	—	—	—
29	69.5	—	—	—
30	66.5	—	—	—
Composite				
1	64.8	57.9	N	N
2	75.9	43.0	N	Y
3	58.5	55.9	N	N
4	74.7	65.4	Y	N
5	82.4	57.3	N	Y
6	78.9	79.4	N	N
7	78.2	49.2	N	Y
8	64.6	50.7	Y	N
9	66.5	62.3	N	N
10	77.6	68.7	Y	N
11	59.6	54.8	N	N
12	71.6	53.9	N	Y
13	66.7	47.0	N	Y
14	62.6	53.0	Y	N
15	63.6	51.7	Y	N
16	81.3	72.0	Y	N
17	66.2	63.2	N	N
18	68.7	56.1	Y	N
19	62.6	47.1	N	Y
20	70.3	44.6	N	Y
21	63.7	44.2	N	Y
22	69.6	35.7	N	Y
23	65.3	47.8	N	Y
24	65.5	45.7	N	Y
25	60.9	59.4	N	N
26	66.9	43.9	N	Y
27	—	—	—	—
28	—	—	—	—
29	—	—	—	—
30	—	—	—	—

Note. Composite = Average of Dyadic Adjustment Scale and Global Distress Scale estimated true scores. Y = yes; N = no. Dash = information not available.

problem with these statistical definitions of clinically significant change. We have been using a discrete cutoff point to separate dysfunctional from functional distributions, without taking into account the measurement error inherent in the use of such cutoff points. Depending on the reliability of the measure, all posttest scores will be somewhat imprecise due to the limitations of the measuring instrument. Thus, some subjects are going to be misclassified simply due to measurement error.

One solution to the problem involves forming confidence intervals around the cutoff point, using RC to derive the boundaries of the confidence intervals. RC defines the range in which an individual score is likely to fluctuate because of the imprecision of a measuring instrument. Figure 5 illustrates the use of RC to form confidence intervals. The confidence intervals form a band of uncertainty around the cutoff point depicted in Figure 5. On the basis of this data set, for the DAS a score can

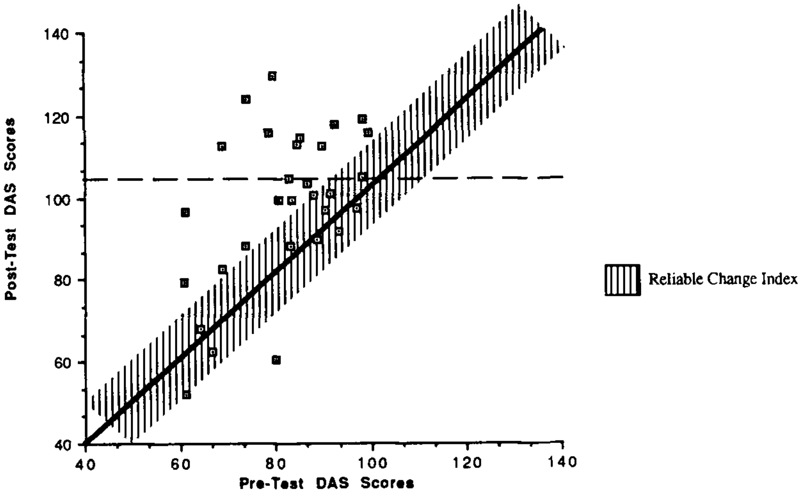

Figure 2. Scatter plot of pretest and posttest scores on the Dyadic Adjustment scale with jagged band showing reliable change index.

vary by as much as 9.83 points and still reasonably ($p > .05$) be considered within the bounds of measurement error. Variations of 10 points or more are unlikely to be explainable by measurement error alone. We then formed confidence intervals around the cutoff point, with the cutoff point serving as the midpoint of the interval. Approximately 10 points on either side of the interval are within the band of uncertainty, but beyond this band we felt confident that the cutoff point had truly been crossed.[2]

As Figure 5 shows, 14 subjects fell within the band of uncertainty created by these confidence intervals. Should these couples be classified as improved, recovered, deteriorated, or uncertain? One possibility would be to add a new category to the classification system: the proportion of subjects who fell within this band of uncertainty. These were couples about whose status we were unsure. If we added this category to our classification system, the revised percentages would be 20% recovered, 47% unclassifiable, and 33% unchanged or deteriorated. Having identified the proportion of subjects about whom we were uncertain, we could use the remainder of the sample and exclude the uncertain subjects in our calculations of proportions of recovered and unrecovered couples. This exclusion would lead to figures of 38% recovered, 19% improved but not recovered, and 44% unchanged or deteriorated. These proportions are probably a more accurate reflection of the true proportion of recovered subjects, inasmuch as subjects within the band of uncertainty are, on the average, going to be equally likely to fall into both categories. In fact, as Jacobson and Revenstorf (1988) noted, this latter suggestion is almost like splitting the difference (i.e., dividing the uncertain subjects equally between recovered and improved but not recovered groups). Although splitting the difference would not reduce ambiguity regarding the status of individual subjects who fall within the band of uncertainty, it would lead to a summary statistic that would include the entire sample. Essentially, such a solution amounts to redistributing subjects within the band rather than ignoring it entirely. When equal numbers of subjects fall on either side of the cutoff point within the band, the proportion of recovered subjects will be identical to that calculated without consideration of measurement error at all. Splitting the difference with our sample data set would have resulted in 43% recovered, 23% improved but not recovered, and 34% unchanged or deteriorated.

Conclusion

In the past decade, the discussion of clinical significance has taken center stage in psychotherapy research. In a recent review appearing in the *Annual Review of Psychology*, Goldfried, Greenberg, and Marmar (1990) referred to it as one of the major methodological advances. There is no doubt that discussion has moved from occasional mention by a group of prescient observers (e.g., Barlow, 1981; Kazdin, 1977) to a lively topic for discussion and debate, as evidenced by the recent special issue of *Behavioral Assessment* devoted to the topic (Jacobson, 1988).

The editors of this special section have asked us to compare the results of using our system with what would have been obtained using standard inferential statistics or other criteria of improvement. When our statistics have been used, the impact has generally been to add additional information rather than to contradict the results of other data analytic strategies. However, the information from these additional analyses has generally led to more modest conclusions regarding the efficacy of the treatment in question. For example, Jacobson and colleagues (Jacobson, Follette, Revenstorf, Baucom, et al., 1984) reanalyzed data from previously published marital therapy outcome studies. Standard inferential statistical analyses yielded results that supported the effects of the marital therapies, in that treatments outperformed various control groups. The reanalyses reported by Jacobson and colleagues addressed the issue of clinical significance, and the results were somewhat disappointing: Fewer than half of the treated couples ended up in the happily married range after therapy on measures of marital satisfac-

[2] As Jacobson and Revenstorf (1988) noted, the use of confidence intervals in this manner is a convenience. The cutoff point is merely a theoretical point, and the actual measurement error logically pertains to individual subject's score.

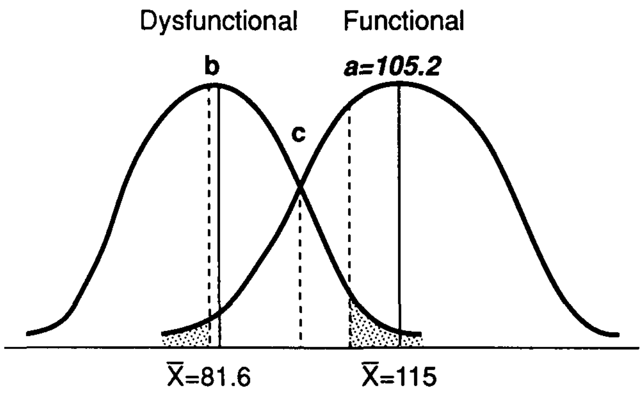

Figure 3. Approximations to the dysfunctional and functional distributions on the Dyadic Adjustment Scale with reference to three suggested cutoff points for clinically significant change (a, b, c).

Table 3
Percentages of Improved and Recovered Couples on DAS, GDS, and Composite Scores

Measure	*N*	% improved	% recovered	% unimproved or deteriorated
DAS	30	30	33	37
GDS	26	12	42	46
Composite	26	27	46	27

Note. DAS = Dyadic Adjustment Scale, GDS = Global Distress Scale of the Marital Satisfaction Inventory, and composite-average of DAS and GDS estimate true scores.

tion. Similar reanalyses based on studies looking at exposure treatments for agoraphobia led to similar results (i.e., treatments outperformed control groups but yielded a relatively small proportion of truly recovered clients; Jacobson, Wilson, & Tupper, 1988).

Experimenters who have used different statistical procedures based on similar principles have often found that clinical significance data make the treatments look less effective than standard statistical comparisons would imply. For example, Kazdin, Bass, Siegel, and Thomas (1989) recently reported on an apparently highly effective behavioral treatment for conduct-disordered children, but a clinical significance analysis suggested that celebration was perhaps premature. Whereas behavioral treatments outperformed a client-centered relationship therapy, comparisons with nonclinic samples revealed that the majority of subjects remained in the dysfunctional range on primary measures of conduct disorder. Similarly, Robinson, Berman, and Neimeyer (1990) recently reported a meta-analysis of studies investigating psychotherapy for depression. Whereas they reported substantial effect sizes for comparisons between psychotherapy and control groups, comparisons with normative samples suggested that subjects remained outside the normal range even after psychotherapeutic intervention.

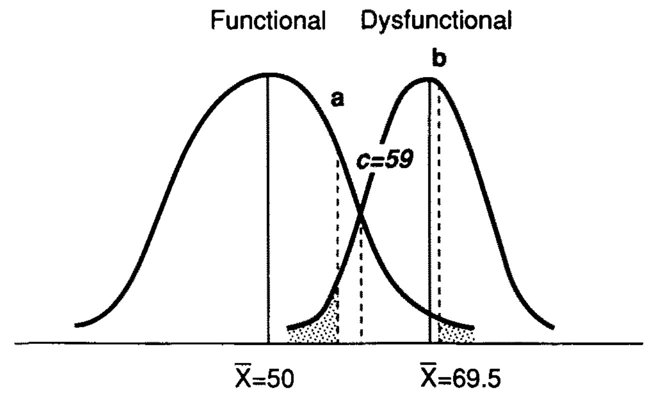

Figure 4. The same approximations of dysfunctional and functional distributions for Global Distress Scale of the Marital Satisfaction Inventory with reference to three suggested cutoff points for clinically significant change (a, b, c).

The approach we have outlined is only one of many possible ways of reporting on clinical significance. On the one hand, our approach has a number of features that we believe should be part of any method for highlighting clinical significance: It operationalizes recovery in a relatively objective, unbiased way; its definition is not tied to a specific disorder, which means that it has potentially broad applicability; because of its general applicability, it could evolve into a convention within psychotherapy research, which in turn would facilitate comparison between studies; and it provides information on variability in outcome as well as clinical significance.

On the other hand, there are a number of unsolved problems that currently limit the generalizability of the method. First, it is unclear at present how robust the method will be to violations of the assumption that dysfunctional and functional distributions are normal. The concept that we have proposed for defining clinical significance does not depend on any formula. The formula is simply one way of determining the midpoint between functional and dysfunctional populations. Even when the formulas for RC and the cutoff points are not applicable, the concept can be applied by determining the cutoff point empirically. However, the formulas discussed in this and other articles assume normal distributions. Second, operationalizing clinical significance in terms of recovery or return to normal functioning may not be appropriate for all disorders treated by psychotherapy. For example, schizophrenia and autism are two disorders in which a standard of recovery would exceed the expectations of most who work in the field. Third, without psychometrically sound measures of psychotherapy outcomes, there are practical constraints that prevent optimal use of our methods, no matter how valuable they might be in theory. In particular, the absence of normative data for functional and dysfunctional populations on many commonly used outcome measures deters the development of standardized cutoff points.

In addition to these and other current problems, there are still a number of subjective decisions to be made regarding optimal use of these statistical methods. These were illustrated in our examples. Only by testing theoretical propositions with real data sets will these ambiguities be resolved. Thus, while it is not premature to expect psychotherapy investigators to report on the clinical significance of their treatment effects, it is far too early to advocate any particular method or set of conventions. Clinical significance has clearly arrived, but the optimal methods for deriving it remain to be determined.

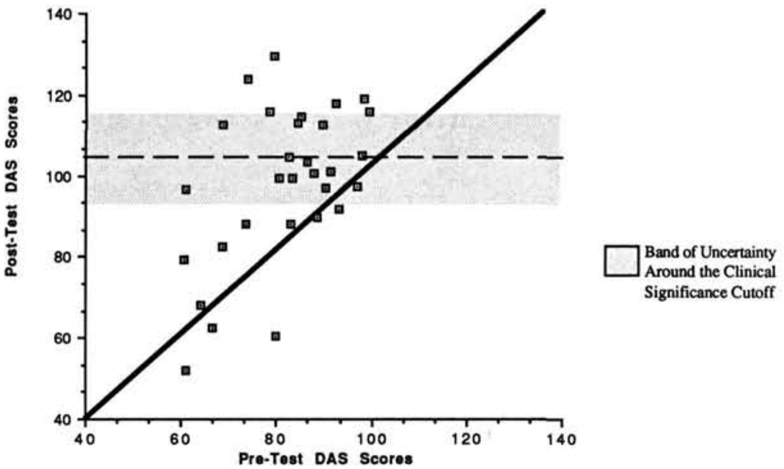

Figure 5. Scatter plot of pretest and posttest scores on the Dyadic Adjustment Scale including band of uncertainty around cutoff point for clinically significant change.

References

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.

Christensen, L., & Mendoza, J. L. (1986). A method of assessing change in a single subject: An alteration of the RC index. *Behavior Therapy, 17,* 305–308.

Garfield, S. L. (1981). Evaluating the psychotherapies. *Behavior Therapy, 12,* 295–307.

Goldfried, M. R., Greenberg, L. S., & Marmar, C. (1990). Individual psychotherapy: Process and outcome. *Annual Review of Psychology, 41,* 659–688.

Hollon, S. D., & Flick, S. N. (1988). On the meaning and methods of clinical significance. *Behavioral Assessment, 10,* 197–206.

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

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., 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., & 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., Schmaling, K. B., Holtzworth-Munroe, A., Katt, J. L., Wood, L. F., & Follette, V. M. (1989). Research-structured versus clinically flexible versions of social learning-based marital therapy. *Behaviour Research and Therapy, 27,* 173–180.

Jacobson, N. S., Wilson, L., & Tupper, C. (1988). The clinical significance of treatment gains resulting from exposure-based interventions for agoraphobia: A reanalysis of outcome data. *Behavior Therapy, 19,* 539–552.

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

Kazdin, A. E., Bass, D., Siegel, T., & Thomas, C. (1989). Cognitive–behavioral therapy and relationship therapy in the treatment of children referred for antisocial behavior. *Journal of Consulting and Clinical Psychology, 57,* 522–535.

Kazdin, A. E., & Wilson, G. T. (1978). *Evaluation of behavior therapy: Issues, evidence, and research strategies.* Cambridge, MA: Ballinger.

Kendall, P. C., & Norton-Ford, J. D. (1982). Therapy outcome research methods. In P. C. Kendall & J. N. Butcher (Eds.), *Handbook of research methods in clinical psychology* (pp. 429–460). New York: Wiley.

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

Mavissakalian, M. (1986). Clinically significant improvement in agoraphobia research. *Behaviour Research and Therapy, 24,* 369–370.

Nietzel, M. T., & Trull, T. J. (1988). Meta-analytic approaches to social comparisons: A method for measuring clinical significance. *Behavioral Assessment, 10,* 146–159.

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

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

Snyder, D. K. (1979). Multidimensional assessment of marital satisfaction. *Journal of Marriage and the Family, 41,* 813–823.

Spanier, G. B. (1976). Measuring dyadic adjustment: New scales for assessing the quality of marriage and similar dyads. *Journal of Marriage and the Family, 38,* 15–28.

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

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.

Yeaton, W. H., & Sechrest, L. (1981). Critical dimensions in the choice and maintenance of successful treatments: Strength, integrity, and effectiveness. *Journal of Consulting and Clinical Psychology, 49,* 156–167.

Received May 16, 1990
Revision received August 4, 1990
Accepted September 7, 1990 ■