Psychotherapy outcome research: Methods for reporting variability and evaluating clinical significance

Article *in* Behavior Therapy · September 1984

DOI: 10.1016/S0005-7894(84)80002-7

William C. Follette
University of Nevada, Reno
105 PUBLICATIONS **3,572** CITATIONS
SEE PROFILE

Dirk Revenstorf
University of Tuebingen
87 PUBLICATIONS **1,769** CITATIONS
SEE PROFILE

BEHAVIOR THERAPY 15, 336–352 (1984)

Psychotherapy Outcome Research: Methods for Reporting Variability and Evaluating Clinical Significance

NEIL S. JACOBSON

WILLIAM C. FOLLETTE

University of Washington

DIRK REVENSTORF

University of Tubingen

The purpose of this article is to suggest some new directions for the presentation and reporting of data in psychotherapy outcome research. Statistical comparisons based on group means provide no information on the variability of treatment outcome, and statistical significance tests do not address clinical significance. Although psychotherapy research has begun to address these issues, it has done so unsystematically. New standards and conventions are needed to serve as criteria for classifying therapy subjects into categories of improved, unimproved, and deteriorated based on response to treatment. A two-fold criterion for determining improvement in a client is recommended, based on both statistical reliability and clinical significance. Statistical procedures for determining whether or not these criteria have been met are discussed.

There seems to be a growing recognition among those who engage in psychotherapy outcome research that many of their standard conventions have helped to perpetuate a gulf between the scientists and practitioners within the field of clinical psychology (Barlow, 1980, 1981; Garfield, 1981; Gottman & Markman, 1978; Hugdahl & Ost, 1981; Jacobson, Follette, & Elwood, 1984; Kazdin & Wilson, 1978; Smith, Glass, & Miller, 1980). Practitioners have long decried the irrelevance of most psychotherapy research for clinical practice (Bergin & Strupp, 1972). More recently, researchers have themselves discussed, debated, and lamented this ap-

Preparation of this manuscript was supported by Grant #5 R01 MH33838-04 from the National Institute of Mental Health, awarded to the senior author. Special thanks to John Gottman, Steve Hollon, Gene Glass, and Bob Bolles for valuable feedback on an initial draft. Comments from anonymous reviewers were also quite helpful. Requests for reprints should be sent to Neil S. Jacobson, Department of Psychology NI-25, University of Washington, Seattle, WA 98195.

0005-7894/84/0336–0352$1.00/0

parent failure of the Boulder model (Hayes, 1981; Hersen & Barlow, 1976; Kazdin, 1977; Leitenberg, 1974; Meehl, 1978; Peterson, 1976; Raush, 1974; Shakow, 1976). As Barlow (1981) put it, "At present, clinical research has little or no influence on clinical practice" (p. 147). While it might be tempting to blame this problem on clinicians for "failing to keep up with the literature," researchers have begun to consider the possibility that the literature itself is to blame (Barlow, 1980, 1981; Hugdahl & Ost, 1981; Raush, 1974).

Two areas which have been subjected to much criticism are the exclusive reliance on group means and statistical significance tests in evaluating treatment effects, and the corresponding deemphasis on data which highlight variability and clinical significance. In this article, our primary purpose is to suggest some alternatives to these conventions for the presentation and reporting of data. As an introduction, a brief statement of the limitations of current methods is provided. This is followed by a justification of new conventions, given these limitations. Following a description of our proposals, we conclude by underscoring our tentativeness regarding these suggestions, and challenging clinical researchers to begin providing information about the effects of psychotherapy which is both clinically meaningful and statistically reliable.

Problems With Current Methods of Reporting Data on Psychotherapy Outcome

The vast majority of published studies evaluating psychotherapy outcome have inferred treatment efficacy on the basis of statistical comparisons between two or more treatment conditions (Kazdin & Wilson, 1978; Meltzoff & Kornreich, 1970; Smith et al., 1980). These comparisons have two conventional properties which limit their utility in psychotherapy outcome research. The first is that they are based on the average improvement score for all subjects and thus provide no information on the effects of therapy for individual clients in that sample (Barlow, 1980, 1981; Garfield, 1981; Hugdahl & Ost, 1981; Kazdin, 1977). The second is that the "significance test" itself imposes a criterion for determining a treatment effect which often has little clinical relevance.

Without some information regarding variability of outcome, the reader has no way of determining the proportion of clients who benefited from the treatment. Yet these proportions are of great importance to anyone interested in estimating the likelihood that a given individual will benefit from therapy. Many have argued, in fact, that descriptive statistics such as the proportion of clients who improve are at least as important as group means in conveying essential information about the effects of psychotherapy (Barlow, 1980; Garfield, 1981; Hugdahl & Ost, 1981; Kazdin & Wilson, 1978).

Whether an experimenter chooses to report summary statistics for the entire group or the proportion of clients who respond to treatment in a particular way, criteria for demonstrating an "effect" need to be established. The convention throughout the behavioral sciences is to use statistical significance tests as the basis for inferring differences between groups. The problem with applying statistical significance tests to psychotherapy research is that the former has little to do with the practical importance of the effect. This limitation is not unique to psychotherapy research (cf. Bakan, 1966; Bolles, 1962; Carver, 1978), but it is a particularly acute problem in applied areas.

Many have begun to advocate adding "clinical significance" as a criterion for evaluating psychotherapy. However, there is little consensus as to what clinical significance is, except the universal agreement that it is *not* merely statistical significance. Clinically significant change has been defined as a large proportion of the clients improving (Hugdahl & Ost, 1981), a change which is large in magnitude (Barlow, 1981), an improvement in the client's everyday functioning (Kazdin & Wilson, 1978), a change which is recognizable to peers and significant others (Kazdin, 1977; Wolf, 1978), an elimination of the presenting problem (Kazdin & Wilson, 1978), and the attainment of a level of functioning which is no longer distinguishable from the client's nondeviant peers (Kazdin & Wilson, 1978; Kendall & Norton-Ford, 1982). All of these criteria have something to do with the notion that the change is of practical importance. None of these criteria is attained by a statistical significance test alone.

The Need for Additional Conventions

A variety of solutions have been proposed to remedy these problems (Barlow, 1981; Bergin & Strupp, 1972; Garfield, 1981; Gottman & Markman, 1978; Hayes, 1981; Hersen & Barlow, 1976; Hugdahl & Ost, 1981; Kazdin, 1977; Kendall & Norton-Ford, 1982; Wolf, 1978; Yeaton & Sechrest, 1981). For example, investigators have begun to report the proportion of clients who improve in response to treatment. However, as yet, no conventions have been developed to define the conditions under which an individual is to be classified as improved, and, as a result, decision rules have varied considerably from one study to another. For example, the agoraphobia treatment outcome literature has frequently included reports of the proportion of clients who improve (Jansson & Ost, 1982), and, in some cases, this literature includes even finer distinctions such as markedly vs. moderately improved (e.g., Hand, Lamontagne, & Marks, 1974). Typically, however, these proportions are not featured prominently in the presentation of results, and the criteria for classifying subjects vary from study to study. The range of criteria is at times unspecified (Barlow & Mavissakalian, 1981); at other times qualitative judgments are made by the principal investigator and based on a priori criteria (e.g., "resumed regular entry into most of the phobic situations") (Hand et al., 1974). Some studies define improvement as a 50% reduction of anxiety/avoidance (Jansson & Ost, 1982); others use a change of 2 or more points on an 8-point scale of anxiety/avoidance (Emmelkamp & Kuipers, 1979). None of these studies attempts to justify the chosen criteria. Despite their arguable face validity, the criteria chosen are either arbitrary (Emmelkamp & Kuipers, 1979; Jansson & Ost, 1982) or highly subjective (Barlow & Mavissakalian, 1981; Hand et al., 1974).

We believe that the field needs agreed-upon conventions for determining what constitutes improvement in an individual, and, if possible, general consensus as to what is meant by clinical significance. These conventions should be applicable to a wide variety of clinical problems, and they should be objective, relatively free of bias, and sound from both a psychometric and a clinical perspective. Without such conventions these data will remain an afterthought, a post hoc exploration acting as a sideshow to the main event which continues to be the group significance test. Conventions protect the field from capitalization on chance, and they guard against the tendency for an investigator to use self-serving criteria. Moreover, currently utilized criteria vary so widely from one study to another that there is little basis for comparison. It seems clear that if reports highlighting variability and clinical significance are to emerge as primary data for evaluating the effects of psychotherapy, the field needs an agreed-upon basis for deciding whether a client warrants being classified as "improved" and when that improvement is clinically significant.

Clinical Significance

We have provided a number of definitions of clinical significance used in the literature. Thus far, the concept remains poorly defined and highly subjective. Operationalizing it to a point where it is standardized across clinical problems will be very difficult, if not impossible. However, we believe that some guidelines can be proposed which will be widely, if not universally, applicable.

First, although clinical significance has something to do with the magnitude of change, there are limitations to pure magnitude criteria. For one thing, they require a decision rule as to how much change is necessary before it will be called clinically significant. Without reference to some standard—such as normal functioning—these decision rules are likely to be subjective and arbitrary. Once such a standard is imposed, we are no longer dealing with pure magnitude measures. Furthermore, pretreatment level of functioning is not taken into account in pure magnitude measures, and as a result for many clinical problems, their predictive validity would be negligible. For example, both Baucom and Mehlman (1984) and Revenstorf, Hahlweg, Schindler, and Kunert (1984) found that the magnitude of change on a measure of marital satisfaction did not predict marital happiness at a 2-year follow-up, whereas posttest levels of satisfaction did. For many clinical problems, the level attained by the end of therapy is considerably more predictive of long-term functioning than the magnitude of change.

Second, clinical significance must refer to a range of possible outcomes rather than simply a dichotomous categorization (Kendall & Norton-Ford, 1982; Patterson, 1974). Successful resolution of the presenting problem has been proposed as a criterion (Kazdin & Wilson, 1978). This definition is conceptually sound but not practical, because it forces the clinical researcher to think in terms of false dichotomies (e.g., one either has a problem or one does not). On what basis would we infer that an agoraphobia has been "eliminated"? How about depression? Although for some clinical problems it may be reasonable to use terms like elimination (e.g., some addictive behaviors), most clinical problems vary somewhere on a continuum of functioning at both pretest and posttest, and therefore, clinically significant changes will involve a range of scores.

Therefore, we propose that a change in therapy is clinically significant when the client moves from the dysfunctional to the functional range during the course of therapy on whatever variable is being used to measure the clinical problem (Kazdin, 1977; Kazdin & Wilson, 1978; Kendall & Norton-Ford, 1982). That is, assuming that both functional and dysfunctional people form distinct distributions on the target problem, clinically significant change can be thought of as movement on the part of an individual from the dysfunctional to the functional distribution. This criterion is composed of two subquestions: Is the level of functioning at posttest sufficiently high that the subject is no longer a member of the dysfunctional population? Is the client now functioning at a level which places him/her within the range of the normal or functional population?

These questions could be operationalized in any of the following ways:

1. Does the level of functioning at posttest fall outside the range of the *dysfunctional* population, where range is defined as extending to two standard deviations above (in the direction of functionality) the mean for that population?

2. Does the level of functioning at posttest fall within the range of the *functional* or normal population, where range is defined as beginning at two standard deviations below the mean for the normal population?

3. Does the level of functioning at posttest suggest that the subject is statistically more likely to be in the functional than in the dysfunctional population; that is, is the posttest score statistically more likely to be drawn from the functional than the dysfunctional distribution?

Thus, we are suggesting that one of these criteria be used to determine whether *each subject* in the sample improved to a clinically significant degree. The result would be a descriptive statistic which would reflect the *proportion* of subjects whose posttest scores indicated clinically significant treatment effects.

Let us apply some hypothetical data from Table 1 to each of these criteria in order to examine its ramifications. The applications are also depicted graphically in Fig. 1. Our hypothetical subject scored 65 on our dependent measure at pretest (x_1) and improved 30 points following therapy ($x_2 = 95$). The dysfunctional and functional populations have the same variance and differ only by their mean. Assuming normal distributions, the cutoff points for clinical significance using each of the above criteria would be:

$$a = \bar{X}_1 + 2s_1 = 80 + 30 = 110$$
$$b = \bar{X}_0 - 2s_1 = 120 - 30 = 90$$
$$c = (110 + 90)/2 = 100$$

Since c is equivalent to the point where the probabilities of belonging to functional and dysfunctional populations are equal, and the variances of

TABLE 1

Hypothetical Data From an Imaginary Measure Used to Assess Change in a Psychotherapy Outcome Study

Symbol	Value
$\bar{X}_1$= mean of both pretest experimental and pretest control groups	80
$\bar{X}_2$= mean of experimental treatment group at posttest	100
$\bar{X}_0$= mean of well-functioning normal population	120
$s_1 = s_0$ = standard deviation of control group, normal population, and pretreatment experimental group	15
s_2= standard deviation of experimental group at posttest	20
$r_{xx'}$= test-retest reliability of this measure	.80
x_1= pretest score of hypothetical subject	65
x_2= posttest score of hypothetical subject	95

the two populations are equal, *c* is simply the midpoint between *a* and *b*. However, if the variances were unequal, one could conceptualize a way to solve for the cutoff *c* as:

$$(c - \bar{X}_1)/s_1 = (\bar{X}_0 - c)/s_0$$

One can then solve for *c* using the equation:

$$c = \frac{s_0\bar{X}_1 + s_1\bar{X}_0}{s_0 + s_1} = \frac{15(80) + 15(120)}{15 + 15} = \frac{1200 + 1800}{30} = 100$$

As Fig. 1 shows, criterion *a* was the most stringent, *b* was the most lenient, and *c* fell in between. Only by criterion *b* would the progress of our hypothetical subject be considered clinically significant.

Which criterion is the best? That depends. Choosing a high cutoff point such as *a* practically ensures that the client falls outside the dysfunctional

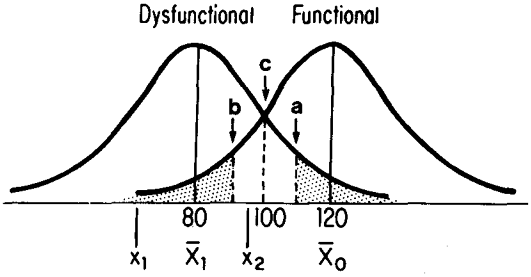

Fig. 1. Pre- and posttest scores for a hypothetical subject (*x*) with reference to three suggested cutoff points for clinically significant change (*a, b, c*).

range if that criterion is attained. A less stringent alternative might be to use one rather than two standard deviations as the cutoff point. Many would accept this more lenient definition of dysfunctional range since only 16% of the dysfunctional clients score above it. Nevertheless, subjects whose posttest scores fall in between this lenient and stringent cutoff are likely to remain somewhat dysfunctional.

The more overlap between the distributions of the functional and dysfunctional populations, the more stringent *a* will be relative to *b*. One could argue that under conditions of high overlap, there is more confusion and ambiguity regarding the determination of functionality, and therefore the more stringent criterion *a* is favored because it is more conservative. On the other hand, although one choice may be intuitively more reasonable than the other given the characteristics of the two populations, ultimately the choice between *a* and *b* is arbitrary.

The third criterion *c* is not arbitrary. It is based on the relative probability of a particular score ending up in one of the two population distributions. Clinical significance is inferred when the likelihood is greater that the score falls within the normal population. Therefore, it seems to be the best choice when dysfunctional and functional population distributions overlap. When overlap is minimal or nonexistent, *c* will be a very stringent criterion for clinical significance.

In order to calculate cutoff points for clinical significance using criterion *c*, norms must be available on the distributions of dysfunctional as well as functional people. Unfortunately, all too often norms are lacking for a sample of well-functioning people. At other times, norms exist but they are based on a sample which includes a combination of well-functioning and dysfunctional people, e.g., when the norms are based on a random sample of the population without excluding dysfunctional people. When there are no available norms, criterion *a* will be the choice for determining clinical significance by default. This criterion can always be estimated as long as pretreatment data are available on a dysfunctional sample. When the norms take the form of a random sample of the population which includes unspecified numbers of functional and dysfunctional people, the investigator has a choice between criterion *a* and a cutoff based on this normative sample. We tend to favor criterion *a* given this choice, since it is unclear where to set the cutoff point for a distribution which includes members of both populations.

We are not the first to recommend using a return to normal functioning as a standard from which clinical significance is determined (cf. Kazdin, 1977): We are simply reaffirming this standard as the best available one, and suggesting ways to use this standard to classify each individual in the sample. Kazdin (1977) has raised some important considerations in adopting normative standards: First, normative functioning, for one reason or another, may at times be an undesirable goal. Second, it is not always possible to identify an appropriate normative group. These considerations should be underscored.

Another possible concern is that for some clinical populations these criteria for clinical significance may be too stringent. For example, a great deal of excitement has been generated recently by the findings that family-oriented aftercare with schizophrenic clients can significantly reduce the rate of rehospitalization during the 1st year following discharge from a psychiatric hospital (Falloon, 1981; Goldstein, Rodnick, Evans, May, & Steinberg, 1978). Since simply remaining outside a psychiatric hospital is a considerably more modest treatment goal than entering the well-functioning population on variables relevant to the diagnosis of schizophrenia, our proposed criteria for clinical significance would exceed the expectations of currently active clinical investigators.

When the currently accepted criteria for clinically significant change are more modest than a return to normative levels of functioning, improvement rates based on these more modest criteria should certainly be reported. However, the rate of return to normative levels of functioning should be reported as well. The incorporation of standards which cut across clinical problems allows for a comparison of the potency of psychotherapy from one area to another, a comparison which is not possible when the definition of clinical significance is completely problem specific. Furthermore, these standards act as a check against the tendency to gradually modify treatment goals according to the limits of our current treatment technology. This may create the illusion of progress when in fact it reflects nothing but changing expectations. These standards help define the limits of our current therapy technology.

Statistically Reliable Improvement

When a client ends up within the range of a normative peer group after therapy, he/she has met one of two necessary conditions for being classified as "improved." The other criterion is that there must have been change during the course of therapy. It is nonsensical to speak of clinically significant treatment effects when no change has occurred, regardless of the level of posttest functioning.

How much change should there have to be for a client who ends up in the normative range of functioning to be considered "improved"? One answer might be that any amount of change would be sufficient as long as the end result meets the criteria outlined in the previous section. This answer would probably suffice when dysfunctional and functional distributions are nonoverlapping, as perhaps in the case of diagnosed schizophrenic disorders. However, when the distributions do overlap, or when the client begins at the high end of the dysfunctional distribution, it is possible for a client to change during the course of therapy, and end up with a normative level of functioning; yet the change may not be reliable in a statistical sense because it is within the margin of measurement error. For example, if the client's pretest score is in the 90th percentile of the dysfunctional population, and by posttest it has moved to the 98th percentile, the amount of change may be minimal and unreliable even though he/she ends up outside the dysfunctional distribution. We could not safely classify this client as improved despite the movement out of one distribution unless the change was of sufficient magnitude to exceed the margin of measurement error. More generally, in order for change to be considered clinically significant it must also be statistically reliable; we must be able to determine that the change is "real." This can be viewed as a purely statistical question, analogous to the correlated *t* test often used to evaluate change on a group basis. Viewed in this way, the question becomes: "For how many clients in this sample was there improvement of a sufficient magnitude to rule out chance as a plausible competing explanation?" For a subset of these clients, the "improvement" which is statistically reliable will also be clinically significant. Hence, we end up with a two-fold criterion for clinical significance.

Similar to recommendations made recently by Nunnally and Kotsche (1983), we propose a *reliable change index* (*RC*) equivalent to the difference score (post-pre) divided by the standard error of measurement:

$$RC = (x_2 - x_1)/S_E$$

where S_E = standard error of measurement; x_2 = the posttest score for a hypothetical subject, and x_1 = that subject's pretest score.

Thus a client's pretest score is subtracted from his/her posttest score, and the difference is divided by S_E. S_E describes the spread of the distribution of repeated performances that would be expected given that no actual change has occurred. An *RC* larger than ±1.96 would be unlikely to occur ($p < .05$) without actual change. Based on the data from Table 1,

$$S_E = s_1\sqrt{1 - r_{xx'}} = 15\sqrt{1 - .80} = 6.7$$

where $r_{xx'}$ = the reliability of the measure. Then:

$$RC = (95 - 65)/6.7 = 4.5$$

Thus this client has changed. *RC* is a good choice as a standardized score for the determination of reliable change. It has a clear-cut decision criterion for improvement which is psychometrically sound. If *RC* exceeds 1.96, it is unlikely that the posttest score is not reflecting real change. *RC* tells us whether the change reflects more than the fluctuations of an imprecise instrument.

The one major disadvantage of *RC* is that it is dependent on the sensitivity of the measure and can be large even if the magnitude of change is small when the measure is highly reliable. However, if used in conjunction with our criterion for clinical significance, this disadvantage is obviated. Other criteria which might be considered include the following:

1. *An index reflecting how far the client has moved during therapy relative to control group clients.* This would be an "effect size" statistic for an individual, with the change score divided by the control group standard deviation. This index would be a pure magnitude of change measure, and it would not be influenced by the effects of therapy on other clients in the treatment group. In our opinion, this index is limited because it does not lead to a decision rule for categorizing people as "improved." The index has no obvious meaning, and the choice would be as arbitrary as it would be using the raw score.

2. *An index locating the client somewhere in the distribution of change*

scores achieved by other clients given the same treatment. This index compares the amount of change achieved by a particular client to that achieved by other clients. The score can be standardized by making it sample referenced (subtracting the mean difference score). In the standardized form, it provides a possible decision rule for designating a client as improved since one could specify a percentile rank based on the client's score relative to others in the sample. The problem with the index, in our view, is that it is sample referenced and therefore changes as a result of the overall treatment effect. How others in the sample do affects the index for an individual.

3. *An index designating how different the client is from what one would predict, using linear regression, based on the pretest score.* This index is similar to one that Cronbach and Furby (1970) discussed as a possible index of individual change. Thus, the group trend toward change is used to predict each client's posttest score. The higher the score, the greater the improvement relative to other clients in the same treatment. Like the index described in the preceding paragraph, this index is a relative gain score and, as such, is limited by its insensitivity to absolute change independent of others in the sample.

4. *An index reflecting the percentage of control group subjects that the client has passed while in therapy.* This index can be derived simply by converting raw scores to z scores and then to percentile scores, based on the control group distribution, and then subtracting a client's pretest ranking from his/her posttest ranking. This index is also independent of the overall group treatment effect. Its interpretation is quite accessible. Its disadvantage is that passing 10% of the control group may be meaningless in the middle ranges of the distribution; near the mean, it takes only a z score of .25 for 10% of the untreated clients to be surpassed. At the extremes of the distribution, however, a z score might have to exceed 3.00 if 10% of the control group is to be passed. Thus, this index may be too lenient in the middle range and too conservative at the extremes.

5. *Single-subject experimental designs.* None of the indices mentioned thus far allows one to attribute change to the treatment. In other words, these criteria allow us to assert only that change has occurred, and we are left in the dark regarding the causes of that change. The establishment of a causal relationship between the treatment and outcome for an individual can only be accomplished by a single-subject experimental design (Hersen & Barlow, 1976). If every client in the sample were treated within the context of a separate within-subject experiment, a reliable change index would be unnecessary. An internally valid single-subject design kills two birds with one stone: It establishes that the change is real and that it is attributable to the experimental treatment. All that remains is to determine clinical significance, and this still requires comparison with a normative distribution. Single-subject designs do, however, obviate the need for determining that the change is statistically reliable.

Unfortunately, single-subject designs are often not practical in outpatient clinical outcome studies: At times, the target behaviors cannot be observed continuously; collecting stable baselines prior to intervention is often impossible; reversal designs are typically precluded either for ethical reasons or because the conditions are not reversible; multiresponse baselines are possible only to the extent that independent target behaviors exist; and quite often insufficient numbers of data points are available for a time series analysis (cf. Kazdin, 1979). Thus, single-subject designs, when feasible, provide viable alternatives to classifying subjects as improved, as a first step toward determining clinical significance, but they are not always applicable to the problems addressed by psychotherapy research. To the extent that innovations in the use of single-subject designs can broaden their applicability, they will continue to play an important role in bridging the gap between clinician and researcher (Barlow, Hayes, & Nelson, 1984).

To summarize, a number of statistical procedures have been considered to help determine whether or not a client has improved, as one step in the two-fold process of identifying the proportion of clients who have manifested clinically significant change. Although all indices have strenghs and weaknesses, we tentatively lean toward the reliable change index (*RC*) as the index of choice.

Conclusion: Extensions and Qualifications

Extensions

We have been advocating the reporting of data on the proportion of clients who have improved (or deteriorated) in a psychotherapy research project. Conventions have been suggested for a two-fold criterion to determine how an individual is classified: one based on whether the change is statistically reliable and the other based on whether the posttest level of functioning places that client within functional limits with respect to the clinical problem. There are many ways to extend the application of these principles even though, thus far, specific methods of reporting variability in immediate treatment effects have been emphasized.

Graphic portrayal of variability. Graphic representation is one potentially useful way to depict the variability in a sample of clients receiving psychotherapy. For example, in Fig. 2, a scatterplot is shown of pretest and posttest scores on a marital satisfaction measure for married couples receiving behavioral marital therapy; this procedure was based on the method for reporting follow-up data in weight reduction research described by Stunkard and Penick (1979). Data for the most distressed spouse in each couple is reported for each of four treatment conditions: behavior exchange (BE), problem-solving training (PS), a complete treatment (CO), and a waiting list control group (WL). Points falling above the diagonal represent improvement, points right on the diagonal indicate no change, and points below the line indicate deterioration. The figure shows no obvious group differences between the three active treatments, but considerable variability between subjects within each condition. Points falling outside the shaded area represent changes that are statistically reliable ($> 1.96\ S_E$); above the shaded area is "improvement," while

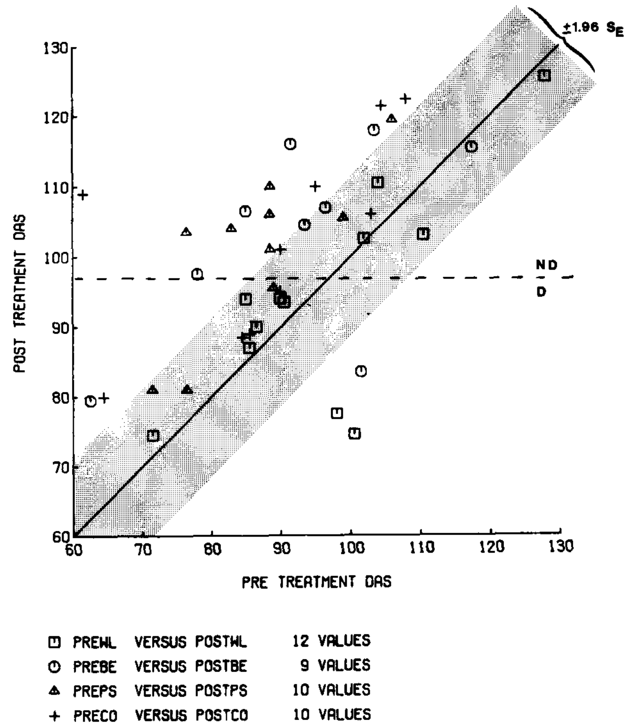

FIG. 2. Scatterplot of pretest and posttest scores on Dyadic Adjustment Scale for spouses treated with different versions of behavioral marital therapy.

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 the index. Finally, the broken line shows the cutoff for the functional range of marital satisfaction. Depending upon which lines one chooses to focus on, sections can be formed which represent various combinations of improvement, statistical significance, and clinical significance. For example, one section represents all subjects whose improvement is both statistically reliable and clinically significant. One can see at a glance how many subjects are excluded when one moves from a single to a two-fold criterion for improvement.

One final advantage of graphic depiction of the data is that aberrant subjects who may be distorting between-group comparisons can be easily spotted. These aberrant subjects may also be of special interest for investigators who want to identify anomalous subjects to develop typologies, understand treatment failures, or produce innovative clinical procedures.

Follow-up data. Follow-up data are typically reported in one of two ways: Within-group correlated *t* tests compare posttest scores with follow-up scores, and the investigator concludes that treatment effects have been maintained if means are not significantly different; or the typical between-groups comparisons are conducted at follow-up, and the results compared with those at posttest, often testing the time × treatment interaction with a repeated measures analysis of variance.

These are remarkably insensitive indices of the persistence of psychotherapy effects. Without reporting variability within the treatment groups, it is impossible to determine relapse rates, or for that matter enhancement effects. The between-group comparisons disguise information in much the same way that they do for typical statistical tests determining treatment effects at posttest.

Treatment effects can either be maintained, enhanced, or eliminated at follow-up. All three outcomes are quite common and of interest. Although psychotherapy is subject to deterioration rates during the course of active treatment, they are quite small compared to the relapse rates from posttest to follow-up. In short, there is more variability to be disguised at follow-up than at posttest, and, therefore, group means are even less satisfactory substitutes for representations of variability. For example, two groups might have the same mean change score of zero from posttest to follow-up, but in one group all subjects score zero, whereas in the other group half improved by ten points and the other half deteriorated by ten points. Although means for both groups indicate overall maintenance of treatment gains, the two groups are responding very differently; these differences will be reflected only in the discrepant standard deviation for the two groups. Without presenting the data in a way that shows, subject by subject, what happened from posttest to follow-up, it is impossible to characterize what the long-term effects of therapy are.

By applying the criteria discussed in previous sections, it is possible to overcome the limitations of these group data.

First, data on the relative proportion of clients whose treatment effects are enhanced, maintained, or reduced can be reported, using criteria that are both statistically reliable and clinically meaningful. If a change in either direction is sufficiently large to suggest an extremely low probability ($p < .05$) that it is a function of measurement error, the subject can be classified as reliably different than he/she was at posttest. The *RC* index will be identical to what it was for posttest calculations, since its parameters are limited to pretreatment variability and reliability. Thus, any subject who improved in excess of 1.96 S_E is said to have continued to improve, whereas a subject whose follow-up score is lower than the posttest score by more than 1.96 S_E can be said to have deteriorated. The clinical significance of the change must be judged by the same criterion that was used from pre- to posttest: Subjects who move from the dysfunctional to the functional range—or vice versa—from posttest to follow-up have changed in status, provided that these changes are also statistically reliable. Using the equal likelihood criterion *c* and considering the data from Table 1, where 100 is the cutoff for normative functioning, clinically significant change will not be inferred at follow-up unless that cutoff is crossed and the amount of movement exceeds 1.96 S_E = 1.96(6.7) = 13.13. Thus, if our hypothetical subject moved from 95 at posttest to 105 at follow-up, we would have insufficient information from which to conclude a clinically significant enhancement of treatment gains, despite passing the cutoff. This two-fold criterion results in a conservative procedure for inferring clinically significant change during the follow-up period. Similarly, a subject may show considerable evidence of continued improvement during the follow-up period—for example, by improving from an 80 at posttest to a 95 at follow-up—and nevertheless not be viewed as improved from the standpoint of clinical significance.

Qualifications and Conclusions

We have suggested that proportions of improved clients be included routinely as part of the report in a psychotherapy outcome study. We have further suggested that standardized criteria be adopted for determining whether the improvement manifested by an individual subject is clinically significant. Let us conclude by underscoring what we believe to be the limits of these proposals.

First, the methods that we have proposed are only as good as the outcome measures available to the field. There is a continuing and serious need for the development of better indices of therapeutic change. The fact that adequate norms do not exist for many widely used measures has already been mentioned as a problem in applying our formulae. Beyond the issue of normative data, however, is the more general question of a measurement's utility: Clinically significant change on a poor measure does not improve our ability to meaningfully interpret the findings from psychotherapy research. Moreover, our discussion does not deal with the common tendency on the part of outcome investigators to use multiple measures. It is not hard to imagine situations in which different measures would diverge in their classification of clients. How does one interpret such discrepancies? Clearly, these and a variety of other issues related to the choice of outcome measures require further discussion.

Second, the criteria for clinical significance will continue to be elusive, despite our best efforts to objectify them. As we have already noted, for some clinical problems the criterion of joining the ranks of the functional population is simply not appropriate. Even when it is appropriate, it may be too stringent. Some will argue that a client who is less depressed is better off than one who is more depressed, even though he may remain depressed: These people, according to this line of reasoning, should be considered improved to a clinically meaningful degree. We consider these arguments persuasive, even though our proposal stands.

Third, for any particular subject considered in isolation from the rest of the sample, our proposed criteria do not allow one to conclude that the treatment caused the improvement. These causal inferences still require either group comparisons or single-subject experimental designs.

Our proposals are not meant to be alternatives to these designs, but simply an unbiased way to aggregate group data which provide clinicians and researchers with more meaningful descriptions of subject samples.

Fourth, these suggestions are not meant to supersede inferential statistics in the traditional sense. Proportions of improved clients should be presented along with between-groups comparisons, and the former should be thought of as data which would help make sense out of treatment effects, or perhaps suggest why there was no effect. When treatment effects are camouflaged by high variability, proportions of improved and deteriorated clients will depict it. Unfortunately, in most such cases proportions must be interpreted cautiously in the absence of a treatment effect, since the significance tests of differential proportions between groups are less powerful than their parametric counterparts.

Fifth, our suggested criteria for pre- to posttest "deterioration" are less satisfactory than those for "improvement." There is no obvious counterpart to our distributional cutoff for clinical significance in the assessment of deterioration rates. Therefore, we limited our proposal to the single psychometric criterion (*RC*). All that can be concluded from a client meeting the *RC* index criterion for deterioration is that the change is statistically reliable. It is hoped that future investigators will suggest some additional criteria so that we can establish reasonable indices of clinically significant deterioration.

Sixth, if our proposals were adopted, we suspect that psychotherapy would look less effective than it typically does at present. Our criteria are realistic, but they are also conservative. Therefore, until the field becomes accustomed to viewing data presented in this manner, it would be a mistake to reject manuscripts for publication simply because the treatment looks ineffective when variability is rigorously reported. It may be necessary to readjust our expectations based on the new data that come in and reevaluate conclusions formed in the past based on improvement that was reported differently. The situation is roughly analogous to the dilemma of evaluating interrater reliability based on the kappa statistic. It is clear that kappa will be lower than straight percentage agreement, but it is not clear how high kappa must be before we can conclude that coders are reliable. Similarly, it is not clear how often a psychotherapy technique must succeed in order for its effects to be worth reporting.

Finally, the thesis of this article is that methods of reporting data should be changed in psychotherapy research, not necessarily that our methods be the chosen ones. It is also suggested that standard conventions must be adopted, but not necessarily ours. We have suggested a number of possible statistics, but even though we have some preferences it would be premature to state those preferences strongly. We view this article as part of an ongoing dialogue, which will hopefully stimulate further discussion and attempts on the part of others to replicate, refine, and improve on our proposals. Some experimentation is in order; the field needs to discover more creative ways of reporting data. The variability and clinical significance of psychotherapy must move from the discussion section to the method and results sections. Perhaps then clinical research will have a greater impact on clinical practice.

REFERENCES

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66,* 423–437.

Barlow, D. H. (1980). Behavior therapy: The next decade. *Behavior Therapy, 11,* 315–328.

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 practitioner: Research and accountability in clinical and educational settings.* New York: Pergamon.

Barlow, D. H., & Mavissakalian, M. (1981). Directions in the assessment and treatment of phobia: The next decade. In M. Mavissakalian & D. H. Barlow (Eds.), *Phobia: Psychological and pharmacological treatments.* New York: Guilford.

Baucom, D. H., & Mehlman, S. K. (1984). Predicting marital status following behavioral marital therapy: A comparison of models of marital relationships. In K. Hahlweg & N. S. Jacobson (Eds.), *Marital interaction: Analysis and modification* (pp. 89–104). New York: Guilford.

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

Bolles, R. C. (1962). The difference between statistical hypotheses and scientific hypotheses. *Psychological Reports, 11,* 639–645.

Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review, 3,* 378–399.

Cronbach, L. J., & Furby, L. (1970). How should we measure "change"—Or should we? *Psychological Bulletin, 74,* 68–80.

Emmelkamp, P. M. G., & Kuipers, A. C. M. (1979). Agoraphobia: A follow-up study four years after treatment. *The British Journal of Psychiatry, 134,* 352–355.

Falloon, I. R. H. (1981). Communication and problem-solving skills training with relapsing schizophrenics and their families. In M. R. Lansky (Ed.), *Family therapy and major psychopathology.* New York: Grune and Stratton.

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

Goldstein, M. J., Rodnick, E. H., Evans, J. R., May, P. R. A., & Steinberg, M. (1978). Drug and family therapy in the aftercare treatment of acute schizophrenia. *Archives of General Psychiatry, 35,* 1169–1177.

Gottman, J. M., & Markman, H. J. (1978). Experimental designs in psychotherapy research. In S. L. Garfield & A. E. Bergin (Eds.), *Handbook of psychotherapy and behavior change: An empirical analysis* (2nd ed., pp. 23–62). New York: Wiley.

Hand, I., LaMontagne, Y., & Marks, I. M. (1974). Group exposure (flooding) in vivo for agoraphobics. *The British Journal of Psychiatry, 124,* 588–602.

Hayes, S. C. (1981). Time series methodology and empirical clinical practice. *Journal of Consulting and Clinical Psychology, 49,* 193–211.

Hersen, M., & Barlow, D. H. (1976). *Single case experimental designs: Strategies for studying behavior change.* New York: Pergamon.

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

Jacobson, N. S., Follette, W. C., & Elwood, R. W. (1984). Outcome research in behavioral marital therapy: Methodological and conceptual reappraisal. In K. Hahlweg & N. S. Jacobson (Eds.), *Marital interaction: Analysis and modification* (pp. 113–129). New York: Guilford.

Jansson, L., & Ost, L. (1982). Behavioral treatments for agoraphobia: An evaluative review. *Clinical Psychology Review, 2,* 311–336.

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

Kazdin, A. E. (1979). Methodological and interpretive problems of single-case experimental designs. *Journal of Consulting and Clinical Psychology, 46,* 629–642.

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

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

Leitenberg, H. (1974). Training clinical researchers in clinical psychology. *Professional Psychology, 5,* 59–69.

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

Meltzoff, J., & Kornreich, M. (1970). *Research in psychotherapy.* New York: Atherton.

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

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

Peterson, D. R. (1976). Is psychology a profession? *American Psychologist, 31,* 572–581.

Raush, H. L. (1974). Research, practice, and accountability. *American Psychologist, 29,* 678–681.

Revenstorf, D., Hahlweg, K., Schindler, L., & Kunert, H. (1984). The use of time series analysis in marriage counseling. In K. Hahlweg & N. S. Jacobson (Eds.), *Marital interaction: Analysis and modification* (pp. 199–231). New York: Guilford.

Shakow, D. (1976). What is clinical psychology? *American Psychologist, 31,* 553–560.

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

Stunkard, A. J., & Penick, S. B. (1979). Behavior modification in the treatment of obesity: The problem of maintaining weight loss. *Archives of General Psychiatry, 36,* 801–806.

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: January 3, 1984
FINAL ACCEPTANCE: April 6, 1984