Methods for Defining and Determining the Clinical Significance of Treatment Effects
Description, Application, and Alternatives

Neil S. Jacobson
Department of Psychology University of Washington
Lisa J. Roberts
Department of Psychology University of Washington
Sara B. Berns
Department of Psychology University of Washington
Joseph B. McGlinchey
Department of Psychology University of Washington

ABSTRACT

This article summarizes and scrutinizes the growth of the development of clinically relevant and psychometrically sound approaches for determining the clinical significance of treatment effects in mental health research by tracing its evolution, by examining modifications in the method, and by discussing representative applications. Future directions for this methodology are proposed.

Preparation of this article was supported by National Institute of Mental Health Grant 5KO2-MH00868-09.

Correspondence may be addressed to Neil S. Jacobson, Department of Psychology, University of Washington, 1107 Northeast 45th Street, 310, Seattle, Washington, 98195.
Received: August 24, 1998
Revised: December 7, 1998
Accepted: December 14, 1998

Jacobson, Follette, and Revenstorf (1984) proposed one of several methods for determining the practical importance of statistical effects found in clinical trials. In that article, as well as in subsequent publications (Jacobson & Revenstorf, 1988; Jacobson & Truax, 1991), Jacobson and colleagues attempted to grapple with two limitations prevalent in statistical comparisons between groups of treated clients. First, such comparisons provide little or no information regarding the variability in treatment response from person to person. Group means, for example, do not in and of themselves indicate the proportion of participants who have improved or recovered as a result of treatment. Thus, statistical comparisons between groups shed little light on the proportion of participants in each condition who have benefited from the treatment. Second, standard statistical comparisons between groups seldom determine the practical importance of the treatment effects. Although previous investigators had tried to improve on standard statistical comparisons by reporting the size of the statistical effect (e.g., Smith, Glass, & Miller, 1980), even effect sizes do not directly speak to clinical significance. Although large effects are more likely to be clinically significant than small ones, even large effects can be clinically insignificant.
There are a multitude of ways that one can characterize variability in treatment response and at least as many ways that one can determine whether changes are clinically significant (see articles in this special section). However the concept is defined, there are a variety of ways to operationalize clinical significance in mathematical terms. It is possible to distinguish the conceptual definition from its mathematical interpretation. Each has its own areas of controversy, and some of the mathematical debate is quite esoteric. Moreover, the mathematical distinctions between the original metric (Jacobson, Follette, & Revenstorf, 1984) and more recent alternatives are all based on assumptions that cannot be tested without an empirical comparison of the methods. The ultimate question is "Which metric works best for a particular data set using a particular measure on the basis of some accepted external criterion?" Because such empirical comparisons have rarely occurred, in this article, we have chosen to concentrate on (a) defining our method, (b) examining its methodology, (c) providing examples of applications and misapplications, and (d) summarizing revised methods that retain our criteria for clinically significant change but proposing different metrics for doing so. We conclude with some recommendations for future research.

Conceptual Definition of Clinically Significant Change

Clinical significance is routinely defined as returning to normal functioning. Although for some disorders this may be too stringent a criterion, it is based on the assumption that consumers enter therapy expecting that their presenting problems will be solved. Even in cases in which this criterion is too stringent, the scientific community, as well as consumers of mental health services, still want to know how often normal functioning is attained. There is a second consideration: The magnitude of change for a given individual should be statistically reliable, that is, beyond the scope of what could reasonably be attributed to chance or measurement error. The final product is a twofold criterion for clinically significant change: (a) The magnitude has to be statistically reliable and (b) by the end of therapy, clients have to end up in a range that renders them indistinguishable from well-functioning people. If the client shows statistically reliable change but ends therapy still somewhat dysfunctional, then the client is classified as "improved but not recovered." If the client ends up in the functional range by the end of therapy, but the magnitude of change is not statistically reliable, then our method cannot determine whether or not the change is clinically significant. Finally, if the magnitude of change is statistically reliable and the client ends up within normal limits on the variable of interest, the client is said to have "recovered." This metric also allows for a determination of how often statistically significant deterioration occurs by identifying those clients who have shown a statistically reliable change in the opposite direction to that indicative of improvement.

By applying our metric to a population of treated clients, one can determine the percentage of clients who improved but did not recover, the percentage of clients who recovered, and the percentage of clients who remained unchanged or who deteriorated in each treatment condition. These descriptive percentages can be compared between groups using contingency table analyses to determine whether any observed differences between groups are statistically significant, or they can simply be used descriptively to augment the standard between-groups comparisons based on mean differences. Either way, the proportions provide valuable information on variability of outcome within each treatment condition and a way of determining the practical importance of statistically significant differences between groups.

To determine whether or not the magnitude of change was statistically reliable, we proposed a reliable change index (RCI; Jacobson, Follette, & Revenstorf, 1984). There were mathematical problems in our original formula for the RCI pointed out by Christensen and Mendoza (1986). In response to their suggestions, the RCI was modified in subsequent discussions of clinical significance (Jacobson &
Revenstorf, 1988; Jacobson & Truax, 1991). We also provided three alternatives for calculating a posttreatment cutoff score on the variable of interest, any of which can be used to categorize clients as recovered or not recovered. If the cutoff point is crossed at the time therapy ends, the client can be said to have recovered (i.e., no longer dysfunctional); if the cutoff point is not crossed, the client can be said to be "still dysfunctional," independently of whether or not the change was statistically reliable. All three cutoff points have advantages and disadvantages, but they all have the same conceptual underpinning of seeking to assess whether clients reached normal functioning as the criterion for clinically significant change.

Criticisms of the assumptions underlying the metric (e.g., Wampold & Jenson, 1986) stimulated an ongoing process of refining decision rules. As alternative formulas for defining reliable change and clinical significance cutoff points have emerged (e.g., McGlinchey & Jacobson, in press), we have examined them and taken them to heart. For example, problems have been identified (Jacobson & Revenstorf, 1988; Jacobson & Truax, 1991) with the use of a discrete cutoff point for clinically significant change. Specifically, because of measurement error, there are bound to be both false positives (those categorized falsely as recovered) and false negatives (those falsely categorized as still dysfunctional). We suggested using the logic of the RCI to establish confidence intervals around the cutoff point to identify participants who were at risk for misclassification. The conservative conclusion would be to simply acknowledge uncertainty about those participants: Because a proportion of participants could not be classified, their status was in doubt. However, we also suggested some strategies for incorporating those participants whose classification was uncertain, for example, by making the assumption that the false positives and false negatives balance out one another. This allowed for an estimation of the proportions recovered and not recovered that made use of the entire sample.

Mathematics aside, the conceptual underpinnings of the method have not changed. Returning to normal functioning is still viewed as a useful criterion for clinically significant change. The contention remains that to be sure that a client is functioning normally at the conclusion of therapy, the degree of improvement must be statistically reliable. However, there are other possible criteria for clinically significant change. Mental health treatment investigators seem to be taking the notion of clinical significance seriously. With increasingly frequency, clinical trials are including one of several methods to report their data in ways that emphasize the practical importance of the findings.

### Operationalizing Clinically Significant Change

#### Cutoff Points for Clinically Significant Change

In examining attempts to define criteria for clinically significant change before 1984, Jacobson and colleagues concluded that there was little consensus for either a particular disorder or general criteria that could be used across disorders. The emerging approach was consistent with others that it would be helpful to objectify the criteria for returning to normal functioning (Kazdin & Wilson, 1978; Kendall & Norton-Ford, 1982). However, most proposed solutions were either arbitrary (e.g., Jansson & Ost, 1982) or highly subjective (e.g., Barlow & Mavissakalian, 1981). Although the field needed conventions for what is meant by clinical significance, it is best if these conventions are objective, relatively free of bias, and psychometrically sound. Furthermore, the conventions should be applicable to a wide range of clinical problems.

Three mathematical criteria were proposed for demonstrating that clients had moved from the dysfunctional to the functional range during the course of therapy on whatever variable is being used to measure that clinical problem. The three proposed methods were (a) cutoff point \( a \), achieved when the level of
functioning fell outside the range of the dysfunctional population, where range was defined as extending to 2 SD s above (in the direction of functionality) the mean for that population; (b) cutoff point b, achieved when the level of functioning fell within the range of the normal population, where range was defined as beginning at 2 SD s below the mean for the normal population; and (c) cutoff point c, when the level of functioning suggested that the client is statistically more likely to be in the functional than in the dysfunctional population. Mathematical formulas for establishing each cutoff point were provided, assuming that the dysfunctional and normal populations were each normally distributed. On the basis of simulations, we demonstrated that, when the dysfunctional and functional distributions were overlapping, cutoff point a was the most stringent, cutoff point b was the most lenient, and cutoff point c fell in between. In such situations, we argued that using a practically ensures that the client is no longer dysfunctional if that cutoff point is crossed by the end of therapy. The more overlap there was in the two distributions, the more stringent a would be relative to c. However, although some might favor a under such circumstances because it was more conservative, ultimately the choice between a and b is arbitrary. In contrast, c is not arbitrary. It is based on the relative probability of a particular score ending up in one population as opposed to another. We concluded that when the two distributions overlap, c was the best choice as a cutoff point. However, when the distributions were nonoverlapping, c might be too stringent. Moreover, to calculate c, one must have available norms on the outcome measure for both dysfunctional and normal populations. When adequate norms are lacking, a must be used because neither b nor c can be calculated. Finally, despite recommending these three cutoff points, we cautioned that normal functioning may not always be the best criterion for clinically significant change (Kazdin, 1977) and that for some clinical populations these cutoff points may be too stringent (e.g., schizophrenia; Falloon, 1981). Nevertheless, we saw no problem with reporting the proportion of participants who end up functioning normally, even when such criteria exceed the expectations of investigators. These cutoff points have the potential to be used for a wide range of clinical problems, allow for a comparison of the potency of psychotherapy from one area to another, and act as a check against the tendency to select treatment goals based solely on the capabilities of current treatment technology.

The RCI

Jacobson, Follette, and Revenstorf (1984) argued that it was nonsensical to call a treatment effect clinically significant when no change occurred during treatment, regardless of the level of posttreatment functioning. The RCI, as a second criterion for clinically significant change, ensures that the degree of change was of sufficient magnitude to exceed the margin of measurement error. We proposed a formula that was modified, based on corrections suggested by Christensen and Mendoza (1986), for defining reliable change (Jacobson & Revenstorf, 1988). This formula involved dividing the magnitude of change during the course of therapy by the standard error of the difference score.

To illustrate an analysis of clinical significance using cutoff points and the RCI, we present the data of 1 participant originally presented by Jacobson and Truax (1991; see Figure 1). This individual's pretreatment score (indicated by a square) was 85, indicating a great lack of adjustment and cohesion in the individual's relationship. However, this person's posttreatment score (indicated by a circle) was 114, which is almost exactly the normative mean for the Dyadic Adjustment Scale (Spanier, 1976). Because the individual's posttreatment score passed the cutoff value of 105.2 (the vertical line in between the two distribution means), and his reliable change score was greater than 1.96, this individual was classified as recovered.

Refinements of the Method

Wampold and Jensen (1986) questioned the assumption of two normally distributed populations used to justify the three cutoff points. They pointed out that most disorders exist on a continuum from dysfunctional...
to normal and are best characterized as one population, with the dysfunctional group at one end of the
distribution. Furthermore, Wampold and Jenson echoed Jacobson and colleagues’ caveat that the calculation
of cutoff points depended on measuring instruments with good psychometric properties, especially the
existence of norms. If distribution parameters were unknown, they had to be estimated based on the sample
of participants found in a particular study, a sample that may not be representative of the dysfunctional
population.

In responding to Wampold's concerns, Jacobson, Follette, and Revenstorf (1986)
argued that regardless of
the nature of the underlying distributions, there were two distinct populations of interest: those seeking
treatment for a particular disorder and those who neither seek nor need such treatment. As long as such
distinct groups exist, it is at least theoretically possible to identify a cutoff point where an individual is equally
likely to be a member of either group. Jacobson et al. shared the concerns about the potential for varying
cutoff points, depending on the nature of the sample. Ideally, cutoff points would be based on carefully
collected norms for both dysfunctional and normal populations. When such norms were in place, one would
no longer need to worry about the representativeness of a particular sample.

Jacobson and Revenstorf (1988) attempted to address some problems, issues, and dilemmas encountered in
applying the 1984 method, above and beyond their 1986 response to the initial commentaries. First, they
identified conditions in which the RCI was either irrelevant or misleading as a criterion for defining clinically
significant change. The RCI is irrelevant for any clinical problem in which exceeding the cutoff point
automatically guarantees change of sufficient magnitude to rule out measurement error as its source. For
example, when all participants begin therapy scoring in the dysfunctional range, and the dysfunctional and
functional distributions do not overlap, it is virtually impossible for a client to cross the cutoff point unless the
magnitude of change is statistically reliable. Furthermore, the RCI is misleading when used in the absence of
cutoff points for clinically significant change. When used alone, the RCI tells one only if the change was real,
not if it was clinically significant.

Second, the method works best when and only when adequate norms are available for both dysfunctional
and normal populations.

Third, the ideal normative sample would not include participants who were dysfunctional; however,
Jacobson and Revenstorf (1988) recommended that outliers from the normative sample (i.e., people whose
scores look dysfunctional) remain in that sample as long as they are not seeking therapy.

Fourth, studies that use multiple measures of outcome could resolve the dilemma resulting from the common
divergence of findings from one measure to another by forming a multivariate composite that retains the true
score component of each univariate measure. Jacobson and Revenstorf (1988) defined a formula for
estimating the true score and then standardizing it. The standardized true scores would then be averaged to
derive the composite. Cutoff points would be established by using norms based on the composite.

Fifth, in mathematically addressing the problem with discrete cutoff points delineated earlier, Jacobson and
Revenstorf (1988) proposed forming confidence intervals around the cutoff point, defining the boundaries of
these intervals using the RCI. Participants who fell within the boundaries (i.e., within a band of uncertainty)
could not be reliably classified.

Jacobson, Follette, and Revenstorf (1984) noted that their methods were only as good as the available
outcome measures. In a study with multiple measures, the ideal cutoff point, $c$, can only be applied when
norms are available for both dysfunctional and normal populations on all measures. When such norms are not
available, either $a$ or $b$ will be needed, even though neither are ideal. Although Jacobson and colleagues originally favored $a$ because it is more conservative with overlapping distributions, there are several hurdles to using $a$. As earlier critics noted, unless norms are available for the population of dysfunctional participants on a given measure, the investigator is forced to rely on the sample in a particular study to derive $a$. Because both means and standard deviations of dysfunctional participants vary from study to study, estimates of clinical significance will also vary accordingly. The determination of clinically significant change should not depend on the vagaries of a particular client sample. Therefore, the absence of norms effectively eliminates cutoff point $a$ from providing investigators with what they want: a standardized estimate of clinical significance that can serve as the basis for comparison across studies.

An additional source of unwanted variability is introduced by the tendency of measurement batteries to vary from one study to another. Cutoff point $a$ will provide different estimates of clinical significance to the extent that investigators use different measures of outcome. This once again gets away from the original purpose of standardizing derivations of clinically significant change.

In fact, Jacobson and Truax (1991) agreed with Hollon and Flick (1988) that cutoff point $b$ was preferable to $a$ when norms were unavailable on dysfunctional populations or when there was no consensus within an area on the appropriate measurement battery. As investigators within a given area (e.g., depression research) move toward consensus on a common assessment battery, and norms are adequately collected on dysfunctional populations, $a$ becomes feasible and arguably even preferable. But the reality is that such consensus has rarely been achieved, and this fact alone argues for the use of $b$ rather than $a$. The only prerequisite to using $b$ is the availability of adequate norms from a group of functional people. Such norms already exist for many outcome measures and thus allow for the applicability of $b$ in those studies. As the psychometric properties of our measures improve, and as the field moves toward common assessment batteries, $c$ will once again be practically feasible and therefore preferable (Jacobson, Follette, & Revenstorf, 1984).

The working definition of clinical significance is also limited to the extent that returning to normal functioning is not feasible, either because the disorder is incurable (e.g., schizophrenia or autism) or the current treatment technology is limited (e.g., some addictive behaviors). For these clinical populations, the method may not be applicable unless the use of cutoff point $a$ is feasible.

The method does not in and of itself establish a causal relationship between the treatment under examination and the outcomes with which this treatment may be associated. The metric for merely describing variability in outcome and clinical significance is silent about both the quality of the experimental design and the reliability of treatment effects. In addition, the method is complementary to, and not a substitute for, the reporting of effect sizes, tests of statistical significance, or internally valid experimental designs.

Finally, our method does not allow for the determination of clinically significant deterioration. Although the RCI can be used to identify deterioration, there are no established methods for determining cutoff points to see whether or not the deterioration is clinically significant. Furthermore, the logic of the method does not allow for such cutoff points.

**Applications and Misapplications**

The method has been applied, as of July 1998, in hundreds of clinical trials spanning the range of severe to mild disorders, both in children and adults. The most consistent and standardized applications were found in
Research on couple therapy outcome (Jacobson & Addis, 1993). This finding is not surprising, given that the method was originally developed in response to a concern about misleading findings from the couple therapy outcome literature. Although study after study showed that various forms of couple therapy were significantly more effective than a control condition (i.e., another type of therapy or no therapy), such apparent empirical support belied the relatively weak performance of these therapies. High percentages of couples were not responding to treatments, but the spontaneous remission rates among control couples were so consistently low that even small mean differences were routinely statistically significant, resulting in apparently effective treatments. When Jacobson, Follette, Revenstorf, et al. (1984) examined the clinical significance of a number of previously published data sets, they documented the disappointing performance of behavioral couple therapy. Since then, as investigators have moved toward a common assessment battery, and norms have been collected on both dysfunctional and happily married couples, investigators have often examined clinical significance in their clinical trials. The studies have established that between 50% and 64% of treated couples improve, regardless of treatment, and between 35% and 50% join the ranks of the nondistressed as a result of therapy (Jacobson & Addis, 1993).

Misapplications

Some studies applied the method inappropriately while comparing outcomes derived from it to arbitrary a priori measures of clinical significance. For example, assume that the method is compared with an a priori criterion for cure following Treatment A for patients with Disorder B. Some investigators have reported that fewer patients are cured than meet the criteria for clinically significant change. They then suggested that high cure was a much more stringent criterion for clinical significance than the cutoff points. However, these results are often misleading because investigators used cutoff point a, apparently because adequate norms were not available for either functional or dysfunctional populations on the measures of interest. When the distributions of the two populations were nonoverlapping, it was possible for clients to move beyond the dysfunctional range but still be far from the normal population, which exhibited little or no dysfunctional behavior. With some populations, given the nonoverlapping distributions, the method is useless without norms because a will inevitably be too lenient, and therefore either b or c would be necessary to approximate returning to normal functioning. When adequate norms are available on measures of these same problems, often the method produces proportions that are more conservative than comparable measures of cure (e.g., Jacobson, Wilson, & Tupper, 1988). The problem with some applications was that, given the nature of the clinical population, adequate norms on the outcome measure were missing. As a result, a criterion for clinically significant change had to be used that was inappropriate for the problem being studied. As we stated, the method is only as good as the psychometric properties of the outcome measures.

Many studies have confused the RCI with clinical significance. Investigators report the proportions of participants who have generated reliable change and conclude from those proportions that the clients' changes are clinically significant. The RCI is necessary when crossing the cutoff point by itself does not necessarily prove that the change is real. However, the RCI is never sufficient for demonstrating clinically significant change. In fact, by itself it has nothing whatsoever to do with clinical significance. It is only when the index is used in conjunction with the cutoff points that it is useful, and even then it is only necessary under some conditions, those specified earlier in this article.

Finally, the method cannot be used to validate a measure. Unfortunately, this was another misapplication. For example, some investigators compared multiple measures of Problem A and found that only one showed reliable and clinically significant change in the treatment group relative to a placebo. The other three measures did not produce significant differences in proportions of participants who benefited from treatment. The authors concluded that the measure that worked for them was more sensitive to change than the other three,
basing their conclusions on the significant group differences in clinical significance. However, there is another, equally plausible interpretation of their findings: It may be that the treatment was ineffective but the measure of sensitivity is actually invalid and artificially inflated the group differences. The method is not intended for validating the sensitivity of outcome measures. There is no logic or empirical basis for using it to that end.

**Appropriate and Innovative Applications**

Despite the many misapplications of the Jacobson, Follette, and Revenstorf (1984) method, the vast majority of applications have been appropriate, even creative, and have moved the field forward by advancing knowledge and at times determining the validity of a priori measures of clinical significance. For example, Ogles, Lambert, and Sawyer (1995) made a valuable contribution to the Treatment of Depression Collaborative Research Program (Elkin et al., 1989) by examining the clinical significance of this multisite comparison of cognitive therapy, interpersonal therapy, pharmacotherapy/clinical management, and pill placebo/clinical management for depressed outpatients. They were able to demonstrate that the findings reported earlier by Elkin et al., based on a priori criteria for recovery, were consistent. The convergence between findings based on the method and those developed by consensus increases confidence in both.

More recently, the method has been extended to areas far beyond the domain that we originally considered: prevention research (Hawley, 1995; Roberts, Neal, Kivlahan, Baer, & Marlatt, 1998), naturalistic studies of psychotherapy (e.g., Howard, Lueger, Maling, & Martinovich, 1993), and health status surveys (McHorney & Tarlov, 1995). Nietzel and Trull (1988) even came up with a way of combining the logic of clinical significance with meta-analysis, so that meta-analysis can be used as an aid in determining the practical importance of treatment effects across numerous studies.

Along with these applications have come critiques of the method. These critiques involve both suggestions as to how to solve problems raised in previous articles and alternative ways of deriving the RCI and establishing cutoff points for clinically significant change. Lively debate has been generated by some of these critiques.

**Critiques, Caveats, and Unresolved Issues**

Hsu (1989) was the first to create an alternative RCI that took into account a specific type of measurement error: regression toward the mean. The earlier model used observed scores and treated them as if they were true scores, whereas all scores on an outcome measure carry with them some baggage that is due to imprecision in the testing instrument. Unless your instrument is perfectly reliable, you will not get the exact same score if you take the test twice. Scores tend to gravitate toward the mean; thus, if you score on the extreme end of a distribution, you have a greater likelihood of being pulled toward the mean.

Hsu, as well as other statisticians, have attempted to estimate true scores, taking into account regression toward the mean. He proposed residualized scores, measuring whether a person's posttest score is larger or smaller than the value predicted for that person using the linear regression of the posttest score on the pretest score. According to Hsu (1989), these scores provide more information than change scores about differences that are due to measurement error alone, using the standard error of prediction to reflect measurement error. With Hsu's formula for the RCI, it isn't always the case that the direction of the change score determines whether the client has improved or deteriorated. A score could go up, and the client could be considered deteriorated if the increase was significantly smaller than what regression toward the mean would lead one to expect.

As Hsu (1989) himself noted, a major problem with this method is that it requires knowledge of which group
mean a client's score is expected to regress toward, and knowledge of this group mean is often unavailable. Nevertheless, Hsu was the first to grapple mathematically with the limitations of change scores in computing the RCI and attempt to partial out change due to regression toward the mean.

Speer (1992) not only illustrated a method of calculating the RCI that takes regression toward the mean into account, but he also compared the improvement and deterioration rates obtained by his revised method to those obtained by our method. He proposed the Edwards-Nunnally method as a means of preventing improvement rates from being artificially inflated by regression toward the mean. This formula uses confidence intervals consisting of 2 \(SD\) s of measurement to set up criteria for determining improvement or deterioration. The interval is centered on the person's estimated true score, which is free of bias on the basis of regression toward the mean, rather than the obtained score, which may well be biased. The confidence intervals created by this model are asymmetrically conservative. The upper limit on the confidence interval is farther away from the pretest score than the lower limit. The impact of this asymmetry is to make it more difficult to change reliably than with our method when the pretest score is relatively close to normal functioning and to make it less difficult when the score at pretest indicates extreme dysfunctionality. Speer (1998) has recently concluded that regression toward the mean, although always a possibility, is not nearly as frequent in actual clinical data sets as some believed. Therefore, it remains an open question whether these revised formulas actually bring investigators closer to true and accurate measures of clinical significance. Speer (1998) has also taken the lead in performing empirical tests on which methods work best, and the results of these empirical tests will be quite informative.

For example, Hageman and Arrindell (in press) suggested a modification of our method by using Cronbach's notion of maximum risk, which represents the likelihood that an individual has been classified correctly. Hageman and Arrindell distinguished between two levels of analysis: the individual classification of a client and the estimations of proportions improved at the group level. For calculating the RCI and establishing cutoff points for individuals, they retained the logic of clinical significance but argued that it provides imprecise and overly conservative methods of classification. They suggested some modifications to the RCI and to the determination of cutoff points, which do indeed create higher percentages of improvement and recovery estimates (McGlinchey & Jacobson, in press). Moreover, given the conservative estimates provided by our method, Hageman and Arindell argued that summing up and obtaining proportions based on these measures will underestimate true clinically significant change. They proposed obtaining group proportions by calculating \(z\) scores.

The Hageman and Arrindell (in press) model requires a significant amount of mathematical computation. Although their calculation of the RCI for individuals is very similar to our calculation (with very reliable instruments, the formulas are virtually identical), their method leads to different cutoff points for each individual in the sample. Moreover, their method requires calculating two \(z\) scores, one that estimates the proportion of individuals who have reliably changed (a group statistic) and one that represents the proportion of participants who passed their individually tailored cutoff points.

These mathematical alternatives are difficult to evaluate because they all make valid arguments, each based on different assumptions or premises. To add even further complexity to the debate, there are recent critiques that raise both conceptual and mathematical issues and grapple with limitations of our method that we have, so far, been unable to overcome (Jacobson, Follette, & Revenstorf, 1984; Jacobson & Revenstorf, 1988; Jacobson & Truax, 1991). For example, Tingey, Lambert, Burlingame, and Hansen (1996) introduced the concept of adjacent samples to distinguish between extreme groups of dysfunctional and normal clients and normative samples of clients who fall somewhere in between. We mentioned earlier that nonoverlapping distributions pose a problem for our method. When dysfunctional and normal
distributions do not overlap, it usually means that the dysfunctional group is extremely dysfunctional. For these types of problems, returning to normal functioning may be viewed as an unrealistic and excessively stringent criterion for clinically significant change.

For example, Tingey et al. (1996) distinguished between inpatient psychiatric patients, outpatient psychiatric patients, normative samples (which included some dysfunctional people who had not sought professional help), and particularly well-functioning people devoid of mental health problems. They then suggested that both reliable and clinically significant change can occur when a client moves from the group that they are in to an adjacent, better functioning group. Thus, an inpatient client who moves into the outpatient (but still dysfunctional) range would be considered improved to a clinically significant degree. Although the concept of adjacent groups requires even better psychometric properties than those necessary for other measures (including normative data that do not exist), at least in principle the application of adjacent samples—if properly normed—would allow for the use of cutoff point $c$. Adjacent samples are overlapping, by definition, and they need to be overlapping for their method to be applicable.

The proposals of Tingey et al. (1996) have generated controversy. For example, Martinovich, Saunders, and Howard (1996) pointed out both conceptual and statistical problems with using currently available measures to generate nomothetic indexes of clinically significant change. They criticized Tingey et al.’s proposals for violating the spirit of clinical significance by using examples involving global constructs (e.g., psychotherapy) measured by catchall instruments such as the Symptom Checklist—90. Martinovich et al. bring investigators back to a more idiographic consideration of particular treatments for particular constellations of problems. It is true that the nomothetic examples used by Tingey et al. create diagnostic and other measurement problems because even clients united by a Diagnostic and Statistical Manual of Mental Disorders classification may differ in ways that are considerably more important than their similarities. For example, two depressed people can differ dramatically in their degree of functional impairment both at work and at home. Assessments of clients’ posttreatment status must somehow incorporate more than mere topographical measures of a presenting complaint and take these into account when evaluating change.

Follette and Callaghan (1996) made a similar point. Most importantly, they bring investigators full circle to the original intent of calculating clinical significance. The goal was to report psychotherapy data in a way that was clinically meaningful, given the expectations that consumers have about psychotherapy. Offering a vigorous defense of returning to normal functioning as a criterion for clinically significant change, Follette and Callaghan quoted extensively from Jacobson, Follette, and Revenstorf’s (1984) article to remind investigators that the original purpose of operationalizing clinically significant change was to provide data that are (a) more meaningful to practitioners of therapy than standard group comparisons and, more importantly, (b) informative to the client as a consumer of mental health services. The ultimate question posed by clinical significance proponents remains the same as it was 20 years ago: How likely is it that a client with a particular problem will leave therapy without that problem?

**Future Directions: Back to the Future?**

The field has a long way to go in documenting the clinical significance of its psychotherapy effects. The original method did not come out of the blue but was part of a call for a more clinically meaningful way of reporting psychotherapy outcome (Barlow, 1981; Kazdin, 1977; Kendall & Norton-Ford, 1982; Wolf, 1977). Clinical significance proponents have been advocating its psychotherapy effects for 2 decades, yet there are still a predominance of conclusions based on small statistical effects of little practical importance.
and a tendency toward overinterpreting group differences that may not satisfy clients but nonetheless are used by researchers to confirm their a priori hypotheses. We contend that most clients enter therapy wanting an end to their suffering, not simply a statistically reliable improvement. As Follette and Callaghan (1996) put it, until therapists and therapy researchers are willing to tell their clients and society at large that they can't return people to normal functioning, this still strikes us as a reasonable criterion to shoot for.

This does not preclude diversity in normative samples. We like the idea of separating representative or random samples of people not seeking therapy from those who are truly healthy (Jacobson & Revenstorf, 1988). Moreover, we applaud the efforts by Tingey et al. (1996) to collect norms on widely used measures of psychotherapy outcome. However, we are struck by how slow the field has been to deal adequately with the subtleties of outcome measurement, subtleties that incorporate the social functions of presenting problems, their implications for performance as workers and parents, and some information about the process of change (e.g., the work of Speer [1998] in debunking a tendency for statisticians to be unnecessarily concerned with regression toward the mean in psychotherapy research). It is still true that any method of defining clinical significance can be no better than the measuring instruments and their psychometric properties.

Finally, we find ourselves hoping that the field, in its attempt to zero in on definitions of clinically significant change, does not get hung up on mathematical minutiae at the expense of conceptual clarity and accessibility. As Follette and Callaghan (1996) noted, without an emphasis on clinical significance, expectations about psychotherapy will drift in the direction of what is currently attainable, not what the client has a reasonable right to expect. We strongly disagree with the notion put forth by Tingey et al. (1996) that psychotherapy is about the extent to which society is satisfied with treatment. Scientists—practitioners should aspire to not only satisfy society but also improve the quality of therapy. Those developing psychotherapies should set a higher standard than those held by consumers, instead of the other way around ("Mental Health," 1995). Not that societal standards are irrelevant; however, as Strupp and Hadley (1977) suggested in a tripartite conceptualization, outcome should be simultaneously assessed from the vantage points of the individual, society, and mental health professionals.

As to which method is best, less mathematical wrangling and more empirical testing is needed (e.g., Speer & Greenbaum, 1995). Ultimately, questions of clinical significance methodology boil down to questions of performance: Which method works best for particular clinical problems and particular groups of patients? For example, we are currently planning to compare various methods (e.g., Hageman & Arrindell, in press; Hsu, 1989; Jacobson, Follette, & Revenstorf, 1984; Speer, 1992) to see which ones tell us most accurately who has actually benefited from psychotherapy on the basis of external, qualitative criteria, such as the relative probability of relapse. Ultimately, it is these kinds of tests that will settle the mathematical questions. Conceptually, clinical significance remains a viable and essential goal for psychotherapy, and its measurement is of major importance.

References


Hageman, W. J. & Arrindell, W. A. (in press). Establishing clinically significant change: Increment of precision and the distinction between individual and group level of analysis. *Behavior Research and Therapy, 60*.


Psychotherapy Research, 6, 124-132.
Mental health: Does therapy help? Consumer Reports, 734—739.

For additional information on outcome measures and norms, see Ogles, Lambert, and Masters (1996).
