CHAPTER 9

ASSESSING CLINICAL SIGNIFICANCE[*]

HOWARD GOLDSTEIN
University of Pittsburgh

Defining and assessing "clinical significance" are particularly thorny issues. In this area of treatment efficacy research, there is not a set of readily agreed upon scientific conventions for demonstrating whether changes attributable to treatments in speech-language pathology and audiology are clinically significant. A number of ideas for assessing clinical significance will be proposed, but a number of decisions for implementing these procedures will remain. It should be clear that many of these ideas require further research.

First, one must recognize that "clinical significance" refers to whether a treatment that meets scientific standards for effecting behavior change is also significant in a clinically or socially relevant way. That is, one is concerned with whether treatment effects reflect *important* and *acceptable* changes in behavior. There is general agreement that criteria for determining statistically significant treatment effects do and should differ from criteria for determining clinically significant treatment effects. The problems are well known: Statistically significant treatment effects may reflect modest changes by all or most clients undergoing treatment; also one may find that there are relatively large changes in some subset of clients undergoing treatment but little or no changes in other clients. Others (Jacobson, 1988) have argued that even if large changes are demonstrated by all or most of the clients undergoing treatment, some standardized criteria for determining whether those "large" changes are important and acceptable are needed.

There are two basic approaches to assessing clinical significance: Normative comparison approaches and social validity approaches. I will argue that both these approaches are potentially useful to both group and single-subject experimental design research.

Normative Comparison Approaches

Normative comparison approaches have sought to demonstrate that outcomes reflected in treatment efficacy research allow clients who evidenced a psychological or a communicative disorder before treatment to be characterized as normal after treatment. That is, after treatment the "clinical population" that underwent treatment now can be characterized as "non-clinical" or within normal limits. This is a stiff test of treatment efficacy. In order to argue convincingly that treatment effects are clinically significant, then clients must appear essentially normal on some relevant outcome measure(s) after treatment. Such outcomes clearly would be *convincing* from the perspective of a skeptical consumer or a skeptical scientist, for that matter. On the other hand, this

[*] Preparation of this manuscript was supported by Grants No. G008730526 and G0087308528 from the U.S. Department of Education to the University of Pittsburgh. Address correspondence to Howard Goldstein, Department of Communication, 1117 Cathedral of Learning, University of Pittsburgh, Pittsburgh, PA 15260.

approach fails to consider the sometime profound differences among clients before treatment. Many of our clients and the significant others in their lives might be and perhaps should be satisfied with improvements in communicative functioning that fall short of normal functioning.

The normative comparison approach, of course, necessitates that "normal functioning" be defined. Perhaps there is a catch here. The normative approach may appear more flexible if careful consideration is given to how and with whom one generates normative data. Who should comprise our normative sample? This question could generate a good deal of discussion, debate, and hopefully research. Let's contrast the selection of two different populations. First, one could identify age-matched individuals who have no communicative disorders and have them perform the same tasks that are being used for our outcome measures. Age-matched individuals do not restrict our sample too much. The intent here is simply to ensure that our sample does not overlap with our clinical population. Thus, the variability of scores should reflect the reliability of the measures, natural behavioral variation in this population, and perhaps even some extreme scores from those individuals with communicative disorders who have not sought treatment (Jacobson & Revenstorf, 1988).

Our second alternative could be to select a much more restrictive sample. Instead of matching only on age, one could take into account cognitive level, and physical and sensory disabilities. In fact, there are numerous variables that might be considered: race, gender, SES, educational level, language background, to mention but a few. Our intent here may be to avoid overlap with our clinical population again and to identify individuals who are functioning well or optimally and who despite certain disabilities are not judged to be in need of treatment. These types of judgments are difficult to make and collecting a large normative sample from such a population may be impractical. But generating peer comparison data even from a small sample may be useful. Peer comparison data may provide a valid way to address clinically significant change among clients who are not expected to function "normally." Take communication board users, for example--Does one have any alternative for judging the effects of treatment for communication board users, in terms of clinical significance?

There are a number of other questions that should be addressed about collecting a normative sample. How similar must the task and the situation for the normative sample be to the clinical task? To what extent should that task reflect generalized performance under natural, conversational conditions? How large a sample is sufficient for the purpose of comparison? Should one develop norms that are appropriate to a clinical population or norms that are appropriate for individual clients? There are methods available for addressing questions about how many observations are typically enough, for example as seen in generalizability theory (Brennan, 1983). But it is difficult to find examples of the use of generalizability analyses in our literatures.

Researchers have proposed a number of alternatives for standardizing the use of normative comparison data. For example, Jacobson and his colleagues (Jacobson, Follette & Revenstorf, 1984, 1986; Jacobson & Revenstorf, 1988) propose the use of two statistics to report the proportion of clients who have changed to a clinically significant degree. First, they formulated a *reliable change index*, a statistic that is defined in terms of the standard error of measurement for a particular measure. Second, they arrive at a *cut-off point* that is midway between the means for norms developed for a clinical and a nonclinical population; alternatively, they suggest that the cut-off points could be one standard deviation above the mean for the clinical group. Note that this proposal requires that norms be developed for the clinical and nonclinical populations; this can be an expensive endeavor.

Others (Nietžel & Trull, 1988) have proposed that norms could be derived from meta-analyses. Meta-analysis converts data derived from a given study to a common metric, effect size, and then summarizes the size of the normative comparison across a group of studies to evaluate a given treatment. Nietzel and Trull (1988) suggested that scores within one or two standard deviations of the normative population (depending on how the normative population is constituted) be considered clinically significant. Meta-analysis does not overcome the methodological deficiencies in the component studies, however. One should be concerned with whether the integrity of treatment was sufficient, whether there was differential client attrition, whether "blind" assessments were conducted, and so on. The stricter the set of criteria used in selecting studies for the normative and clinical data set, the fewer studies that can be included. Given the paucity of well-controlled treatment studies that might contribute to a worthwhile meta-analysis at this point, time might be better spent on the independent development of norms. Furthermore, to the extent that meta-analysis is used in assessments of clinical significance, dependent measures used to evaluate treatments must be restricted to those with norms.

Single-subject researchers also have used normative approaches. For example, Osnes, Guevremont, and Stokes (1986) developed norms for the amount of talking among preschoolers during free play by observing the seven classmates of two socially withdrawn target children. Doyle and his colleagues (in press) obtained data from age-matched individuals to determine the rate of requests for information one might expect for their conversational task situation (plus and minus one standard deviation). These researchers judged the clinical significance of their treatment effects by the degree to which the clients demonstrated communication skills within normal limits as demonstrated by their nonclinical populations. These two examples use to their advantage the collection of data in situations nearly identical to the treatment situation. On the other hand, one might quibble with the representativeness of the nonclinical sample selected and whether adequate numbers of comparison subjects were assessed. Nevertheless, these studies illustrate that norms, even if preliminary in nature, can be developed based on peer comparisons in an economical manner. Experimenters need not rely on impressions or arbitrary cut-off points to judge the clinical significance of their treatment effects. Normative comparison data offer a viable approach to assessing the degree to which results are clinically significant.

Social Validity Approaches

A second approach to assessing clinical significance is through social validation (Kazdin, 1977; Wolf, 1978). Social validity assessments have come to be commonplace in many behavioral journals. In a recent review, Schwartz (1989) found that 40% of the articles in the 1987 volumes of *Journal of Applied Behavior Analysis* and *Behavior Modification* reported social validity measures. Social validity assessments are far from commonplace in our speech and hearing journals at this time. Social validity is the evaluation of treatment effects based on feedback from people other than the experimenters. These other people will be called "consumers." Social validity assessments sample the opinions of relevant consumers of clinical treatments. Consumer opinions should evaluate the importance of treatment effects based upon the social context in which the client functions. This information could be used not only to evaluate the importance of treatment effects, but also can be used to modify intervention programs. In fact, social invalidity can be particularly revealing if one can trace sources of disapproval or dissatisfaction on the part of consumers.

Three types of questions have typically been asked in social validity assessments (Wolf, 1978): questions about the social significance of goals, questions about the social appropriateness of intervention procedures, and questions about the social importance of treatment effects. Questions about the social significance of goals allow consumers to react to that selection of behaviors that are changed or to be changed through direct intervention. Are the behavior changes relevant to the desired outcome of assisting people to communicate more effectively and more independently? Consumers or competent communicators might be asked to simply list along certain parameters the most desirable behaviors for a communicative situation. For example, Whitney and Goldstein (in press) had judges listen to tape recordings of clients with mild aphasia before treatment; they were asked to note what, if anything, made the speakers sound "different," "unusual," "distracting," or "hard to listen to." Nine characteristics were identified, which were later used to evaluate the clinical significance of the treatment effects. A similar approach was used by Runco and Schreibman (1987, in press), in which they asked school children to view videotapes of autistic children and to identify "unusual behaviors." Alternatively, it might be possible to ask consumers to view videotape samples, judge what behaviors have changed, and rate the importance of each of these apparent changes to improvements in the communicative situation.

The second question seems to be addressed relatively infrequently. Questions about the social appropriateness of procedures invite consumers to rate the desirability of the intervention procedures themselves. The more desirable consumers find the procedures the more likely they are to choose to implement those procedures or be an advocate for their use. If consumers find the procedures to be distasteful for whatever reasons then another course of action is called for, from educating consumers to creating more palatable treatment alternatives.

The final question is the one that usually attracts most interest among applied researchers: What is the social importance of a treatment effect? This involves consumer's satisfaction with behavior changes. These behavior changes could include those that were predicted or targeted as part of the intervention, as well as those side effects (be they positive or negative) that may not have been anticipated.

Hayes and Haas (1988) argue that social evaluation is "a humble, straightforward measure, based on the common-sense view that important changes can be seen." A common way to assess social validity is to ask parents and teachers of clients to rate improvement. This is a simple form of obtaining consumer satisfaction ratings. For example, Koegel and his colleagues (1988) asked parents and teachers to rate changes in a child's speech as (a) showing marked improvement, (b) showing improvement, but with some remaining errors, or (c) showing no improvement. Such global appraisals may reflect whether behavioral changes had an impact on how significant others viewed the child. They also are more likely than more specific measures to reflect bias, however. Consumers are likely to rate behavior as improved, when they are led to believe that performance will improve with training (Garfield, 1983). Some care should be taken to ensure that the consumers doing the evaluation are basing their ratings on therapeutic effects, not on preconceived notions.

Perhaps a better way of assessing social validity is to have consumers make relative judgments about two recordings (audio or video recordings) that sample a client's communicative behavior. Raters can observe one recording from before treatment and the other from after treatment without knowing which one occurred first. Care must be taken to maximize the degree to which "blind" judgments are obtained. Therefore, the order of these recordings for various clients would be systematically balanced so that posttreatment recordings would not always follow pretreatment recordings.

Another set of critical decisions involves selecting samples of clients' behavior in a way that will not bias results. It is important to consider to what degree selected samples represent typical performance. For example, it would not be fair to take the worst of baseline performance and compare it to the best of posttreatment performance. Alternatively, one might select recordings that reflect the average of the last few baseline sessions and the average of the last few treatment sessions. This alternative is an attempt to ensure that more representative samples are chosen. It also might be advisable to select several samples of pretreatment and posttreatment performance. Finally, one should consider whether samples from treatment tasks or perhaps from generalization tasks should be selected.

The next step in the process involves the task of the judges. On what dimension(s) should the judges rate the samples? Typically, investigators have used Likert-like scales with multiple dimensions. The questions should be relevant and specific to the behavioral changes expected, but should be broad enough to ask about related effects that may not have been anticipated. Scales should allow for a wide variation in responses; for example, a 7-point scale might be more revealing than a 4-point scale. In addition, for control purposes, it might be useful to include ratings of normal individuals or ratings of two samples that should not differ. Other types of scales with a more objective basis for judgment might also be useful. Or consumers might simply be asked to make a forced-choice among two or more samples to select the one they find most acceptable. With so many alternative approaches to conducting social validity assessments, it should be clear that much research is needed to identify parameters that may bias or otherwise affect the evaluation of consumers. Such research hopefully will provide a basis for outlining a set of guidelines for conducting reliable social validity assessments.

Finally, a most critical question is who should decide whether treatment effects are important. Who are the consumers? There are many potential consumers. They include: (a) the recipient of the treatment, our clients; (b) the referral agents or all those who bring to our attention complaints about the client's communication skills, (i.e., parents, spouses, and friends, as well as teachers, doctors, and other professionals); (c) the larger immediate or extended community, (possibly including spouses, teachers, neighbors, peers, or others not necessarily familiar with the client); (d) people from future environments, such as prospective teachers, peers, neighbors; and (e) professionals (e.g., speech-language pathologists) who are familiar with the population and might be especially sensitive to the changes and adequacy of changes in communication skills. One could argue that a wide spectrum of consumers should participate in assessments of social validity. In fact, if our treatment efforts are successful and that success is readily perceptible to a variety of consumers then the participation of a wide spectrum of consumers would have a number of benefits. Social validity assessments could facilitate higher rates of adoption of our treatment approaches and perhaps help to solicit community support for effective treatment programs.

Although there may be benefits to including a number of consumers in social validity assessments, one still needs to consider whose input is most central to our task of convincing journal readers (another set of consumers) of the clinical significance of treatment effects. The consumers who most need to participate in social validity assessments are those whose perceptions need to change to stop the complaint of a communicative disorder (Baer, 1988). If for example, one considers an adult voice case, the most relevant consumer might be the client; clients are interested in treatment outcomes, because they are interested in their own well-being. In other cases, it might be a set of teachers who are concerned about whether they can handle children in their regular education classes next year when these children have received special education services in the past; this may reflect larger society's interest in treatment outcomes and whether behavior improvements adhere to social norms (e.g., standards of conduct for students). In other cases, professionals who are largely responsible for deciding whether stroke patients are likely to benefit from additional therapy might be the most credible raters; professionals may make judgments that take into account theoretically prescribed aspects of behavioral functioning (e.g., brain-behavior relationships). Baer (1988) points out that clinicians will try to change a client's behavior when someone's complaint about the client's behavior and the clinician's values about the behavior are in concordance. The bottom line approach is that a good case for clinical significance can be made if people are convinced that continued treatment, or at least continued treatment for the communication problems originally identified, is unnecessary. Thus, it is with some care that one decides on the consumers who should judge the social validity of treatment effects.

Recommendations

The most important message reiterated throughout this conference is that more treatment efficacy research needs to be done in our discipline. In child language disorders, for example, most of the treatment research has been done by psychologists and special educators. I am thankful for their contributions, but lament the fact it has taken so long for those of us most responsible for service delivery to contribute empirically derived treatment procedures to our own discipline. As part of our commitment to generating high quality treatment research, we must recognize that good treatment efficacy research includes assessments of clinical significance. Assessments of clinical significance are important supplements to our primary means of analyzing data, be it through statistical analysis or through visual inspection of outcome data. Frankly, we should believe and expect that clinicians who are consumers of our research should be convinced that they should adopt treatment procedures only when researchers report that other relevant consumers have deemed outcomes to be clinically significant.

Two basic approaches to assessing clinical significance have been discussed, normative comparison approaches and social validity approaches. Assessment practices from both approaches should be adopted. The strongest cases for clinical significance will be made when treatment outcomes are evaluated with respect to the *goal of* having particular behaviors fall within a normal range and with respect to consumers' evaluations of whether important changes in behavior have occurred.

It is important to point out that both approaches are compatible with both group and single-subject experimental designs. As noted above, normative approaches are being used with increasing regularity regardless of design. In group designs, a comparison of group means still remains the accepted way of reporting treatment effects. In the future, it is important to report the percentage of clients who achieve clinically significant changes (Blanchard & Schwarz, 1988). In single-subject designs, the percentage of observations that fall within a normative range should be readily apparent as depicted visually in figures.

Social validity assessments have become rather commonplace in single-subject design research. It should be pointed out, though, that group designs do not preclude the use of social validity assessments. The problem is one of practicality. Given that it may be impractical to conduct social validity assessments for all subjects, one might select a representative sample of subjects and have raters judge pre- and post-treatment observations.

In the past investigators often relied on clinical opinions as to what treatment objectives were important and testimonials as to how effective a treatment appeared. It is important to note that assessments of clinical significance are more than the subjective evaluation of treatment outcomes. This discussion of assessment techniques emphasized how a reliance on empirical data can be fostered. These evaluation methods are not free of problems. The problems inherent in applying normative comparison and social validation approaches should be viewed as important areas where further research is needed. These problems should not be viewed as obstacles that prevent us from assessing clinical significance. In the process, we can expect to gain a greater understanding of behaviors that make for effective communicative functioning in everyday life and we can expect to gain a greater understanding of treatment approaches that are successful and acceptable to a variety of potential consumers and advocates.

References

Baer, D. (1988). If you know why you're changing a behavior, you'll know when you've changed it enough. *Behavioral Assessment, 10,* 219-223.

Blanchard, E., & Schwarz, S. (1988). Clinically significant changes in behavior medicine. *Behavioral Assessment, 10,* 171-188.

Brennan, R. (1983). *Elements of generalizability theory.* Iowa City, IA: American College Testing Program.

Doyle, P., Goldstein, H., Bourgeois, M., & Nakles, K. (in press). Facilitating generalized requesting behavior in Broca's aphasia: An experimental analysis of a loose training procedure. *Journal of Applied Behavior Analysis.*

Garfield, S. (1983) Some comments on consumer satisfaction in behavior therapy. *Behavior Therapy, 14,* 237-241.

Hayes, S., & Haas, J. (1988). A re-evaluation of the concept of clinical significance: Goals, methods, and methodology. *Behavioral Assessment, 10,* 189-196.

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

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

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

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

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

Koegel, R., Koegel, L. K., Van Voy, K., & Ingham, J. (1988). Within-clinic versus outside-of-clinic self-monitoring of articulation to promote generalization. *Journal of Speech and Hearing Disorders, 53,* 392-399.

Nietzel, M., & Trull, T. (1988). Meta-analytic approaches to social comparisons: A method for measuring clinical significance. *Behavioral Assessment, 10,* 159-169.

Osnes, P., Guevremont, D., & Stokes, T. (1986). If I say I'll talk more, then I will: Correspondence training to increase peer-directed talk by socially withdrawn children. *Behavior Modification, 10,* 287-299.

Runco, M., & Schreibman, L. (1987). Socially validating behavioral objectives in the treatment of autistic children. *Journal of Autism and Developmental Disorders, 17,* 141-147.

Runco, M., & Schreibman, L. (in press). Children judgments of autism and social validation of behavior therapy efficacy. *Journal of Applied Behavior Analysis.*

Schwartz, I. (1989). *Social-validity assessments: Is current practice state-of-the-art*? Unpublished manuscript, University of Kansas, Lawrence.

Whitney, J., & Goldstein, H. (in press). Using self-monitoring to reduce disfluencies in speakers with mild aphasia. *Journal of Speech and Hearing Disorders.*

Wolf, 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.