Strategies for Evaluating Clinical Change: Implications for Practice and Research

(clinical practice, occupational therapy, research design)

Kenneth Ottenbacher    Jennifer York

The methods and procedures of evaluation research using single-system or small-N designs are defined and described. Various components of single-system evaluation research, including specifying the problem, measuring and recording the data, selecting an appropriate design, and analyzing the data, are presented and discussed. The argument is made that single-system evaluation research methods are ideally suited to document clinical change on an individual basis and can provide a mechanism for establishing therapeutic accountability. The relationship between single-system evaluation research and the more traditional experimental procedures is briefly explored, and the implications for establishing an empirically derived basis for clinical practice are discussed.

Recent advances in the field of evaluation research provide therapists with the procedures and strategies necessary to document treatment outcomes accurately. Evaluation research may be defined as the systematic application of applied research procedures to assess the conceptualization, design, implementation, and effectiveness of therapeutic intervention programs. Evaluation research, as presented in this paper, is considered an individualized clinical activity. The hallmark of evaluation research is the "systematic" application of procedures for the continuous collection of reliable information designed to determine the efficacy of therapeutic intervention services.

One method of evaluation research involves the use of single-system designs. Bloom and Fischer (4) argue that single-system designs, also referred to as single-subject, small-N, or idiographic research, are the preferred method for evaluating clinical change in most human service environments. These researchers contend that single-system designs are practice based and practitioner oriented. The method provides continuous assessment and outcome information that can be used by the therapist to monitor patient progress and even adjust the intervention, as indicated based on data interpretation. The flexibility of single-system evaluation methods allows the therapist to modify the treatment, depending on the individual needs of the patient or situation. This is generally not possible with more traditional group comparison designs in which the treatment or independent variable remains the same throughout the investigation.

The single-system model of evaluation research provides a method for incorporating empirical procedures into clinical practice not available in traditional research methods (4-7). Classical group comparison experimental designs are traditionally employed.
to test hypotheses or theoretical predictions important in the establishment of a clinical knowledge base. There are, however, a number of practical problems with applied or field applications of these traditional designs. First, many potential investigators often cannot assemble a large group of patients and then randomly assign them to experimental and control conditions. This is particularly true when the population of interest consists of heterogeneous, widely distributed individuals with handicapping conditions. Furthermore, information provided by group comparison research may not, in some circumstances, be useful for clinical practice. Group comparison designs are often aimed at determining differences between the average change in two or more groups. The clinical therapist, however, is usually not concerned with average change in a group but with change in an individual patient who possesses a unique set of characteristics. The practitioner is not only interested in overall effect at the end of treatment but also in the specific course of that clinical change over time. Using a single posttest on a dependent measure, the usual way of calculating change in groups, does not provide such information. As Strupp and Bergin (8) state, the therapist needs to know "what treatment, by whom, is most effective for this individual with that specific problem, under which set of circumstances" (p 111). Because of their idiographic structure, single-system designs allow the clinician to begin to answer these important individual questions.

Single-system designs permit the therapist to collect data during an intervention period while the treatment is being delivered. These data are then compared with those collected when the intervention is not in effect. The no-treatment period is commonly referred to as the baseline. In this design, the patient serves as his or her own control and provides evidence for a standard comparison against previous performance. The single-system methodology permits the therapist to achieve two major goals: the monitoring of patient progress and the individual determination of intervention effectiveness. The basic components of single-system designs used to achieve these goals are outlined later; these are followed by specific clinically based examples of how single-system designs can be integrated into practice.

**Specifying the Problem and Method of Measurement**

It is crucial that the problem to be treated be identified as clearly as possible and that some behavioral referent for the problem be defined. It is also vital that the therapist and patient or the patient's caretakers agree on the nature of the problem and the most appropriate intervention. Traditionally, standardized assessments have been used to help specify or diagnose a problem; however, standardized norm-referenced assessments may be of limited value in delineating the qualitative aspects or functional relevance of a particular dysfunction. Brown and his associates (9, 10) have pioneered the use of strategies designed to identify the skills needed by handicapped persons to function maximally in their natural environments. These strategies involve the construction of an environmental or ecological inventory that includes all of the settings in which the individual with a handicap interacts on a regular basis. A detailed analysis of that environment is then conducted to determine the specific skills and abilities needed to achieve and maintain the highest possible functional level within each environment. The ecological inventory strategy is designed to develop, organize, and refine information that can be used to determine the exact problems encountered in the handicapped person's environment and then clearly identify the corresponding skills that are required to ameliorate functional deficits.

Once the specific skills that the patient will need to function in his or her natural environments have been identified, these skills are broken down into units of behavior that can be observed, counted, or measured in some way. Each behavior unit can be defined in terms of how often it occurs (frequency) or how long it occurs (duration) and also in terms of qualitatively desirable components. This process will be possible only if the problem to be the focus of intervention is clearly defined and specified.

After the problem has been clearly defined, a procedure for measuring and recording data must be identified. Specifying the rate of behavior is a particularly popular method of measuring patient performance. Rate is generally defined as the frequency of a behavior divided by the time frame in which it occurred. For example, a behavior that occurred 10 times (frequency) within a 2-min period (time frame) would have a rate of occurrence equal to five behaviors (sometimes referred to as movements) per minute. Rate of behavior as a measure of performance has a number of advantages (11). For the purpose of de-
testing changes or trends, rate is a particularly sensitive measure because it has no theoretical upper limit that could create an artificial ceiling effect (12).

Rate is often recorded as a percentage of correct responses in clinical or educational settings (13). Percentages are straightforward and easy to compute. There are, however, many ways in which percentages can be misleading; therapists who use the rate of correct responses, recorded as a percentage, should be aware of these limitations. One limitation of percentage-correct scores is that they do not say anything about the actual number of times a patient has performed correctly. For example, the patient’s performance was 50% correct responses on the first day, which means he or she was successful 50% of the time. If the patient were still 50% correct on the second day, how many trials would he or she complete successfully? Half of 5 is 2.5. The patient, however, cannot complete half of a trial successfully. Either the patient does or does not achieve the target behavior (lip closure). There are no half behavioral units. In other words, the patient, who is provided five trials, cannot demonstrate 50% accuracy. He or she may successfully complete two trials (40% correct performance) or three trials (60% correct performance). When compared with the first day’s data, the patient must look like he or she has improved or gotten worse. A no-change statement is impossible because the score achieved on the first day cannot be achieved on the second day.

Recording performance in terms of rate, frequency, or percent correct response is not as simple or straightforward as it may appear initially. A therapist using this method must be aware of the possible limitations so that misinterpretation of the data recorded does not occur.

Once data have been recorded, they must be organized and presented in some manner. Perhaps the most convenient and desirable way to organize the recorded data is to put them on a chart or graph: Charting results in a pictorial representation of changes in the behavior units from the onset of recording to the termination of intervention. Charting or graphing also assists in the assessment or evaluation process, because data changes reveal patterns that provide valuable information concerning various factors that may have affected performance.

Charting or graphing may involve the construction of relatively simple line graphs, frequency charts, or a more detailed standardized chart, which is commercially available. The Standard Behavior Chart, an example of this more detailed chart (see Figure 1), allows the therapist to chart recorded data in a systematic standardized manner (14).

The chart in Figure 1 is divided into semilog units; this format is selected because of its utility in estimating linear trends in data and the ease with which it can be used by practitioners. Behaviors with extremely high or low rates can be recorded on the chart. The rates of behavior can vary from .000695/min or once in 24 hours to a maximum of 1,000/min. Behavior can be continuously recorded on the chart for up to 20 weeks. The wide range of rates that can be accepted by the chart enhances the accuracy of the recordings, because virtually no data are lost due to ceiling or floor effects. In spite of the apparent complexity of the chart, it is really quite simple to use. White and Haring (14) provide detailed instructions on how to use the Standard Behavior Chart. In addition, Carr and Williams (15) have provided some therapeutic examples of how the chart may be used to record data in a clinical setting.

An issue closely related to recording patient performance is the training and reliability of the recorder. It makes little sense to develop a sophisticated systematic
method of recording performance if the person doing the recording is poorly trained or unreliable in his or her recording activities. The behavior unit to be recorded should be clearly defined, and the recorder must be familiar with how, when, where, and how long the behavior should be observed and recorded. Various procedures have been developed to train observers and to determine the reliability and accuracy of their recording efforts (16, 17).

Selecting a Design
The interpretive logic of single-system designs rests on a comparison between phases when the intervention is in effect and when it is not. The sophistication of single-system, small-N designs has increased dramatically over the past decade. The discussion of various designs presented here is meant to be illustrative rather than exhaustive. Therapists wishing to employ single-system designs as a method of evaluation research should consult one of the excellent resources now available for more detailed information (4, 7, 18-20).

AB Designs. The AB design is the most basic form of single-system design. The A represents the baseline or no-treatment condition (phase A). Subsequent interventions are indicated by different letters of the alphabet. If an intervention or baseline is repeated, then the same letter is used to represent the repeated phase. For example, Figure 2 displays the effect
Response pattern of a mentally retarded patient showing changes in weight over a baseline period and a period of oral motor intervention.

A, baseline period; B, period of oral motor intervention.
Figure 3. Response pattern demonstrating the ability to swallow liquid without loss during periods of oral motor intervention and a period of no intervention.

(Phase A) displays the average percent correct responses (per week) for self spoon feeding in a mentally retarded patient. During phase A, the patient received assistance from the teacher in orienting and guiding the spoon, using a compartmentalized lunch tray. In phase B, the teacher eliminated any physical assistance and provided verbal cues to the patient. Finally, in the third phase (phase C), treatment involved the use of a scoop dish instead of the regular lunch tray. No teacher assistance was provided during phase C. An additional point to be made in this example is that the decrease in performance during phase B was readily apparent via the charting of the data, and as a result an alternate intervention was introduced.

In some cases where the treatment or intervention cannot be achieving a stable criterion level, the criterion is increased. In the next phase, a new and more difficult criterion level is established, while the intervention is continued. When behavior reaches this new criterion level and is maintained across trials, the next phase, with its more difficult criterion, is introduced. For example, Figure 4 demonstrates a patient's progress in developing independent standing ability. In phase A, the criterion was 2 sec of independent standing. In phase B, the criterion was raised to 5 sec, and in phase C, the criterion was 10 sec.

Another alternative is to introduce successive variations in treatments over the phases. Such a design may be referred to as a successive-intervention design (4). Figure 5 illustrates a successive-intervention design. The first phase
withdrawn or withdrawing the treatment would not result in a return to preintervention performance levels, a multiple-baseline design may be appropriate. In a multiple-baseline design, baseline and intervention data are collected on several individuals for a similar behavior or setting, or data may be collected across several different behaviors or settings for a single individual. In other words, there may be multiple-baseline designs across individuals for the same behavior or setting, or across different behaviors or settings for the same individual. In a multiple-baseline design, the intervention phase is generally introduced in a staggered manner across individuals, behaviors, or settings. For example, Figure 6 illustrates a multiple baseline across individuals to assess the effectiveness of an adaptive cup combined with a program of applied behavioral management to teach lip closure to three severely multihandicapped students. During the baseline, information on the ability to maintain lip closure (in seconds) was recorded across trials. The intervention was then administered to each of the three subjects in a sequential, staggered schedule, and the amount of lip closure per trial was recorded. The consistently increasing performance of each individual as the intervention phase was introduced indicates the effectiveness of the procedure across the students. As noted previously, the design examples presented are meant to be illustrative. Many other design variations are possible using a single-system framework. A detailed examination of the many possible design modifications is beyond the scope of this paper. All of the previous examples used are actual clinical cases and were selected to demonstrate the versatility and usefulness of single-system methods for documenting therapeutic outcomes. Therapists desiring additional information on design variations and related methodological and procedural considerations should consult these resources (4, 18-20).

Analyzing Data Collected
Graphic analysis and visual inspection are the traditional analytic tools used to present and interpret the results of single-system evaluation research. However, Christensen (22) has observed that "as single subject designs have become more popular in the applied research areas, there has been an increased emphasis on the need for statistical analysis of the collected data" (p 261). The use of various statistical procedures with single-system designs is a controversial topic. Several authorities feel it is inappropriate to employ statistical techniques with single-system designs and argue that the application of such procedures confuses the issue of clinical significance with that of statistical significance (23, 24). Hersen and Barlow (18) present an excellent overview of the arguments surrounding the issue of clinical significance and statistical significance as they relate to single-system designs. As the use of single-system designs increases in clinical environments, the value of quantitative analysis receives more attention. Gottman and Glass (25) provide a dramatic illustration of the interpretive difficulties that may re-

The American Journal of Occupational Therapy
Figure 6. Multiple baseline design across subjects demonstrating the effect of an adaptive cup and applied behavioral management program to teach lip closure to three severely handicapped students.

Data collected in single-system designs are generally of a repeated nature. That is, measurements on a variable are taken repeatedly based on the performance of an individual or small group over time. The fact that repeated measurements are gathered from a patient means that a degree of serial dependency or auto correlation exists within the data collected. Serial dependency refers to the fact that sequential responses emitted by the same individual will be correlated. The higher the correlation between responses, the better performance can be predicted. The result of this serial dependency is to reduce the variability in the observed responses and thereby bias "the estimates of behavioral score properties like stabilities, variabilities or averages" (26, p 155). The presence of serial dependency within data adds to the interpretive difficulties based solely on visual analysis (27). Cook and Campbell (28) present a detailed discussion of the problems caused by serial dependency and describe several quantitative procedures designed to deal with the effect of auto-correlated data.

We agree with the position of Elashoff and Thoresen (29), who argue that statistical and visual methods should be "partners in the analytic endeavor" (p 291). These researchers contend that statistical and visual analyses should be used jointly to provide a clearer interpretation of single-systems evaluation research.

Components of Visual Analysis

Visual analysis involves the ability to interpret data presented in a graphic format and derive the appropriate clinical implications. Several components or properties of the data have been identified as meaningful in visual data interpretation. The first variation to look for in a graphic presentation of single-system data is a change in level (see Figure 7a). Level, according to Bloom and Fischer (4), refers to the magnitude of the data. When the data evidence an abrupt increase or decrease immediately following intervention, then a change in level has occurred. This change in level is also referred to as a discontinuity in the data.
A second visual property of graphic data is a change in trend. Trend refers to the direction of the data. Figure 7b demonstrates a linear trend, whereas Figure 7c displays a curvilinear or quadratic trend across phases.

The final component of visual analysis discussed in this paper is slope. Slope refers to a change in the pitch or angle of a linear trend. Figure 7b presents data that demonstrate change in slope across phases for a linear trend. Figure 7d represents both a change in level and a change in slope across the two phases, each of which are displaying a linear trend. The terms trend and slope are not defined consistently in the literature and various authorities interpret them somewhat differently (4, 18).

This discussion presents only a superficial overview of some of the basic components that are important in visual analysis of single-system data. For a more detailed discussion of the method and procedure of visual inspection, the reader is referred to the work of Bloom and Fischer (4), Kazdin (20), and Parsons and Baer (30).

Statistical Analysis Procedures

Changes in level or trend may be relatively easy to detect visually; however, changes in slope are much more difficult to interpret using visual inspection alone. Several statistical procedures have been developed to help single-system design users interpret the data they collect. Some of these techniques are relatively complex, such as the autoregressive integrated moving average (ARIMA), and require computer assistance (32). Other procedures, however, are simple to compute and require little familiarity with statistical methods. Such procedures can easily be employed in the clinical setting and provide a degree of quantitative confidence that can enhance the results achieved through visual inspection. Two of these semistatistical procedures with particular clinical relevance will be reviewed. The reader interested in additional information on these or related procedures should consult the appropriate references, particularly the works by Bloom and Fischer (4), Kratcochwill (19), and Kazdin (31).

The Celeration Line Method. The celeration line approach for determining statistical significance is one of the simplest analytic tools to use with single-system designs. A line is computed that divides the data so that half the baseline observations will fall on or above the line and half on or below the line. This line, referred to as the celeration line, indicates whether the pattern of behavior observed during the baseline condition is accelerating, decelerating, or stationary (14). The celeration line is then extended into the treatment or intervention phase to predict what course the patient’s behavior or performance will take in the absence of any intervention (see Figure 8).

Theoretically, the same proportion of data points should fall above (or below) the celeration line during the intervention phase as during the baseline phase if the intervention has no effect. If there has been a change in the patient’s response pattern because of the intervention, then the proportion of data points (observations) above or below the celeration line will change during the intervention phase. If the proportion of data points above or below the celeration line during the intervention phase is sufficiently different from that occurring during the baseline condition, then a probability statement may be made, indicating that a statistically significant change in response patterns has occurred across the phases.

Figure 8 illustrates the use of the celeration line to demonstrate a change in the slope of postrotary nystagmus durations for a patient receiving sensory integrative treatment. The reader may consult White and Haring (14) for more detailed instructions in computing the celeration line. Ottenbacher (32) has provided an example of the use of the celeration line approach in a clinical research study.

The Standard Deviation Band Method. This technique was originally presented by Shewart (33) and has been described in detail by Gottman and Leiblum (34). The standard deviation band method requires a little more statistical ability than the celeration line approach but remains a relatively easy procedure to compute. To
employ this method, the therapist must first compute a mean for a set of baseline data. Then the values associated with two standard deviations above or below the baseline mean are determined. A horizontal “band” is then drawn on the graph, representing plus and minus two standard deviations from the mean of the baseline. The lines are extended into the intervention phase (see Figure 9). If at least two successive observations (data points) during the intervention phase fall outside the band created by the standard deviation lines, then Gottman and Leiblum (34) argue that a statistically significant change has occurred; this is because the chance of a data point occurring outside the two standard deviation band is less than 5 in 100 if no change took place from baseline to intervention. Figure 9 illustrates the use of the standard deviation band method when applied to the same set of data used in Figure 8. Both statistical procedures detected a significant change in response patterns from baseline to intervention.

Statistical procedures should be interpreted with caution and used as an adjunct to visual analysis. The presence of serial dependency or unusual trends within the data may compromise the results of a purely statistical analysis of data from single-system designs.

Implications of Evaluation for Practice and Research

Until recently the primary professional concern related to research in occupational therapy centered around the need to get it done. The emphasis was on providing support, education, and financial encouragement to therapists who would produce a research literature relevant to occupational therapy. The resources of the professional association were channeled toward the goal of research production. These efforts appear to have been effective, because there has been a noticeable increase in certain types of data-based research reports within occupational therapy literature (35). Now that a research literature is emerging in occupational therapy and methods of disseminating research findings have been developed (36), an expansion of AOTA’s research related efforts is indicated. More emphasis should now be devoted to integrating and assimilating occupational therapy research into clinical practice.

One obvious method of applying research findings to practice situations is to extrapolate the findings from a single study and apply them directly to the clinical setting. While this approach may have a certain degree of direct intuitive appeal, it is not the most desirable method available for synthesizing research and practice. Burr and others (37) refer to this approach as the “empirical method.” They explain that the direct application of results from an isolated research study to a particular clinical setting has some serious limitations. The empirical method is likely to be limited by the absence of an exact fit between patient problems and those reported in the investigation. A second limitation is the general inability of clinicians to replicate precisely the intervention strategy em-
Figure 9.
Example of standard deviation (SD) band method to evaluate changes in postrotary nystagmus (PRN) duration associated with a program of sensory integrative activities.

A, baseline period; B, period during which subject received program of sensory integrative activities.

Theoretical method is described as "the process of going from technical research literature to 'theory' and from 'theory' to practice rather than trying to directly apply research literature to practical situations" (p. 19). Using this approach, a therapist who encounters a clinical problem would translate that problem into a question related to theory and attempt to derive the solution from an analysis of the theory, rather than seeking the answer directly from a specific research study. Clearly, the relevant theory the therapist is relating the clinical problem to must be supported by previous research. An unsupported or unverified theory is of limited clinical use to the practicing therapist. The role of traditional deductive empirical research is congruent with the "theoretical method" described by Burr et al. (37). The researcher employing traditional group comparison methods typically develops a hypothesis or prediction based on some theoretical perspective and then submits the hypothesis to an empirical examination. If the research hypothesis is supported, then the investigator has provided another piece of evidence supporting the theory. If the research hypothesis is not supported, then the theory may have to be modified or revised. This interaction between classical research procedures and theory is well established in many fields in the behavioral and biomedical sciences.

The value of research literature in terms of the theoretical method is obviously related to the evidence it provides for or against a particular theoretical position. Thus, if the results of several research studies support a particular theoretical position, the therapist can
The theoretical method of integrating research and practice clearly illustrates the nexus between theory development, testing, and traditional deductive research. Therapeutic practice based on a synthesis of theory and research provides a firm foundation for clinical intervention that is removed from trial and error intuition. This interplay between theory and research is a complex process, which generally occurs over a prolonged period of time. It also assumes that theories within a field are developed to a level that they can be supported or refuted, and thus begins to provide guidelines for practice (39).

Payton (40) states that “research is important to therapy because it is the major tool available to us for validating our services to people in our role as therapists” (p IX). The preceding discussion illustrated that the general validity of a therapeutic procedure, as applied to a particular patient population, is gradually established over a period of time via the accumulation of research evidence supporting a specific theoretical position. While such generic validation of particular treatment programs is essential to professional development, it is not the only form of therapeutic validity. As clinicians, we must also demonstrate the validity of specific therapeutic procedures as applied to individual patients in isolated clinical environments. The procedures described in this paper are designed to evaluate and measure systematically changes in patient performance on an individual or small-group basis. The procedures are not designed to be inferential or deductive in the traditional sense, and their primary purpose is not to verify theory- or test theory derived hypotheses.

The application of single-system designs and related procedures described in this paper will assist therapists in systematically evaluating individual clinical change and thereby provide a mechanism for establishing therapeutic accountability. Yet, single-system methods cannot be the sole type of research relevant to validating practice claims. As noted, the overall validation of therapeutic programs employed with particular patient populations can only occur through the accumulation of evidence in support of a particular theoretical orientation. The development of such a body of evidence generally requires the use of traditional group comparison (large-N) research designs. For the most part, this is due to the difficulties inherent in extrapolating to other individuals or other clinical environments from single-system designs. One of the major advantages of traditional group comparison procedures employing large numbers of subjects is that the rules of control and generalization are clearly specified.

The single-system procedures may be viewed as ideal for monitoring and evaluating practice effectiveness on a day-to-day individual basis, whereas the more traditional large-N or group comparison procedures may be better suited to demonstrate the generic effectiveness of an intervention program and support or refute the theoretical rationale for an intervention.

The relationship between single-system evaluation and traditional large-N or group comparison methods can evolve into a system that will develop the knowledge and produce the evidence necessary to validate clinical practice. It can be argued that not every therapist should produce traditional group comparison deductive research designed to develop or refine theory or to test theoretically derived hypotheses. However, every clinical therapist does have a responsibility to document the effectiveness of the services provided to individual patients in a systematic and objective manner. The use of single-system designs can be used by clinical therapists to fulfill this responsibility and can also provide initial data for the generation of theoretically derived hypotheses suitable for large-N experimental research.

Evidence from several single-system evaluations of a particular clinical problem may provide a reasonable basis for expecting similar outcomes from a large-N or group comparison research investigation, which could be conducted by academic or other personnel having access to the appropriate resources. The results of traditional experimental research will provide the basis for making decisions regarding the general effectiveness of a particular intervention strategy for a given patient population and will help establish a professional practice founded on valid scientific evidence.

REFERENCES
3. Gillette, NP: Nationally speaking—A


