Discriminative Versus Evaluative Assessment: Some Observations on Goal Attainment Scaling

Kenneth J. Ottenbacher, Anne Cusick

Key words: assessment process, occupational therapy

Goal attainment scaling has been described in the rehabilitation literature as a method of evaluation designed to monitor client progress, structure team conferences, and provide information on overall program evaluation. The appropriate use of goal attainment scaling in clinical environments depends on a clear understanding of its strengths and weaknesses. The primary strength of goal attainment scaling is the ability to evaluate individualized longitudinal change. Goal attainment scaling is not a methodology designed to assess functional status. Confusion occurs when evaluation tools designed to discriminate between persons (determine status) are used to assess change in individual patients over time. This paper describes the strengths and weaknesses of goal attainment scaling as a methodology for evaluating clinical change. The process of evaluative assessment with goal attainment scaling is compared to discriminative assessment where the purpose is to distinguish between persons or to classify persons into predefined categories.

In a recent series of articles, Guyatt and colleagues attempted to classify evaluation procedures according to their purpose (Guyatt, Deyo, & Charlson, 1989; Guyatt, Kirshner, & Jaeschke, 1992; Kirshner & Guyatt, 1985). Kirshner and Guyatt (1985) identified three primary purposes for assessment in health related fields: discriminative, predictive, and evaluative. Discriminative assessment is used to differentiate (or discriminate) between people at a single point in time when no external criterion or gold standard is available. Predictive assessment is useful when the purpose is to classify persons into a set of predefined measurement categories when a gold standard is available. Evaluative assessment determines the magnitude of longitudinal change in an individual or group. Guyatt and associates (1989, 1992) argued that the choices a clinician or researcher makes in selecting an appropriate assessment strategy depend on whether the assessment purpose is discriminative, predictive, or evaluative.

Problems arise in clinical practice and research when assessment strategies and instruments are used for purposes for which they were not originally intended. A common example is the use of an instrument originally designed to discriminate between people at a single point in time being used to evaluate within-subject longitudinal change. Guyatt and colleagues provided detailed information on the characteristics and attributes of discriminative, predictive, and evaluative assessment tools (see Guyatt et al., 1989; and Kirshner & Guyatt, 1985, for additional details).

Review of the criteria provided by Guyatt et al. (1992) shows that most of the instruments currently available in rehabilitation generally, and occupational therapy specifically, are discriminative or predictive. Few assessment strategies being used in rehabilitation include the characteristics of an evaluative instrument as defined by Kirshner and Guyatt (1985). These characteristics include evaluative items that (a) are related to individual functional performance, (b) demonstrate little variation between replicate evaluations, and (c) are responsive to change. An evaluative strategy must also have longitudinal construct validity (see Guyatt et al., 1992; Kirshner & Guyatt, 1985).

Although there are few assessment instruments in the rehabilitation fields that have been specifically designed for the purpose of evaluation as defined by Kirshner and Guyatt (1985), goal attainment scaling is one methodology with a strong evaluative emphasis. As originally proposed by Kiresuk and Sherman (1968), goal attainment scaling is designed to evaluate longitudinal change.

Lewis, Spencer, Haas, and DiVittis (1987) described goal attainment scaling as "a technique for evaluating program effectiveness on the basis of the extent to which individualized patient goals, established at intake, have been achieved upon termination or follow-up" (p. 408).
In a previous paper (Ottenbacher & Cusick, 1990), we presented the details of goal attainment scaling, including an example relevant to occupational therapists. The specific steps involved in completing the goal attainment process will not be repeated here. A brief review, however, is necessary as a background for the examination of the strengths and weaknesses of goal attainment scaling.

Features of Goal Attainment Scaling

Using a format referred to as the goal attainment guide, the therapist (preferably with client and family participation) decides on the expected level of outcome for a particular goal. The expected level of outcome provides a (behavioral) description of the anticipated performance of the client after an intervention program. Outcomes that are both more or less favorable than the expected outcome are also determined for each goal. The expected level of final performance is at the center of the guide. An outcome level somewhat better than expected and a level much better than expected represent scale positions above the expected level; a level slightly worse than expected and a level much worse than expected represent two-scale positions below the expected level. For example, the goal of community participation has been selected for a particular client (see Table 1). The expected level of outcome after intervention is “participates in community activity 1 to 2 times per week.” The guide contains better than expected and worse than expected outcomes (see Table 1). Each level of performance is associated with a numeric value ranging from +2 to 0 with 0 associated with the expected level of outcome positioned at the center of the guide. When at least three goals have been identified and placed into this format, a goal attainment follow-up guide has been completed.

Operational definitions of the outcome criteria (e.g., the operational definition of participates in community activity), and the ability of the examiner to accurately and reliably record the various levels of outcome must be established before using goal attainment scaling in any evaluative study.

Relative weights are assigned to each of the three or more goals identified for the client in the goal attainment scaling scoring process. There is no standard procedure for determining how each goal is weighted. Assigning a weight is ideally achieved by consensus among the client, therapist, teacher, family members, and other persons concerned with the client’s performance. Generally the weight simply reflects a prioritizing or ranking of the goals. If four goals are prioritized, the most important goal might be assigned a ranking (weight) of +4 and the least important goal a weight of +1. The weights must be determined in the goal planning stage and not in the analysis phase. If goals are assigned weights during the analysis, the weights might reflect the priorities of the clients or therapists as they look back on the outcome of the program and consider its strengths and weaknesses. This situation would introduce the possibility of systematic bias into the evaluation process.

After the termination of the treatment program, the progress toward achieving the outcomes selected for each goal is determined. Ideally, this determination is done by an independent examiner without previous involvement in the treatment program or goal setting process. Using the sample goal in Table 1, the examiner might find that after intervention the client was participating in community activities 3 to 4 times per week, his or her level of attainment on that goal would thus be +1 (greater than expected outcome). After the examiner determines the client’s status on each goal, this information is used to calculate a goal attainment score using the formula originally proposed by Kiresuk and Sherman (1968):

$$T = 50 + (10 \sum W_i X_i) / \sqrt{(1-r) \sum W_i^2 + r (\sum W_i)^2}$$

where $X_i$ represents the outcome score for each behavior (a value from $-2$ to $+2$) and $W_i$ represents the weight assigned to a particular goal. The $r$ value in the formula reflects the estimated average intercorrelation between outcome scores. Kiresuk and Sherman (1968) argued that an $r$ value of .30 can safely be assumed and used as a constant in the formula.

A final goal attainment scaling $T$-score of 50 corresponds to the zero point on the guide, that is, the expected level of outcome (see Table 1). A $T$-score of greater than 50 represents performance above the expected level and a $T$-score of less than 50 reflects performance below the expected level. Even when clients differ considerably in the nature of the goals they are attempting to achieve (which is the usual situation in rehabilitation), the goal attainment score allows comparison of a client’s relative success in achieving his or her unique series of goals.

### Table 1

<table>
<thead>
<tr>
<th>Predicted Attainment</th>
<th>Goal: To increase participation in community activities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Most unfavorable outcome likely (-2)</td>
<td>Participates in no community activities and is unable to identify 5 appropriate community activities</td>
</tr>
<tr>
<td>Less than expected outcome (-1)</td>
<td>Participates in no community activities but is able to identify 5 appropriate community activities</td>
</tr>
<tr>
<td>Expected level of outcome (0)</td>
<td>Participates in community activities 1 to 2 times per week</td>
</tr>
<tr>
<td>Greater than expected outcome (+1)</td>
<td>Participates in community activities 3 to 4 times per week</td>
</tr>
<tr>
<td>Most favorable outcome likely (+2)</td>
<td>Participates in community activities 5 times or more per week</td>
</tr>
</tbody>
</table>

Note: Operational definitions for participates and community activities must be developed by the therapist (and client or family members).
Applications of Goal Attainment Scaling

Goal attainment scaling has been applied in a variety of clinical environments where the purpose of assessment has been evaluative, as defined by Guyatt and associates (1989). Recently, several applications of goal attainment scaling have appeared in the rehabilitation literature (Maclec, Smigielski, & DePompolo, 1991; Palisano, Haley, & Brown, 1992; Stephens & Haley, 1991). For example, Maclec et al. (1991) used goal attainment scaling to evaluate work and related outcomes for persons with traumatic brain injury enrolled in a comprehensive rehabilitation program. The authors reported that the "results of this study and case analyses support goal attainment scaling as a quantifiable, individualized measure that is useful for (1) monitoring client progress, (2) structuring team conferences, (3) ongoing rehabilitation planning and decision making (4) concise relevant communication to family referral services and funding sources, and (5) overall program evaluation when used in the context of other objective outcome measures" (p. 138).

Advantages of Goal Attainment Scaling

The goal attainment scaling system is a flexible set of procedures for evaluating change in individual (or group) performance in a variety of areas. An advantage of the system is that it is not bound to any theoretical orientation or particular type of treatment or outcome assessment tool. Another strength is that goals can be individualized and are specifically designed to represent realistic expectations concerning client performance. Along this line, the goal attainment scaling strategy actively encourages cooperative goal setting. Input from clients, family members, and other health service providers is important in establishing and prioritizing the goals in determining realistic levels of expected performance.

In evaluating the advantages of goal attainment scaling, it is important to recognize that it is not a methodology designed to determine the status of clients relative to a particular trait of interest, such as gross motor performance or activities of daily living (ADL). Rather, goal attainment scaling is a set of procedures designed to evaluate change. The focus of goal attainment scaling is on process (evaluating change) not product (determining status).

Despite the positive aspects of goal attainment scaling, it is not a panacea for the difficulties encountered in clinical documentation and evaluation. The procedures may be misused or misinterpreted and therapists using goal attainment scaling must be aware of its limitations.

Cautions Concerning Goal Attainment Scaling

Goal attainment scaling was not developed as a traditional research methodology. The intent of the process is not to establish causal inferences between independent and dependent variables. Goal attainment scaling represents a set of evaluative procedures designed to assist professionals in assessing client change. The procedures associated with goal attainment scaling can, however, be incorporated into many standard research designs. If goal attainment scaling is used in a traditional research context, attention must be devoted to ensuring the internal validity of the investigation by randomly assigning clients to groups, blindly recording the level of outcome, and establishing the reliability of the dependent outcome criteria. Lewis and associates (1987) provided valuable suggestions and guidelines to therapists interested in using goal attainment scaling in a research context.

Concerns and criticisms of goal attainment scaling generally fall into two major categories: (a) conceptual problems and disagreements related to the application of psychometric principles to goal attainment scaling (Crytenbaum, Ginath, Birdwell, & Brandt, 1979; Goodyear & Bitter, 1974), and (b) practical difficulties resulting from the conceptual concerns. Issues associated with the conceptual limitations of goal attainment scaling are reflected in the controversy regarding whether the purposes of assessment are discriminative or evaluative.

Many of the attempts to examine the psychometric properties of goal attainment scaling (reliability, validity, etc.) consider it as an assessment procedure intended to evaluate a person's status or ability in a particular area of interest. For example, Stephens and Haley (1991) reported a study in which the motor function of children was evaluated with goal attainment scaling procedures and scores on the Peabody Developmental Motor Scales (Folio & Fewell, 1983). Such attempts assume that the two assessment methods are testing the same construct, in this case fine and gross motor function. The ability of goal attainment scaling to test a construct, such as gross motor function, will vary according to the selection of individual items and their relative weighting. In contrast, the items from a standardized, norm-referenced test are fixed and cannot vary across individual subjects or examiners. No test necessarily tests the claimed underlying construct—this idea differs from the idea that goals vary among subjects but tests do not.

If the goal attainment scaling score is to be considered a valid test of client status or ability, then correlations with scores of other status or ability oriented (discriminative) assessment tools are of psychometric interest. If, on the other hand, goal attainment scaling is viewed as an evaluative assessment designed to evaluate change over time, then the correlations of interest for testing reliability and validity are those with other defensible assessments of individual longitudinal change. Comparing goal attainment scaling scores, which are based on individualized goals that have been uniquely weighted, with standardized tests reflecting normative performance (i.e., using the same items for every subject) will result in very low correlations.
Relevance and Reliability of Selected Goals

Several authors have identified concerns about the goals identified as part of the goal attainment scaling process (e.g., who should be responsible for stating them, how they should be weighted). This issue can be considered a question of validity, that is, are the problems conceptualized by a committee or team relevant for the client? Or, alternatively, this issue can be considered a question of reliability. Will two teams or different therapists identify the same goals and assign the same weights for a given client (see Cytrynbaum et al., 1979; Seaberg & Gillespie, 1977)? These reliability and validity concerns reflect the normative approach generally used in developing standardized assessments designed to make between-subject discriminative comparisons. Discussions and criticisms of goal attainment scaling based on traditional psychometric assumptions associated with discriminative assessment do not reflect the evaluative nature of the goal selection and development process that is central to the goal attainment scaling methodology. From the perspective of an individualized or evaluative process, the selection of goals and their scaling is more a sampling issue than one directly associated with validity and reliability. A client receiving intervention generally presents several different problems (potential goals). Each problem area may be conceptualized into scaled goals in a variety of different ways. For example, a client who has experienced a cerebrovascular accident and lost function in her right upper extremity may not be able to either dress or feed herself. One team (therapist) may identify dressing as the goal of interest, another team (therapist) may identify feeding as the problem that should be addressed. Content analysis would find the goals dissimilar and the reliability would be low. Both goals, however, are addressing the same problem domain, activities of daily living.

From the perspective of standardized, normative (discriminative) assessments, the construction of identical goals across raters is an important psychometric requirement. From the evaluative perspective, however, the construction of similar goals and weights across raters is not of primary importance. In fact, insistence on identical goal construction eliminates the ability of goal attainment scaling to provide evaluative assessment that is individualized to the client’s unique situation and needs. The issue of reliability is an important one in relation to the actual evaluation of client performance. Estimates of the interrater reliability should routinely be computed and reported. Because the goals are individualized, the interrater reliability of assessing performance must be established for each distinct goal. The process of establishing interrater reliability for the outcome measures used in goal attainment scaling is similar to that required by other evaluative procedures with an individual focus (e.g., single-subject designs).

Another important issue related to the collection of outcome data with the goal attainment scaling approach is how the data will be collected. To reduce bias and help insure the integrity of the goal attainment scaling score, it is important that the evaluation of the progress in achieving goals be done blindly. That is, the person collecting the final outcome information should not be aware of the client’s initial status or of the group to which a client was assigned (if groups are used in the study).

Computation of the Goal Attainment Scaling Score

In their review and critique of the goal attainment scaling methodology, Seaberg and Gillespie (1977) identified a problem associated with assigning numeric values to levels of performance. The manner in which the values (-2 to +2) are assigned indicates that they are ordinal level data. The problem of analyzing ordinal data with parametric level statistical procedures is not unique to goal attainment scaling. This issue has been widely discussed and debated in the behavioral sciences literature. Recently, the issue has been discussed in relation to functional assessments in medical rehabilitation (Merbitz, Morris, & Grip, 1989). Kiresuk and Sherman (1977) agreed that the goal attainment scaling process produces ordinal data, but they argued that this level of data is not dissimilar to that produced by Likert or other ordinal scaling procedures commonly analyzed with parametric statistical techniques. The controversy regarding ordinal versus interval data levels of measurement and data analysis represents an issue that is much larger than the computation of goal attainment scaling scores and has not been clearly resolved either philosophically or statistically (Gaito, 1980; Gardner, 1975).

Another computational issue related to goal attainment scaling concerns the T-score itself. In developing the T-score formula, Kiresuk and Sherman (1968) emphasized that it represents a within-subject standardization. That is, the distribution reflected by the T-score is not that of the trait being assessed, but of the change in performance (i.e., change on that trait within the subject).

If the scale is correctly constructed, this change in performance is evenly distributed with a mean of 0 and a standard deviation of 1 (assuming the -2.0 to +2.0 scaling). This approach was required because the goal attainment scaling is an individualized evaluation approach. The within-subject standardization enables comparisons in dimensionless units across subjects. The T-score proposed by Kiresuk and Sherman (1968) is based on location and scale standardization, not normative standardization. Heavlin, Lee-Merrow, and Lewis (1982) noted that “by location and scale standardization, what is meant is relativization of a [score] against the dispersion it is likely to assume by chance—a T-score transformation” (p. 231). Kiresuk (1973) reported that, on the basis of the
evaluation of more than 900 goal attainment scaling assessments averaging four goals (scales), a symmetrical distribution for the scale had been found (mean T-score = 50, SD = 10).

The final T-score obtained from the goal attainment scaling process reflects not only the change in performance for a goal, but also the accuracy of the prediction made at the beginning of intervention. If the various goals are not realistically scaled so that the expected level of performance is not equal to 0 and the standard deviation is not equal to 1, then the mean and standard deviation of the T-score may be considerably different from the anticipated values. Kiresuk and Sherman (1968) noted that this will not affect comparisons of T-scores in studies where subjects are randomly assigned to conditions, but it may influence examination of T-scores that are compared longitudinally.

Evaluative studies are likely to make longitudinal comparisons. To address this problem Cytrynbaum et al. (1979) suggested that extensive training may be necessary to achieve adequate interrater reliability in some areas of goal setting and assessment. For instance, the community participation goal presented in the previous example (see Table 1) may present problems in the area of interrater reliability if raters are not properly trained, or if enough operational detail is not provided. The reliability problem in assessing client performance is not unique to goal attainment scaling. Whenever judgments are made in relation to human performance, care should be taken to ensure that the judgments are as consistent (reliable) and accurate as possible. Reliability may be improved by including multiple evaluation periods, a defined training program for raters, and explicit definitions or examples of client performance. Detailed procedures to determine the reliability of observations and assessments are beyond the scope of this paper. Methods of enhancing observer reliability are widely available in the applied behavioral science research literature (Hartman, 1982; Webb, Campbell, Schwartz, & Sechrest, 1966).

A second area of concern identified by Cytrynbaum et al. (1979) is the possibility of floor effects in the goal attainment scaling evaluation process. If the expected level of performance after intervention is assigned a 0 value on a scale ranging from -2 to +2, it is probably safe to assume that many clients will begin the program at a lower level, i.e., -2. When this happens there may be no room for regression. The worst possible outcome for the client is that he or she remains at the -2 level during the final evaluation. Such a finding would suggest that the client's status has remained unchanged during the period of intervention. No allowance is made for clients who may have regressed during the time the program was in effect. Although this is not a major problem with goal attainment scaling, it is a limitation that therapists should be aware of, particularly if they are working with a client population where deterioration in performance is expected.

Conclusion

Despite the concerns described above, goal attainment scaling represents an improvement over the subjective and anecdotal evaluations that typify clinical evaluative assessments. Goal attainment scaling is a flexible evaluation methodology that can address the documentation and accountability concerns facing health care providers, including occupational therapists. Perhaps most important, goal attainment scaling is a method that is practice based and practitioner oriented. The noted evaluation methodologist Daniel Stufflebeam once observed that "the purpose of evaluation is to improve, not to prove" (Isaac & Michael, 1985, p. 1). When properly applied, goal attainment scaling is an evaluation method that can improve the clinical services provided by occupational therapists.

References


The American Journal of Occupational Therapy


