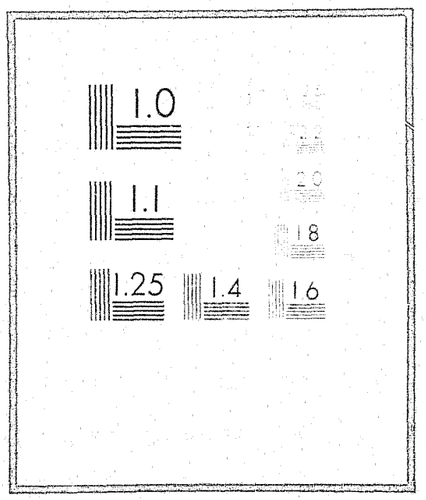

Center for Social Welfare Research
School of Social Work
University of Washington

Center for Social Welfare Research
School of Social Work
University of Washington

by

James R. Seaberg

David F. Gillespie

Goal Attainment Scaling: A Critique and an Alternative*

April, 1976

*This research was supported by a grant, 90-C-430, from the National Center on Child Abuse and Neglect, Children's Bureau, Office of Child Development, Office of Human Development, Department of Health, Education and Welfare.

GOAL ATTAINMENT SCALING: A CRITIQUE AND AN ALTERNATIVE

ABSTRACT

Goal Attainment Scaling is a recently proposed procedure for measuring the outcome of mental health services. It has come into wide application without the benefit of critical review from outside the group who developed it. Such a review is a central theme of this article. Serious problems with the GAS procedure include basic conceptual problems, prediction statement problems, and computational problems. An alternative procedure for Individual Problem Rating (IPR) is presented with a discussion of several issues which must be studied in detail prior to widespread adoption of the alternative procedure.

GOAL ATTAINMENT SCALING: A CRITIQUE AND AN ALTERNATIVE*

Goal attainment scaling (GAS) was first reported by Kiresuk and Sherman (1968) as a new method for evaluating mental health programs. GAS was developed in response to the problems associated with standardized measures of mental health adjustment (Kiresuk and Sherman, 1968: 443-445). The objective sought was a measurement procedure suitable for evaluating programs but also sensitive to the peculiar mental health needs of individuals on an individualized basis. The response has been remarkably favorable; Kiresuk (1973:16) reports that over 70 programs have adopted GAS. The extensive use of GAS can be partly attributed to the massive dissemination efforts of the Hennipen Program Evaluation Project (HPEP), and partly to the individualized approach characterizing GAS. It is apparent that contemporary evaluators and clinicians alike favor an assessment tool which is geared to the particular problems of program clients.

In spite of its widespread adoption, however, reports on GAS results outside of HPEP are limited and the only critical assessments of GAS known to these authors are unpublished (Barlow and Ravneberg, n.d.; Clayton, 1975). These points suggest that perhaps there has not been enough time elapsed for clinicians and evaluators to gain sufficient experience with GAS to report or publish their findings. In view of this situation, it seems a critical assessment of GAS is timely.

The first part of this paper, therefore, will provide a critical overview of GAS in terms of its viability as a program evaluation tool. The

---

*This research was supported by a grant, 90-C-430, from the National Center on Child Abuse and Neglect, Office of Child Development, Office of Human Development, Department of Health, Education and Welfare.

critique will seek to answer two questions: (1) do the assumptions of GAS meet the accepted conventions in psychometry and sociometry, and (2) can GAS be reasonably used in the evaluation of large-scale demonstration programs such as the current set of national demonstration programs aimed at affecting the child abuse and neglect phenomena.* As will be demonstrated in the following sections, our assessment results in negative answers to both of these questions. The second part of this paper thus presents an alternative individualized measurement procedure for evaluating social service programs.

An Overview of Goal Attainment Scaling

The major elements of the Goal Attainment Scaling procedure include: (1) selection of an individual's problems for which improvement is desired, (2) the assignment of weights to each problem in the set, reflecting the relative importance of each problem (or alternatively each problem may be considered to be of equal importance), (3) the development of scales for each problem which are essentially behavioral referents of expected change (goals) in the problem areas for a given time interval, along with referents for more and less than expected change. These three steps are organized into a format called a "Follow-up Guide." Progress on the problems is assessed at given intervals (often three months). Each of the five levels of each scale is given a standard numeric score and at follow-up scores for each problem are combined to derive a GAS score for the individual.

Of course, there is much greater detail to the procedure and many subtle nuances to its understanding and implementation which can only be gained from

*There are currently 36 demonstration projects underway nationally sponsored by the National Center on Child Abuse and Neglect, Office of Child Development, all of which will be evaluated to determine their effects on various aspects of the child abuse and neglect problem.

consulting the original sources (e.g. Kiresuk and Sherman, 1968; Kiresuk and Garwick, 1975; Garwick, 1974c).

Unfortunately, it is not possible to address Goal Attainment Scaling as a fixed entity. As it has been used over the past few years, it has been modified. For example, as originally presented patient goals were set by a committee, but several variations on goal setting have been tried since that time. Likewise, originally, patients were randomly assigned to treatment modalities, but this is no longer practiced (at least at the HPEP). A critical assessment of a procedure which is in such a state of evolution is complicated by the transiency of the subject. There are, nonetheless, a number of features about the more stable elements of GAS which can be examined.

Validity and Reliability

Fundamental to measurement theory is the necessity of establishing, empirically if possible, the validity and reliability of the measurement device. Validity refers to the relevance of the device, reliability to the precision of the device. The validity and reliability of Goal Attainment Scaling are, then, a major concern for persons or organizations considering adoption of this measurement procedure.

The only statements on the validity of GAS known to the authors are those of Garwick (1974a) and Mauger (1974). In his report, Garwick presents an argument for the "construct validity" of GAS, the basic construct underlying GAS being the "outcome" or "attainment of expectations" (1974a:5). A variety of arguments and results of empirical study are presented to support the notion of construct validity. These basically take the form of posing hypotheses about factors which might account for variation in GAS scores.

The list of such factors, which is lengthy, is admittedly incomplete and the data available to investigate the possible effects of these factors are sketchy. As Garwick notes, "these and other theoretical considerations imply that any summary comments about the construct validity and Goal Attainment Scaling should be cautious" (1974a:8). We concur.

Such caution is further warranted by the findings of Mauger (1974) that GAS scores and GAS change scores correlated only slightly with MMPI mean-change scores (.285 and .306 respectively). This suggests rather low "concurrent" validity. But Garwick apparently dismisses this kind of empirical result by noting that ". . . Goal Attainment scores are not intended to have a particularly high correlation with other treatment outcome measurement devices, since Goal Attainment Scaling is such a radically different evaluation system (1974a:11)." The case for this argument is not clear.

Two other issues arise in relation to the validity of GAS scores. As mentioned above, the issue of which problems are to be assessed and who should state these is problematic, and thus can be thought of as a validity question. Are the problems conceptualized by a therapist or a committee relevant for the client? Are the problems as conceptualized by a client relevant for the therapist who will be attempting to assist the client in modifying these problems? That there is tension in these conceptualizations has been pointed out in one study. The study compared client-prepared versus clinician-prepared follow-up guides, and a significant difference in the GAS scores resulting from these two guides was observed with the client-prepared guides resulting in much higher scores (Garwick, 1974a:9). But the question still remains as to whose conceptualization of the problems is most relevant for program evaluation. The second issue on the validity of GAS involves the movement toward semi-standardized scales (Sherman, et al., 1974:8-9).

Without going into great detail, such movement seems to be away from the original source of validity which was based upon a highly individualized procedure. Both of these latter two issues will be further developed in subsequent sections.

The reliability of GAS is reported in a summary statement prepared by Garwick (1974b) as well as in a study by Sherman, et al. (1974). Reliability in the case of GAS is argued to be a matter of inter-observer reliability; that is, agreement between two or more raters of the same client concerning follow-up guide construction and client level of functioning. Garwick's (1974b:7) report revealed an interesting feature concerning the effects of the number of problems (scales) designated and the correlation of individual scale scores with the total follow-up guide scores: As the number of scales increased, the mean scale scores, the mean goal attainment scores, and the correlation coefficient decreased. From the Sherman, et al. (1974:7-8) study it was estimated that 18% of the variance was due to follow-up interviewer errors in scoring or observation; 17% was due to choice of follow-up guide material; 15% was due to short term client changes or follow-up bias fluctuations; and 50% was due to client long term deviation from expectation. The reliability coefficient was .57 which is modest at best (Nunnally, 1967:226). This was considered reasonable reliability, however, "Given the severity of the test and the unique advantages of GAS . . . (1974:8)."

The reported validity and reliability studies even though not conclusive and not consistently encouraging are based on a device whose basic assumptions are questionable. To the extent that one finds disparity between these assumptions and the actual device, current validity and reliability statements are negated. As already alluded to we perceive a number of problems with Goal Attainment Scaling, conceptually and operationally. We turn now to

a review of these problems and, then, propose a strategy which may lead to an alternate form of individual goal setting and assessment which is free of these problems.

PROBLEMS WITH GOAL ATTAINMENT SCALING AND ITS APPLICATION

There are four basic types of problems with Goal Attainment Scaling and its application: (1) conceptual problems entail the ambiguous meanings attached to the GAS conceptual framework; (2) prediction statement problems include the issues of (a) who states the problems, (b) how many problems should be formulated, and (c) what level of abstraction should be used in stating problems and in preparing scale referents; (3) computational problems refer to (a) the fuzzy assignment of weights indicating problem area importance, (b) the use of numeric values representing equal intervals on the Goal Attainment scales, and (c) the practice of aggregating different scales; and (4) evaluation design problems deal with the difficulties of achieving conclusions concerning program treatment effects. The following discussions highlight the nature of these problems and the difficulties they pose in using GAS as a means to evaluating social service programs.

Conceptual Problems

Perhaps the most fundamental source of confusion in understanding and using GAS is the loose equivalency given to a variety of concepts and constructs which include "goal definition and measurement," "outcome," "attainment of expectations," "specific predictions for a series of outcome levels," and so forth. It is not possible for us to speak of the concept or construct upon which GAS is founded but, as outside observers, we can indicate which construct emerges most strongly from the literature. We concur with Clayton

(1975), among others, who has observed that GAS is based on the ability of a selector to identify client problems and predict the level of functioning for the client on those problems within a given time interval, assuming certain types of therapeutic input. This is somewhat affirmed by Garwick (1974) when he identifies "attainment of expectations" as the construct underlying GAS. The point to be made, however, is that assessing the accuracy of predicted levels of functioning is conceptually different from the specification of a goal or set of goals and measurement of progress toward those goals. This point will emerge more sharply when we discuss computational problems.

Prediction Statement Problems

There are several problems which attend the statement of predicted outcomes in the GAS format. The first is who should conceptualize the problems for which predictions are to be made. The GAS procedure has varied; sometimes it is an intake screener or committee, other times the therapist, and on occasion it is the client. At least one study, as mentioned above, revealed a significant difference between GAS scores for therapist versus client generated follow-up guides (Garwick, 1974a:9). Obviously there are deficiencies in either the therapist or the client as the sole source of the conceptualization of the problems to be evaluated. In this regard, Kiresuk and Sherman (1968:450) have noted that ". . . therapists are biased . . . for particular modes of therapy, prefer to deal with certain kinds of patients and problems, tend to conceive of their role and purpose in ways that will emphasize certain problem areas and exclude others." On the other hand, the clients may be in such a state of confusion, agitation, depression, or of limited mental capacity that they can neither clearly conceptualize nor articulate their problems.

If the problems are selected by a committee additional problems arise; namely, that the problems identified may be different from those treated by the therapist. Here Kiresuk and Sherman (1968:451) assume ". . . that if the goal selectors and the therapists are all well trained, reasonable professionals, the goals that would be chosen would be reasonably comparable." But this is inconsistent with their recognition that "one goal selector may perceive a patient's problems in terms of intrapsychic symptoms and psychodynamics while another may see them in terms of his relationship to others." The inconsistency, however, is presumed to be more apparent than real because it is further assumed that therapy promotes a "general therapeutic effect" (Kiresuk and Sherman, 1968: 451). The general therapeutic effect implies that improvement in one problem area carries over to other problem areas not directly related to the therapy. A general therapeutic effect may exist, but the nature and extent of such an effect is an empirical matter which has yet to be assessed. One piece of evidence suggests this may not be the case; specifically, recall the above reference to the noted tendency for the mean scale scores and mean GAS scores to decrease consistently as the number of problems increased (Garwick, 1974b:7).

A second major problem in the statement of client problems is the question of how many problems should be stated and, correspondingly, from the universe of problems which set should be stated. Although the procedure allows for the statement of as many problems as necessary, in practice there appear to be usually no more than five stated problems. There is a high degree of subjectivity at this stage of the procedure. Clearcut criteria for determining which problems should be the focus of evaluation are absent. Guidelines such as: "The most significant, relevant problem area should be selected for inclusion," ". . . specification of the major areas where change

would be feasible and helpful. . ." (Kiresuk and Garwick, 1975:3), and so on, are given. Significant and relevant may be quite different from feasible. The previously mentioned finding that client developed follow-up guides resulted in significantly higher scores (Garwick, 1974a:9) may be an indication that clients specify problems that are more relevant or more feasible or both, but we cannot be sure which. Problems reflecting expected outcomes which are obtainable in the short term might be a criterion of equal importance, but there is considerable room for subjective interpretation of the criteria.

The procedure for the setting of referents (goals/expectations/predicted outcomes) by which progress or deterioration in the level of functioning is assessed involves a high degree of subjectivity. The criteria for setting the referents are less than definitive. Kiresuk and Garwick (1975:5) state: "The expectations ought to be pragmatic, so that the expected level of each scale reflects what outcome actually 'could' be attained by the follow-up data, not necessarily what 'should' be attained." Elsewhere the directive to use ". . . any form of objectively determinable event" (Kiresuk and Sherman, 1968:447) is given. Of course, a bias toward stating too easily attainable goals could pervade this procedure regardless of who set them. Among intake interviewers and therapists this is apparently not a significant problem since Sherman, et al. (1974:7) demonstrated a reliability coefficient between these groups on follow-up guide construction of .83. However, this probably reflects agreement between persons with essentially the same training and experiential background (professional socialization). Agreement between clients and persons of a clinical orientation is apparently not as great given the score differences observed when comparing client versus clinician prepared follow-up guides (Garwick, 1974a:9). It should be noted that our

critique of problem selection and referent selection may be less serious than portrayed here, if the primary source of variation in these aspects of the procedure can be attributed to the person doing the selection. If this is the case, then stabilization of the source of problem selection and referent selection should be of primary concern in development of client goal oriented measurement techniques.

Computational Problems

If the basic procedural elements of a measurement technique are less than exacting, even the purest computational procedures for generating scores become meaningless. In this instance, we find portions of the computational elements of Goal Attainment Scaling to rest on faulty assumptions. The first problem of this order involves the assignment of weights to indicate the relative importance of different selected problems. The weights need not total any fixed value and they need not even be assigned if the relative importance is not clear (Kiresuk and Sherman, 1968:447). Intuitively, the assignment of weights is a reasonable notion, but on close inspection considerable ambiguity is revealed.

The major problem here is that the criteria for judging importance are not specified. Differential importance could be ascribed to different problem areas according to a variety of criteria. For example, "importance" might be (a) the estimated difficulty for a particular client to resolve the problem, (b) the correspondence between problems selected and therapies available, (c) a rank order of social desirability based upon clinician values or community norms, (d) acknowledged importance to the client, (e) the length of time that a problem has persisted, (f) importance to client relatives or significant others in the client's environment, and so on. It is unlikely that the use

of different criteria of importance would result in the same rank order of importance for a given problem set. It can be surmised, moreover, that different selection sources and perhaps the same selection source at different points in time use various combinations of these criteria. The consequence of different or shifting criteria is that it reduces reliability; if weights are used scores will vary depending upon the weights attached, and the weights vary according to the criteria. Our perceived tendency in the application of GAS to eliminate assignment of weights is probably wise. As noted by Nunnally (1967:543) ". . . weighted and unweighted summative scores usually correlate very highly." And, a classic study by Likert (1932) produced a correlation of .99 in comparing weighted and unweighted scale scores.

Another computational problem involves the assignment of numeric values to each of the referents indicated for each client problem. The manner in which these values are assigned and treated mathematically clearly indicates an assumption that the distance between the referents is equal for each problem specified and across all problems specified. This assumption of interval level measurement seems totally unwarranted and at best the procedure meets the assumptions of ordinal level measurement. The interval level assumption would require assuming, for example, that the distance between "dating/petting" and "some satisfactory intercourse" is the same as the distance between "some satisfactory intercourse" and "regular dating/regular satisfactory intercourse/ marriage" (Kiresuk and Sherman, 1968:446). If one is persuaded that Goal Attainment scales are not interval level data, then another computational problem follows forthwith; namely, the formulas for calculating goal attainment scores (Kiresuk and Sherman, 1968:448-449) apply improper mathematical operations to ordinal data. Moreover, this violation

of mathematical propriety is magnified as the scoring procedure moves from single problem "scale" scores to "follow-up guide" scores, and on to aggregated scores for entire service programs.

In commenting on the problems of aggregation for GAS scales, Garwick (1975:2) has observed ". . . since the aggregate's components are of vastly different scales, the psychometric (as opposed to statistical) questions are numerous. . . there are enigmas and mysteries everywhere." We agree with Garwick, except even more than this, it is our judgment that along with the unresolved psychometric issues, there are also some major conceptual and statistical problems to be resolved.

Evaluation Design Problem

While not a problem of the Goal Attainment Scaling measurement procedure per se, we draw attention to what might be a problem resulting from misunderstanding of the proper inferences which can be drawn about the effects of counseling based on the "design" recommended by Kiresuk and Sherman (1968:445). They recommend a random assignment to different treatment modalities as a means of determining the relative effects of those modalities. This is quite properly stated. Once again, however, the lack of conceptual clarity in the discussions of GAS might pose a problem. Terms such as "outcome," "change," and "expectations" are interspersed with enough frequency that we must be careful to guard against the temptation to slip into inferences about the absolute effectiveness of treatment, i.e. inferences based on a comparison of treatment versus no-treatment. The Kiresuk and Sherman "design" does not have a no-treatment group and therefore inferences about the absolute effectiveness of treatment cannot be made.

In later revisions of their recommended procedure, the random assignment to treatment modalities was dropped. A change score was calculated contrasting client functioning at time-one with that at time-two. Of course, change scores derived from before-after comparisons are subject to questions of internal validity based on the substantial number of viable alternate hypotheses (other than treatment effects) as to the cause of the observed change (Campbell and Stanley, 1963:7-12). This problem is somewhat lessened if single-subject designs such as reversal (ABAB) or multiple-baseline are used (Howe, 1974). The latter are not part of the recommended "design" by Kiresuk and associates, yet Goal Attainment Scaling is described as an evaluation device.

DISCUSSION

We initially set out to assess the utility of Goal Attainment Scaling as a measurement procedure which could be used in evaluating the effects of large-scale social service demonstration programs. To do this, a careful examination of the quality of the measurement procedure was requisite. As discussed above, there are many problems with Goal Attainment Scaling as presently constituted. Indeed, it has little to recommend it as a program evaluation tool. Thus we will not discuss any of the issues involved in its adaptation as a measurement device for large-scale evaluation; for example, the extent and frequency of training in its use in order to attain adequate reliability.

The idea of a measurement device based on individualized client problems, however, is both appealing and challenging as a direction for overcoming the measurement and evaluation problems noted by Kiresuk and Sherman (1968:443-445). Stimulated by the work of Kiresuk and his colleagues in developing

and testing Goal Attainment Scaling, and also our critical review of the procedure, we have undertaken the conceptualization of an alternative measurement procedure which we believe overcomes some of the inadequacies of GAS. The procedure we propose is at this time untested in the field and, therefore, represents a hypothesized device. Various plans for testing this new procedure are suggested as part of its presentation.

AN ALTERNATIVE INDIVIDUALIZED MEASUREMENT PROCEDURE

The measurement procedure proposed here is based upon the concept of individual (or client) problems rather than goals. The goal concept has inspired researchers for many years, but the problems attending its use have overshadowed its potential. Although client problems contain many of the same pitfalls as goals, by limiting our attention to only client problems, we at least reduce the number of conceptual and methodological difficulties. Moreover, goals are intended or desired future states (to be contrasted with "predicted states") and, as such, they are essentially implicit in the statement of a problem, i.e. remove or lessen the problem to a tolerable degree. At the same time there is a considerable tradition in psychometric theory and practice which supports the notion of persons being able to rate themselves on attitudes, behavior, performance, etc. It is thus reasonable to ask persons to identify and rate their individual problems. First we will present the proposed procedure, then discuss several issues which must be carefully researched before it can be adopted for evaluation purposes.

Individual Problem Rating (IPR)

The Individual Problem Rating is based on two primary assumptions: first, that individuals referred for social services are able to describe the

problems which led to the referral (self or otherwise), and second, that individuals can distinguish among their problems in terms of the relative importance of each problem to them. IPR requires that individuals do the following: (1) in the course of an intake interview, specify the problems for which they would like help; (2) for each problem listed, rate the severity of the problem* on a scale of numeric values from 1 for "not at all severe" to 100 for "extremely severe;" and (3) among the set of problems listed assign weights of importance* as portions of 100 percent which total 100 percent. A basic IPR score is calculated by the formula:

$$\text{IPR score} = \frac{PS_1PI_1 + PS_2PI_2 + \ldots + PS_iPI_i}{N_p 100}$$

where: PS = severity of problem
PI = importance of problem
N_p = number of problems listed

In Figure 1 an example is cited.

Figure 1

Example of Rating of Individual Problems

Problem/s	Problem Severity (1=not at all severe to 100=extremely severe)	Problem Importance (portion of 100 percent)
1. (fictitious problem)	87	55
2. (fictitious problem)	39	30
3. (fictitious problem)	48	15

$$\text{IPR score} = \frac{(87)(55) = (39)(30) + (48)(15)}{(3)\ 100} = 22.25$$

*With appropriate instructions to treat the distance between each interval as equivalent.

and testing Goal Attainment Scaling, and also our critical review of the procedure, we have undertaken the conceptualization of an alternative measurement procedure which we believe overcomes some of the inadequacies of GAS. The procedure we propose is at this time untested in the field and, therefore, represents a hypothesized device. Various plans for testing this new procedure are suggested as part of its presentation.

AN ALTERNATIVE INDIVIDUALIZED MEASUREMENT PROCEDURE

The measurement procedure proposed here is based upon the concept of individual (or client) problems rather than goals. The goal concept has inspired researchers for many years, but the problems attending its use have overshadowed its potential. Although client problems contain many of the same pitfalls as goals, by limiting our attention to only client problems, we at least reduce the number of conceptual and methodological difficulties. Moreover, goals are intended or desired future states (to be contrasted with "predicted states") and, as such, they are essentially implicit in the statement of a problem, i.e. remove or lessen the problem to a tolerable degree. At the same time there is a considerable tradition in psychometric theory and practice which supports the notion of persons being able to rate themselves on attitudes, behavior, performance, etc. It is thus reasonable to ask persons to identify and rate their individual problems. First we will present the proposed procedure, then discuss several issues which must be carefully researched before it can be adopted for evaluation purposes.

Individual Problem Rating (IPR)

The Individual Problem Rating is based on two primary assumptions: first, that individuals referred for social services are able to describe the

problems which led to the referral (self or otherwise), and second, that individuals can distinguish among their problems in terms of the relative importance of each problem to them. IPR requires that individuals do the following: (1) in the course of an intake interview, specify the problems for which they would like help; (2) for each problem listed, rate the severity of the problem* on a scale of numeric values from 1 for "not at all severe" to 100 for "extremely severe;" and (3) among the set of problems listed assign weights of importance* as portions of 100 percent which total 100 percent. A basic IPR score is calculated by the formula:

$$\text{IPR score} = \frac{PS_1PI_1 + PS_2PI_2 + \ldots + PS_iPI_i}{N_p 100}$$

where: PS = severity of problem
PI = importance of problem
N_p = number of problems listed

In Figure 1 an example is cited.

Figure 1

Example of Rating of Individual Problems

Problem/s	Problem Severity (1=not at all severe to 100=extremely severe)	Problem Importance (portion of 100 percent)
1. (fictitious problem)	87	55
2. (fictitious problem)	39	30
3. (fictitious problem)	48	15

$$\text{IPR score} = \frac{(87)(55) = (39)(30) + (48)(15)}{(3)\ 100} = 22.25$$

*With appropriate instructions to treat the distance between each interval as equivalent.

The identical procedure would be repeated at standard intervals, possibly every four weeks. If the problem list does not change, the same set of problems is rated again. If the problem list does change, the set from the previous rating is rated again, the new problems added to the set and the enlarged set rated. This procedure allows for a comparison between each successive time interval. Because of the interval level nature of these data all arithmetic functions are permissible. Therefore, depending on the evaluation design employed, appropriate parametric statistics could be applied. For example, in the case of treatment vs. non-treatment designs, comparisons between groups at time-one, time-two, etc. could be made; in the case of a pre-post, single group design* changes within the treatment group could be examined and possibly a comparison of treatment modalities if random assignment to treatment modality was made.

Obviously the real world is not so simple that persons so inclined could simply print instructions and forms for applying IPR and march off to evaluate their social service program. IPR does have the advantage of being based on two reasonably clear criterion concepts, severity of problem and importance of problem; the psychometric tradition of self rating is well established; and the interval level of measurement makes scores much more interpretable and easy to manage mathematically and statistically. Nonetheless, there are a number of problems in using IPR which must be carefully investigated before the procedure can be considered a valid and reliable measurement form.

Issues to be Studied

One of the issues to be studied relates to the type of individuals for whom this form of measurement is appropriate. We have designated the client

*This design is, of course, subject to the limitations mentioned earlier in the critique of GAS.

or patient as the source of the identification and rating of problems. This decision was based on the assumption that the individual experiencing the problems can make the most valid statement of problems for which help is sought. This eliminates such factors as clinical bias and halo effects in perceiving client problems, and it also reflects a trend in treatment toward greater involvement of the client in the treatment process and evaluation. This assumption runs into problems on at least two fronts: first, individuals who are out of touch with reality (psychotic) probably could not identify that condition or any other with any accuracy; and second, individuals who are participating in the treatment under coercive circumstances may give distorted responses in any of the three IPR tasks. There also may be variation on performing the tasks caused by such factors as level of education and level of income. The issue, then, is essentially, who among the set of individuals seen for social and mental health problems is an appropriate respondent to the IPR procedure? We plan studies around this issue.

Our first effort will be a feasibility study, involving the use of IPR with a sample of clients from an agency specializing in child abuse and neglect cases or with a substantial caseload of these cases.* After clients have completed the procedure, we will conduct an open-ended interview to determine the problems encountered in performing the procedure. We will compare these responses to such variables as education, income, type of problems (at least intrapsychic vs. relational) and condition of referral (e.g., voluntary vs. non-voluntary). If there is no significant variation caused by these variables we will consider the procedure invariant with regard to client characteristics and condition of referral. If, however,

*We also intend to test the procedure with other less specialized problem groups such as in community mental health agencies, family and child counseling agencies, and so on.

significant variation is associated with one or more of these variables, the procedure will be qualified as indicated. and the conditions under which it can be used appropriately will be thus restricted. But variations to the basic instrument will be proposed. For example, with psychotic clients the referral source might have to rate the problems, resulting in a different procedural format than that employed when clients are capable of stating their own problems.

Another issue involves the level of abstraction at which problems are stated. Instability in the problem list which is merely a reflection of variation in the level of abstraction at which the problems are stated poses a critical problem whether the exercise involves stating problems or stating goals and this is so irrespective of who is making the statement. We anticipate increased instability as the level of abstraction is reduced. For example, the problem of "conflict in the marital relationship" is probably a reasonably stable problem, but this problem is composed of a set of subproblems, e.g. "not spending enough time with spouse," "not sharing in the care of the children," and so on, which may fluctuate depending on momentary, situational events. Variations might be related to the ability of the individual to articulate problems and this ability might be assumed to improve with practice and as a result of the socializing effects of interaction with a clinician. Studies in this area should result in some operational guidelines which will reduce to a minimum effects on scores which might result from these sources.

On the level of abstraction our first concern for study will be variation within clients rather than variation among clients. That is, if clients are reasonably consistent over time in the level of abstraction at which problems are specified, we would have little concern. But, if clients

change levels of abstraction over time, we will need to adjust the procedure or its instructions to reduce this type of variation. We will study this issue by analyzing the levels of abstraction produced from the first sample of clients who use the procedure. We also will be checking the level of abstraction in terms of the problem severity ratings and the problem importance ratings.

Closely related to the issue of the level of abstraction is the issue of the interaction or overlap of problems listed. An example of such a situation would be a IPR problem list which included "conflict in the marital relationship" and "conflict in parent-child relationships." Rarely are these viewed as independent phenomena. Again, this is an issue attendant of goal statements as well as problem statements. To a certain extent it may be controlled operationally by establishing clearcut guidelines. Likewise it may be controlled mathematically via corrections to the formula for calculating the IPR score. The corrected IPR formula being:

$$\text{IPR Score Corrected} = \frac{PS_1PI_1 + PS_2PI_2 + \ldots - PS_1PI_1(PS_2PI_2)}{N_p 100}$$

where: PS = severity of problem
PI = importance of problem
N_p = number of problems

Obviously, however, carefully thought out guidelines will have to be developed, and to the extent possible empirically tested to establish for which combinations of problems the correction factor is necessary.

The interaction issue will be studied by including in our open-ended interview, mentioned above, a question about the degree of interaction between each combination of problems. We may also draw upon clinical judges as a

means of assessing the degree of interaction between problems. It is anticipated that the interaction may be a function of the level of abstraction. If the degree of interaction is significant, we may institute a routine of always using the corrected-formula (subtracting the interaction) for every set of problems. The alternative is to develop guidelines for using the corrected-formula based on combinations of problems known to be interactive as a result of accrued experience with clients making problem statements.

Another issue is how to handle changes over time in the problem list. Our solution is the guideline that problems may be added, but never deleted. At the extreme, if a problem was completely resolved it should still be retained as an element of the problem set by giving a severity rating of "1", an importance assignment of "1" percent, and it should be counted as a part of N_p. The reason for this strategy is that to give credit for problems which in the rare case are totally resolved, the problem must be retained in the denominator of the formula without adding substantially to the numerator. One times one is one, of course, which we submit does not substantially distort the numerator, the exception being the unlikely event where an individual totally resolved a large number of problems. We believe this solution corrects for this problem sufficiently that it need not be investigated further.

Finally, the validity and reliability issues must be approached empirically also. We plan to investigate the concurrent validity of the changes observed over time by comparison with an accepted measure of interpersonal and intrapersonal adjustment such as the MMPI, CPI, and so forth. Reliability will be determined by a derivative of coefficient alpha appropriate to linear combinations (Nunnally, 1967:232-235).

The several studies of issues attendant to the testing of the IPR procedure which have been mentioned are presented as the beginning points in studying the application of this proposed procedure. We fully anticipate that as these studies develop and as experience with the procedure develops, other studies will be suggested and mandated. Those, however, cannot be anticipated specifically at this time.

We conclude by again underscoring the tentative nature of the measurement device. It is presented here as a conceptual proposal so it might be tested by other researchers prior to any wholesale adoption by practitioners of clinical evaluation.

SUMMARY

The major concern of this paper was a critical assessment of Goal Attainment Scaling as a measurement procedure both in terms of its psychometric quality and its potential utilization as an evaluation tool in large-scale demonstration programs. We found GAS less than conceptually clear and mathematically in error. For these reasons it obviously could not be recommended for the evaluation of any program, large-scale or otherwise. We submit that Goal Attainment Scaling was adopted in practice well before a variety of issues attendant to its conception and operationalization were appropriately investigated. In what was intended as a constructive derivative of this critique, we have proposed an individualized measurement procedure which is both a conceptual and a mathematical advance beyond Goal Attainment Scaling. This procedure, called Individual Problem Rating, is a problem rather than goal based method. A number of issues which must be empirically resolved prior to the adoption of this procedure in evaluation practice have been outlined, and it is our hope that others will join us in refining or refuting the IPR alternative.

REFERENCES

Barlow, S., & Ravneberg, R. Goal attainment scaling: Procedures, problems and modifications. Tacoma, Washington: Office of Research and Evaluation, Comprehensive Mental Health Center, undated.

Campbell, D.T., & Stanley, J.C. Experimental and quasi-experimental design for research. Chicago: Rand McNally, 1966.

Clayton, S.C. Goal attainment scaling critique. Salem, Oregon: Department of Human Resources, Mental Health Division, May 1975.

Garwick, G. A construct validity overview of goal attainment scaling. Minneapolis: Program Evaluation Project Report, June 1974a.

Garwick, G. An introduction to reliability and the goal attainment scaling methodology. Minneapolis: Program Evaluation Project Report, June 1974b.

Garwick, G. Guidelines for goal attainment scaling. Minneapolis: Program Evaluation Project Report, October 1974c.

Garwick, G. (written correspondence), September 10, 1975.

Howe, M.W. Casework self-evaluation: A single-subject approach. Social Service Review, 1974, 48, 1-23.

Kiresuk, T.J. Goal attainment scaling at a county mental health service. Evaluation, 1973, Special Monograph No. 1, 12-18.

Kiresuk, T.J. & Garwick, G. Basic goal attainment scaling procedures. Minneapolis: Program Evaluation Project Report, April 1975.

Kiresuk, T.J., & Sherman, R.E. Goal attainment scaling: A general method for evaluating comprehensive community mental health programs. Community Mental Health Journal, 1968, 4, 443-453.

Likert, R.A. A technique for the measurement of attitudes. Archaeological Psychology, 1932, 140.

Mauger, P. A study of the construct validity of goal attainment scaling. Minneapolis: Program Evaluation Project Report, June 1974.

Nunnally, J.C. Psychometric Theory. New York: McGraw-Hill, 1967.

Sherman, R.E., Baxter, J.W., & Audette, D.M. An examination of the reliability of the Kiresuk-Sherman goal attainment score by means of components of variance. Minneapolis: Program Evaluation Project Report, August 1974.

END