**THE FUNCTIONS OF ASSESSMENT: IMPLICATIONS
FOR SELECTION AND DEVELOPMENT OF DEVICES FOR
ASSESSING REPERTOIRES IN CLINICAL, EDUCATIONAL,
AND OTHER SETTINGS**

ROBERT P. HAWKINS

WEST VIRGINIA UNIVERSITY

An attempt is made to identify the many different functions that assessment of an individual's repertoire can serve. Implications of these functions for the character of and evidence about assessment devices are suggested. The functions fall into two general groups, those which influence decisions regarding an individual learner, and those which influence policy, program development, and scientific knowledge. The first group of functions is presented in a rough chronological sequence such that they form a "behavioral assessment funnel," beginning with functions involving broad-band assessment to identify likely persons and skill areas, and narrowing to the precise pinpointing, monitoring, and follow-up functions. The contribution of behavior analysis and behavior therapy to assessment methodology in this sequence is identified as well as the areas where more traditionally conceived methods are still useful. The second group of functions and behavioral contributions to it are then discussed.

DESCRIPTORS: behavioral assessment, functions of assessment, selection of assessment devices, traditional assessment, clinical assessment, educational assessment, phases of assessment, construction of assessment devices

Applied behavior analysis has made profound contributions to the technology for assessing repertoires. At the same time, behavior analysts have often rejected or ignored valuable existing ("traditional") knowledge and technology for clinical and educational assessment (Cone, 1977; Goldfried, 1977; Hawkins, 1975; Nelson and Bowles, 1975; Hartmann, Note 1). The purposes of the present paper are three. The first is to identify several different phases and functions in assessment. The second is to put the contributions of behavior analysis in context, showing where the contribution thus far has been greatest and where previously existing knowledge and technology can be of value

to a behavior analyst. The third is to illustrate how the present analysis holds implications for selection and construction of assessment devices, including implications regarding generalizability (Cronbach, Gleser, Nanda, and Rajaratnam, 1972), type of output needed, and costs.

First, a discrimination that appears useful is between assessment designed to influence directly decisions regarding an individual and assessment designed for one or more of three other functions: classification for administrative records, research for evaluating interventions, and collection of normative data. Those functions serving the general purpose of influencing decisions regarding an individual will be presented first.

This paper is based partly on one presented at the meeting of the American Psychological Association, September 1976, Washington, D.C., under the title "Relevance for what? It depends on the assessment goal," in a symposium chaired by D. H. Barlow and entitled *Behavioral assessment: The relevance of traditional psychometric procedures*. Reprints may be obtained from the author at Psychology Department, West Virginia University, Morgantown, West Virginia 26506.

**PHASES AND FUNCTIONS
INVOLVED IN ASSESSMENT
INFLUENCING DECISIONS
REGARDING AN INDIVIDUAL**

Texts on assessment generally are designed to help train professionals for work with indi-

vidual students, clients, or other learners. The general goal of such work is usually improvement of the learner's adaptation to his or her environment in some way, such as through acquisition of academic skills, learning to interact more effectively with members of the opposite sex, or relearning functional speech or locomotion. Thus the general purpose of the individual assessment taught by such texts is, as Linehan (1977) has described, "to figure out what the client's problem is and how to change it for the better (p. 31)"; or, in terms more descriptive of educational assessment, "to figure out what the client's present skills are and what should be learned next."

But the process of figuring out "what the client's problem is" or "what should be learned next" is not a unitary process; it is one consisting of several sequential phases, each with a somewhat different function and each placing somewhat different requirements on the assessment methods employed. As Anastasi (1968) put it, speaking of validity, "No test can be said to have 'high' or 'low' validity in the abstract. Its validity must be determined with reference to the particular use for which the test is being considered" (p. 99). Yet assessment texts, including Anastasi's, generally do not distinguish between the different functions and their different implications, and most texts completely omit some of the functions.

THE "BEHAVIORAL ASSESSMENT FUNNEL" AND THE FUNCTIONS PREDOMINANT IN EACH PHASE

The phases of individual case assessment can be considered to form a funnel, with an initially broad scope and a subsequent narrow and constant focus (Cone and Hawkins, 1977). The five phases are as follows: (1) screening and general disposition, (2) definition and general quantification of problem or achievement, (3) pinpointing and design of intervention, (4)

monitoring of progress, and (5) follow-up. Though these phases describe a typical chronological sequence, the functions attributed to them here are not so firmly limited in time. For example, the functions of the first phase are never totally eclipsed and an initial pinpointing of the problem in phase three might well be revised during a later phase.

Screening and General Disposition

Functions predominant in screening and disposition. When a potential learner (client or student) first appears at the door of a clinic or educational institution, the first assessment function to be performed is *determination of whether the case is appropriate* for the agency. The question is "Does this case belong here?" This assessment is often conducted partly or wholly by clerical staff and may take only a few minutes. In the schools the person's age and previous grade in school may suffice, though colleges often require a certain level of performance in previous schooling and/or on an entrance examination. In clinics the person answering the telephone will often obtain some gross description of the client's problem and will refer those clients who obviously need a different agency's service.

If the answer to the initial screening question is "No, the learner does not belong here," the assessor must make a *determination of where to refer the learner* for more appropriate services. But if the answer to the initial screening question is "Probably," the next assessment function is the *determination of what further assessment may be relevant*. In the clinic this second screening and disposition question might be phrased "What general kind(s) of problem(s) may this person be having (and thus what kinds of further assessment information will I need)?" In a school the question might be simply "Is it likely that this student's academic or social skills are unusual enough for his or her age that some program out of the

ordinary might be called for?" The whole screening and disposition phase may take 5 minutes or several hours.

Characteristics desirable in methods used for screening and general disposition. The screening and disposition phase requires a broad-band assessment (Cronbach, 1970) capable of detecting any likely kind of problem, even if it does not satisfactorily specify the nature of the problem. Thus interviewing, a case study, a "mental status" examination, incomplete sentences, and the like might be appropriate for clinical cases, because each is very flexible and can readily survey a broad range of possible problems. For example, a clinician assessing a medically referred case involving complaint of chronic headache might initially attempt to determine what neurological assessment has been done, what dental assessment has been done, what consequences might normally ensue when the client complains of headache, and how demanding the client's job and home life have been. In educational environments, interviews, the Wide Range Achievement Test (Guidance Associates of Delaware, 1978), an AML (Cowen, Dorr, Clarfield, Kreling, McWilliams, Pokracki, Pratt, Terrell, and Wilson, 1973), the STS Youth Inventory (Scholastic Testing Service, 1971), or the Mooney Problem Check List (Psychological Corp., 1950) might be appropriate, depending on whether academic, personal, or social skills are of interest.

The functions of this phase of assessment are a crude sorting of people into "serve" or "send elsewhere," and "assess for X" or "do not assess for X." These are obviously predictions, and any assessment device that improves the accuracy of these predictions is worth considering. For example, if one had evidence that the MMPI improved detection of potential suicides in a particular population, it would seem foolish not to use the MMPI as part of the screening of a client representing that population. One need not subscribe to the theoretical assumptions regarding "personality" that typify

users of a device like the MMPI in order to take advantage of any predictive value it may have.¹ On the other hand, neither must one accept the traditional correlation coefficient as the measure of predictive validity. A more satisfactory evaluation would be a series of contingency tables for each of several promising cutoff scores. Each table would show how many true positives, false positives, true negatives, and false negatives a particular cutoff score would produce with a given population.

The kind of output needed for screening and disposition functions is simple. The assessment must simply classify the learner as appropriate or inappropriate for the agency and identify further assessments that may be useful. Commensurate with the simplicity of output is the allowable cost. Usually screening and disposition devices must be inexpensive, particularly in terms of professional time.

Contributions of behavioral assessors to screening and disposition assessment. Of the five assessment phases, this one has probably received the least attention by both traditional and behavioral assessors. Certainly the SORC analysis of assessment by Kanfer and Saslow (1969) made a significant contribution to a behavioral conception of both this phase and the second phase of assessment. Interview guides (Holland, 1970) and descriptions (e.g., Morganstern, 1976) similarly contribute to both phases. The computer-assisted clinical interview described by Angle, Hay, Hay, and Ellinwood (1977) is designed to determine in what areas of adjustment clients may be having sufficient difficulty to warrant further inquiry. This interview has the advantages of being inexpensive in terms of

¹Suicide risk is intentionally used in the example to emphasize the fact that a cost, risk/benefit is implicit in deciding to conduct any assessment. In this case the risk is not detecting a high probability, negatively valued event. The magnitude of such risk and the magnitude of benefit from detecting a problem (e.g., from ensuing treatment), would determine the cutoff scores used and the tolerable cost of the assessment.

professional time (though not the client's), very broad in scope, and completely standardized (and thus readily researchable).

Definition and General Quantification of the Problem or Achievement

Functions predominant in definition and quantification. This phase is the most familiar to clinical and educational assessors, and traditional texts on testing deal almost exclusively with it (e.g., Anastasi, 1978). This phase is comparable to the aspects of assessment that Mash and Terdal (1976) label "diagnosis," in reference primarily to clinical problems. It takes up where screening and disposition left off, with rough hunches as to the nature of the problem. It proceeds with a careful evaluation of those hypotheses and ends with a commitment to certain hypotheses about the functional relations involved in the problem. Further assessment and treatment is then based on these hypotheses. In this process the problem will often have been quantified, if only in gross terms such as "mild" or "severe," or in terms of a prognosis.

Thus two major processes are required in this phase—hypothesis formation and measurement—and the clinician will often cycle repeatedly between hypothesizing and measuring (including probing for the simple presence or absence of hypothesized factors) until some tenable formulation of the problem is attained. In clinical settings a great deal of knowledge is often required about behavior pathologies if the assessor is to be effective in generating reasonable hypotheses.

The functions of this phase, then, are at least the following: (1) *identification of the skill or problem areas in which help is definitely to be offered*, (2) *quantification of the general level of skill in the relevant areas*, and finally (3) *determination of where and how, in general, the needed learning can best be accomplished*. Put more generally, the functions are to set the general goals of intervention and select the environment for their achievement.

The kinds of questions asked in this phase might include "How serious is this particular problem?" "What is this student's level of reading comprehension?" "Should intervention be attempted?" and "What program or person can be of greatest help?" Note that the outcome of this phase is *placement* in a general learning environment—such as a second-grade classroom or with an assertiveness training expert—and often, within that environment, at a particular level of skill, such as a third-grade reader or an advanced assertion group. Many clinicians and educators terminate serious assessment efforts here.

Characteristics desirable in methods used for definition and general quantification. The first outcome of this phase is definition of the general nature of the problem. This might be achieved through a series of assessment activities, each with its own validity requirements. For example, the problem of a client whose initial complaint was chronic headache might have been identified in this phase as one involving insufficiently assertive behavior on a job in which the new boss was unreasonably demanding. The devices used in arriving at this hypothesis might have included interviews that explored several questions, such as when the headaches began and how often they occur, whether there had been any trauma to the head, and what precedes and follows a headache. Ideally one would have evidence that these questions, individually or collectively, are good discriminators between persons known to have psychosomatic headaches and persons known to have headaches based on other factors. The assessment might also have included some form of neurological examination, and here again one would look for criterion-related validity showing that the type of examination given was a good discriminator between two groups, particularly between persons with neurologically based headaches and persons with headaches based on other factors. Finally, the assessor may have used a structured assertiveness test, in which case it would be desirable to have evidence that this test discrim-

inates well between persons whose behavior is insufficiently assertive in the face of unreasonable demands in the natural environment and those whose behavior is effectively assertive (e.g., McFall and Lillesand, 1971).

The final outcome of this phase is a recommended placement of the learner, which is essentially a prediction that that person (or perhaps others) will fare better with such a placement than with alternative placements, including no special treatment or education. Evidence for the validity of the placement could be program evaluation data showing that similar learners had performed better during placement in the similar programs than without those programs. The similarity of the present learner to those with whom such predictive validity was established might be on the basis of the type of problem to be remedied, demographic variables, or various traditional test scores. So long as there is good evidence that a particular variable predicts success in a program, the behavior analyst should feel free to use that variable as a predictor. The fact that it is a traditional assessment device with little value in *identifying the problem* should not dissuade the behavior analyst from using it in *determining placement*. Any assessment device may have high validity for one function and low validity for another.

The cost of conducting assessment in this second phase may be considerably higher than in the screening and disposition phase, particularly with complex clinical problems. The output from this phase may be as simple as the statement, "This student's academic skills are progressing at the expected rate, except in the area of written expression where they are beyond expectation; continuation of the present program is recommended." On the other hand, the output could be a complex formulation describing how the natural environment may have established a set of interrelated, maladaptive behavior patterns and may continue to maintain them. In addition, the output will include a general recommendation regarding the type of intervention suggested.

Contributions of behavioral assessors to definition and general quantification. Behavioral contributions to this phase of assessment have come primarily from behavior therapists whose interest is individual clinical problems. In contrast, contributions to the next three phases have come primarily from applied behavior analysts with a great diversity of interests.

Behavioral assessment for this phase may use a variety of methods, including indirect assessment of behavior, direct assessment in controlled environments, and direct methods in natural environments (Cone and Hawkins, 1977). The behavioral interview—both a direct and an indirect assessment method, depending on the behaviors of interest—has been developed by a number of behavior therapists. Morganstern (1976) provides analysis of the procedures used, while Goldfried and Pomeranz (1968), Kanfer and Saslow (1969) and others focus more on the content. The interview is used to assess a wide range of clinical problems, as are self-report questionnaires. Examples of self-report questionnaires used by behavior therapists are the Beck Depression Inventory (Beck, 1972), the Rathus Assertiveness Schedule (Rathus, 1973), and the Menstrual Symptom Questionnaire (Chesney and Tasto, 1975).

Direct assessment in controlled environments has been reviewed by McFall (1977) and by Nay (1977). It includes measurement of a client's avoidance of a feared object through the behavioral avoidance test, the assessment of a client's heterosocial skills through a contrived meeting with a confederate of the opposite sex, the assessment of how assertively a client behaves through a telephone call involving a series of demands, and measurement of sexual arousal to verbal descriptions of situations (Abel, Blanchard, Barlow, and Mavissakalian, 1975).

Direct assessment in the natural environment can be carried out by the behavior therapist or by his agent, as when Hawkins, Peterson, Schweid, and Bijou (1966) observed parent-child interaction in the home to determine what

the problem was. But self-monitoring in the natural environment might also be considered direct assessment to the extent that one can have confidence in the reliability of the data. As Mahoney (1977) and Nelson (1977) have illustrated, self-monitoring can include immediate recording of a specified event, a structured diary in which antecedents and consequences are recorded along with a particular class of event, immediate ratings at preselected or random times, or tape recordings of verbal descriptions of behavior and associated events. Any of these could be very valuable in defining and quantifying a clinical problem to determine an appropriate type of intervention.

In educational settings the most familiar assessment device for these functions is the standardized, norm-referenced achievement test, such as the Iowa Test of Basic Skills (Houghton Mifflin, 1974). A norm-referenced test can assist in determining what group a student fits into best. Such a test describes the general level of a student's skills in each of several skill areas; however, it is of limited value in determining precisely what skills to teach the individual child next, a major function of the next phase of assessment. By contrast, a criterion-referenced test (Livingston, 1977) can perform both these functions.

It should be noted that one form of assessment in the definition and quantification phase can be to "short circuit" much of the process and to place the learner in a likely program for a trial period. If the costs and risks of such a placement are likely to be smaller than those of another form of assessment, the trial period approach seems advisable. It has the advantage of providing the best possible prediction as to how well the learner will perform in the program: his or her actual performance; but the assessor must actually monitor the learner's performance during the trial period and not prematurely rule out alternative placements. The trial period approach was employed by Hansen and Lovitt (1976) in deciding what reading text was the most appropriate for each child in

a special classroom. They monitored both "words correct" and "words incorrect" per minute in oral reading from texts at each of several difficulty levels.

Pinpointing and Design of Intervention

This phase corresponds closely to the "design" aspect of assessment described by Mash and Terdal (1976). It involves functions that educational and psychological tradition have given relatively little attention. The functions are: (1) *selection of specific behavioral objectives*, (2) *analysis of specific existing contingencies that may be maintaining the subject's ineffective behavior or may otherwise impede achievement of the objectives*, (3) *identification of specific resources (social, physical, financial, behavioral) for achieving the objectives*, (4) *design or selection of a teaching sequence (curriculum, in education's terms) to achieve the objectives*, and (5) *determination of precise quantitative levels of key measurable events*.

The focus of this phase is very narrow and constitutes the apex of the behavioral assessment funnel. Baseline data on a learner would be a final outcome of this phase, as the phase ends the pre-intervention assessment.

Questions addressed in this phase include, "What would the stimulus control and topography of a more effective performance look like and how does this person's performance deviate from that?" "How regular and adequate are the current antecedent stimuli (opportunities and demands for performance) and consequences?" "What interfering behaviors are occurring and why?" "Under what antecedent and consequence contingencies is the desired performance to be maintained eventually?" and "What specific resources (*e.g.*, antecedents and consequences) promise to be useful in changing the repertoire?"

Characteristics desirable in methods used for pinpointing and design of intervention. The kind of generalizability that appears most important in this phase is content validity. If the definition of the problem in the previous phase

has been correct, the assessor's job in pinpointing is to determine what specific behaviors truly make up that problem in this individual case (thus the relevance of task analysis; *cf.* Schwartz and Gottman, 1976). The assessor needs devices that sample the relevant behaviors, antecedents, and consequences in such a way that the typical quality of each is assessed.

The typical norm-referenced assessment device will be of little use. Instead the assessor needs devices that are criterion referenced in the sense that they assess how well relevant aspects of the repertoire approximate repertoires known to be *effective* at the tasks identified. Though the decision as to what levels of performance are effective may be based on data that are normative in the sense of describing a criterion group (*e.g.*, Walker and Hops, 1976), the ideal basis is experimental demonstration of the effectiveness of various levels of performance (Hawkins, 1975).

Because repertoires and environments are so individual in terms of what skills are critical and what levels of performance are effective, direct observation in the natural environment (Cone and Hawkins, 1977) or at least in controlled, naturalistic situations is very appealing.

The documentation of the content validity of such assessments would be difficult, at best. A very relevant body of information would be research validating task analyses like the one being used in pinpointing. For example, shy males were once taught to make greater eye contact while conversing with women, but it has been discovered in recent experiments that what is probably important is not the *amount* of eye contact but *when* it occurs (McFall, 1977). Through a number of such studies experimentally testing the effects of teaching certain skills to persons with a certain type of problem, we can discover an increasing number of the necessary and sufficient skills for effective performance.

Where contrived situations are used to pinpoint, evidence is also needed that the performance observed is related to the natural

environment performance that the assessor assumes is being evaluated. This is particularly critical when neither the antecedent stimuli nor the form of response are identical to those in the natural environment. For example, Freedman (1974) (see also McFall, 1977) developed an Adolescent Problem Inventory in which a youth reports *verbally* what he or she would do in a series of problem situations that are *verbally* described. The specific situations described to the youth were selected by a series of judgments made by Freedman, by delinquent youth, and by experts. Freedman found the device to discriminate significantly between delinquents and nondelinquents, but that is weak evidence that the 44 items represent truly critical situations or that the verbal responses obtained represent the behavior that the subjects would actually emit in the natural environment. Thus the device seems of very limited promise for the pinpointing functions, though it may be useful for other functions.

A familiar form of validity evidence relevant in this phase is that involving the accuracy of the predictions involved. The intervention design generated in this phase is a prediction that if the design is implemented the behaviors and environmental factors will change. These predictions are at least partially verifiable. If the environment and behavior do not change, the decisions of this phase (and perhaps earlier ones) are invalid at least to the extent of being insufficient. If the desired changes do occur, the decisions receive at least partial support, though placebo and other effects may actually account for the changes.

Finally, interobserver agreement is relevant for this phase, though it will not be discussed further until the next phase. Here let me point out an interesting form of interobserver agreement that is particularly appropriate to the pinpointing and design phase. It is a form of social validation (*cf.* Wolf, 1978) and would consist of asking the learner to what extent the process and outcome objectives set during this phase seem appropriate to the problem or skill level

with which he or she entered. This is common practice among some clinicians, but below the college level it appears to be rare among educators.

Contributions of behavioral assessors to pinpointing and design of intervention. Behaviorally oriented educators and clinicians have made innumerable and profound contributions to the functions of this phase. In education, learning hierarchies are being constructed (e.g., Resnick, Wang, and Kaplan, 1973) that form the basis for assessment as well as curriculum development. Criterion-referenced tests for assessing such academic skills as reading are beginning to be commercially published (e.g., Harcourt Brace Jovanovich, 1975). "Prescriptive" behavior checklists such as those of Popovich (1977) and Walls, Zane, and Werner (1978) are being developed that appear to perform all of the functions of this phase for certain types of learners.

Clinical assessment devices for this phase include a variety of direct observation procedures (Bijou, Peterson, and Ault, 1968; Hawkins, Axelrod, and Hall, 1976; Mahoney, 1977) applied to a wide range of problems. Barlow, Reynolds, and Agras (1973) directly observed the topography of sitting, standing, and walking behaviors in a young male transsexual and, based on comparative observation of normals, targeted very specific aspects of these topographies. Schwarz and Hawkins (1970) videotaped a very shy adolescent in a classroom setting and selected three specific behaviors for modification: voice loudness, posture, and face-touching. Trap, Milner-Davis, Joseph, and Cooper (1978) used transparent overlays to assess seven different accuracy characteristics of cursive letter writing in first-grade students.

Oral and written verbal reports are also very useful in the pinpointing and design phase. LoPiccolo and Steger (1974) use a well-structured questionnaire to target specific sexual interactions of spouses having sexual adjustment problems. Stuart and Stuart (1976) designed a set of detailed questionnaires for families to assess

positive and negative behavior change goals and the family resources for achieving them. Questionnaires have also been developed for assessing only potential reinforcers (e.g., Daley, 1976).

Monitoring of Progress

Functions of monitoring. As the name implies, the primary function of assessment in this phase is *determination of the current level of performance* so that the intervention can be evaluated and changed as needed. Of all assessment functions this may be the one to which behavior analysts have made their greatest contribution. Being accustomed to precise, continuous measurement in the laboratory, pioneering behavior analysts addressing human problems developed methods for similarly measuring behavior in the field (e.g., Ayllon and Michael, 1959; Ferster and DeMeyer, 1962; Hart, Allen, Buell, Harris, and Wolf, 1964; Lindsley, 1959). The precise, daily (or frequent) measurement procedures resulting from these pioneering efforts have many advantages over global, norm-referenced or infrequent measurement (Eaton and Lovitt, 1972; Hawkins, Axelrod, and Hall, 1976).

Another function that has recently emerged (Wolf, 1978) is *evaluating the palatability of the intervention to the learner and other consumers*. As Phillips, Phillips, Wolf, and Fixsen (1973) have indicated, the ideal program is one that is not only maximally effective but also *preferred* by the learners.

Because the measurement used to monitor progress is narrow band and constant over time (except when target behaviors change), this phase of assessment represents the long, narrow "neck" of the behavioral assessment funnel. The questions addressed in this phase include, "Are the desired changes in key events taking place?" "Are events whose change is undesirable staying stable?" and "Is this as humane a process as it can be?"

Characteristics desirable in assessment methods for monitoring progress. The devices useful

in monitoring progress will be those that are criterion referenced or "edumetric" (Carver, 1974) in the sense that they assess only behaviors being taught (content validity) and are sensitive to change. If, in past research, a device has failed to show change when it is reasonable to expect change (e.g., Lobitz and Johnson, 1975), it would not be a good candidate for monitoring progress. Further, the device should not be responsive to spurious influences such as retesting effects or instrument decay (cf. Baer, Wolf, and Risley, 1968; Kent and Foster, 1977; Wildman and Erickson, 1977).

As with almost any assessment device used in any phase, evidence of high interobserver (inter-examiner, interscorer, interrater) agreement is desirable in the monitoring of progress. When progress is monitored through direct measurement of the target behaviors in the natural environment, this may seem to be the only relevant information, because there is no better criterion with which to compare the data. However, it is also desirable to know that the data represent the typical performance of the learner in situations and at times other than those in which the data are being obtained (Cone and Hawkins, 1977).

To the extent that a device differs from direct measurement in the natural environment, evidence of a high correlation between the device and such direct measurement becomes important. However, this should probably be qualified by pointing out that it is sometimes more efficient and more socially valid to measure the functional outcome or "impact" of the target behaviors, rather than the target behaviors themselves, such as measuring the frequency of dating in a client who has been taught several specific hetero-social skills to solve a problem he or she has identified as an inability to get dates. In fact, Gilbert (1978) would probably argue that measuring the behavior itself is trivial without assessment of what we might call the "impact" of the behavior change.

A further characteristic that is necessary if an assessment method is to be used routinely

and repeatedly with all learners being served by a program (as opposed to use in time-limited research) is economy. Mahoney (1977) has described some economical methods for obtaining clinical data, and others (Hall, 1971; Hawkins, Axelrod, and Hall, 1976; Kubany and Sloggett, 1973) have described some for use by school personnel, but further development of economical, readily disseminated methods is needed.

In assessing the palatability of the intervention, perhaps the primary criterion a device should meet is that one index of palatability agree with another index from the same subjects. For example, if one is administering a questionnaire to assess a student's satisfaction with a learning program, it would be beneficial to know that the student gave a similar report when the same questions were posed by others not obviously connected with the program. Even more convincing would be evidence that when repeatedly given the choice to participate in that program as opposed to others, the learners would choose that program, as Phillips *et al.*'s (1973) youth did.

Contributions of behavioral assessors to monitoring. The technology of using repeated measurement, particularly daily (Eaton and Lovitt, 1972) or very frequent measurement (e.g., Bushell, Jackson, and Weis, 1975), has been highly developed by applied behavior analysts. This technology is well illustrated throughout the *Journal of Applied Behavior Analysis*. Behavior analysts are working to make the technology practical for everyday use in clinical and educational settings, and the technology for maximizing causal interpretability of the results is well developed (e.g., Hersen and Barlow, 1976).²

²The author would argue that the primary responsibility of the practitioner is to document the direction (favorable *vs.* unfavorable) and amount of change occurring in the repertoire *during* teaching or treatment, and, in many clinical cases, the durability of that change. The use of research designs implies a research and development function, to be discussed later, not service delivery. Documenting that the service being provided is causing the change, particularly through

Follow-up

Function predominant in follow-up. The primary function of a follow-up assessment is *evaluation of the durability and sufficiency of behavior changes*. Evaluation of durability is usually accomplished through continuing the measurement used during the previous phase, but making it less frequent so that follow-up can continue for an adequate period without excessive cost. Thus the end of the assessment funnel's "neck" would represent the durability evaluation.

The questions of interest include, "Is the effect lasting?" "Is it carrying over to the desired situations in which training did not occur?" "Is it satisfactory to the persons who requested the change?" and "Are there unexpected effects in the repertoire or the environment?"

The funnel simile may be violated somewhat when the sufficiency of the behavior change is evaluated. Because the monitoring might have sampled only parts of the problem for the sake of economy, the follow-up may be broader band in order to see whether the whole program has been solved. A more inclusive follow-up could also detect side effects of the intervention.

Characteristics desirable in devices used for follow-up. In general the same characteristics needed in the pinpointing and the monitoring phases are useful in follow-up, but at least two additional characteristics are often of value here. The first is breadth of coverage. To assess side effects, one must sample a wider range of behaviors and environmental factors than those assessed during the monitoring phase. Use of a broad-scan behavioral checklist before and after treatment might be useful, for example. A second useful characteristic is economy. Once a learner is no longer intensively engaged in a treatment or education program, he or she is unlikely to be willing to invest the needed time,

labor, or inconvenience for frequent or difficult assessments.

One form of evaluation of the sufficiency of a behavior change is measurement of consumer satisfaction (Wolf, 1978). The learner and/or others who are in a position to be affected by the learner's behavior are asked such things as whether the behavior change was relevant and great enough. If their verbal reports on a consumer satisfaction questionnaire correlate with such measures as their verbal reports to others on the same topic, their referral of the learner for similar help again, their hiring the learner, or their admitting the learner to more advanced programs, the questionnaire is shown to be serving its purpose. Such validity evidence on consumer satisfaction measures does not seem to be available yet.

The Multiple-Phase Utility of Some Devices

It is interesting to note that many behavioral assessment devices appear to serve numerous functions that cross several phases. For example LoPiccolo and Steger's (1974) Sexual Interaction Inventory would probably be useful in every phase but screening and general disposition. Similarly, a comprehensive, criterion-referenced series of tests of mathematical skills organized according to a learning hierarchy (Resnick *et al.*, 1973) would be useful for all phases except screening and disposition. The characteristics that seem to make a device useful for so many functions appear to be these: (1) the device is designed to assess performance of a fairly well-defined task or task-sequence;³ (2) the device samples most of the skills needed to perform that task; (3) it does not sample many other skills (or if it does they are clearly identified with a different task); (4) it includes quantification of each of the skills sampled; and (5) one can readily measure the targeted skills individually without constructing a new assessment device.

use of designs more complex than AB (Hersen and Barlow, 1976), goes beyond efficient service delivery and falls in the research and development function described later in this paper.

³The word "task" is used here to mean a combination of a stimulus array (situation) and a setting event (goal, motivational state; cf. Bijou and Baer, 1978).

FUNCTIONS THAT INFLUENCE POLICY, PROGRAM DEVELOPMENT, AND SCIENTIFIC KNOWLEDGE

Classification for Administrative Records

It is common practice in mental health institutions to classify individuals for the sake of record keeping. Such information is not necessarily used to make decisions regarding the individual's treatment, and if it is used in this way, the earlier discussion regarding definition of an individual's problem is applicable. Though classification is often distasteful because of the oversimplification it entails, the shortcomings of most classification schemes, and the possible damage to the learner (Hobbs, 1975), some form of classification seems important because it is essential in epidemiological research, in prevention efforts, and in program description and evaluation. It is unfortunate that mental health has been burdened with a classification scheme—the American Psychiatric Association Diagnostic and Statistical Manual (1968)—in which it is difficult even to get agreement between experienced clinicians (Spitzer, Cohen, Fleiss, and Endicott, 1967); the ethics of using such a scheme is questionable.

Characteristics desirable in devices used for administrative classification. Perhaps the first requirement of devices for this purpose is interobserver agreement. Major policy decisions are often influenced by data from classifications. For example, if a federal agency is considering allocation of monies for education of "learning disabled" children, it is highly desirable that the agency have data first on the reliability with which children can be classified "learning disabled" and then on the incidence of learning disabilities among children. Similarly, it is desirable that a community mental health center have data on the incidence of developmental retardation or depression in its service area before launching programs to serve those problems.

It is also desirable that classification using different methods arrive at approximately the same result. Thus a device intended for identi-

fying schizophrenics should be able to discriminate reliably between persons in a program for schizophrenics and persons in a general hospital, a program for alcoholics, or a program for sex offenders. It should be pointed out here that such concurrent validity is *not sufficient for clinical use in placing individuals*, even though it is often cited as evidence of "the validity" of a device, as though validity were unitary.

Contribution of behavioral assessors to classification. Three clearly behavioral efforts in this area are those of Adams, Doster, and Calhoun (1977), Cautela and Upper (1975), and Ross (1974). Perhaps one of these approaches will lead to a system sufficiently reliable to be worthy of widespread use.

Evaluating Interventions during Research and Development

During the development of new interventions it is important that they be evaluated extensively and intensively. The methods used for this can be ones that would be infeasible for use in routine delivery of the same intervention. For example, it is difficult to envision routine use of the facsimile cocktail lounge for assessment of alcoholic behaviors (Mills, Sobell, and Schaefer, 1971) in many clinics treating alcoholics or routine use of vaginal blood volume measurement of sexual arousal (Geer, Morokoff, and Greenwood, 1974; Geer, 1977) with women having sexual adjustment problems. Only if the benefits of such devices are repeatedly shown to far exceed those of other devices will their use become routine. Yet their use in research is quite appropriate.

Program evaluation—in which an ongoing program is evaluated in many of its dimensions, including learner variables, process variables, outcome variables, and cost variables—has its own set of questions and requirements. Typically measurement of many variables is desirable, thus making cost per individual assessment a critical factor. In addition, the risks from errors of individual measurement are greatly reduced, freeing the evaluator to use devices that

would be unacceptable in individual evaluation for anything but screening.

The number of questions an evaluation researcher can address is infinite. Some examples are "What kinds of individual problems do the persons served by the program have?" "What kinds of environments do they come from?" "How much change in the target behaviors occurs while the learners are in the program?" "Is this due to the program or to other factors?" "What kinds of individuals are affected most readily by the program?" and "Under what circumstances do the effects endure?"

Characteristics desirable in assessment methods for research and development. In experimental research the most uniformly important characteristic of an assessment device is probably sensitivity to change. On the other hand, cost is usually of less concern than in the routine delivery of services. Other characteristics would be difficult to discuss because the research and development function is probably not one function. The value of including it here is to point out that many devices appropriate for research are not appropriate for individual case decisions.

In intensive study of a program it is appropriate to commit many more resources to assessment than in routine delivery of the same program, after it has been adequately developed and tested. In fact, it would probably be an unwise practice to revalidate continuously the program in any way that costs more than 5% of the program's resources, depending on such factors as how much the program is still developing and the number of factors militating against maintenance of quality programming (*i.e.*, social influences, response effort). Thus the investment that goes into assessment must depend upon the stage of the intervention's development, the sources of funding and their objectives, the size of the group being assessed, and similar factors.

In order to find learner or environmental variables that may *predict* success in a program, a researcher may be wise to use assessment de-

vices that tap a rather broad range of variables or are norm-referenced. Personality test scores or IQ scores may be very good predictors of success during or after a program, for example, and if so, they should be considered for use in future admissions. Until more behaviorally conceived devices are developed that serve all such predictive functions well, traditionally conceived devices will be needed.

Contribution of behavioral assessors to research and development. In addition to the behavioral assessment devices already mentioned, behavioral assessors have devised a wide array that would be useful in research. They range from a checklist for assessing street-crossing skills taught at school (Yeaton and Bailey, 1978) to a simple textual change that assesses reading comprehension (Guthrie and Tyler, 1976), and from a nonverbal measure of activity as an index of depression (Williams, Barlow, and Agras, 1972) to a partially controlled social interaction as a measure of interpersonal skills (Goldsmith and McFall, 1975).

Obtaining Normative, Descriptive Data

Though this could probably be treated as part of the research and development function, it differs in that no evaluation of a program is implied. The researcher is simply attempting to describe, to answer the question "What normally happens?" Either learner behavior or environmental factors can be measured.

Characteristics desirable in devices for normative data. Any assessment device that is useful for other functions can be useful for describing repertoires or environments. There appear to be no characteristics that are unique to this function.

Contribution of behavioral assessors to normative, descriptive research. Behavior analysts have tended to devalue normative data while developing criterion-referenced approaches and functional analyses, but there is a legitimate place for norms in behavior analysis, as Nelson and Bowles (1975) and Kazdin (1977) have pointed out. For example, Jones and Eimers

(1975) obtained rather precise normative data on the behavior of some very successful teachers and used these norms to set criterion levels of performance for less successful teachers. Walker and Hops (1976) obtained precise data on the classmates of school children who were being phased out of a special treatment program to see how closely the treated children's behavior approximately these norms, and Busse (1974) did the same kind of thing for the same purpose, using a more general behavior checklist. Finally, White (1975) measured teachers' rates of approval and disapproval, finding, among other things, that approval rates decline rapidly after the primary grades.

In some cases normative information serves as a temporary substitute for a functional analysis. Knowing what effective persons do in a particular situation can suggest what responses are functional in making them effective (Goldfried and D'Zurilla, 1969; Hawkins, 1975), though the evidence is correlational and a functional analysis is preferable. Similarly, if educators and clinicians working with developmentally retarded individuals have data on the normal sequence in which specific skills were learned, they are in a better position to construct a tentative learning sequence.

In other cases normative data are important because social norms *define* effectiveness. "For example, a male's walking or sitting in a particular 'masculine' manner will receive social acceptance (reinforcement) because it is the *norm*, while walking or sitting in a feminine fashion will reduce the overall reinforcement available to the male" (Cone and Hawkins, 1977, p. 389).

CONCLUSIONS

The technical characteristics and evidence of generalizability required of an assessment device vary greatly according to the functions it is to serve. The same device may be appropriate for one function and inappropriate for another. If the functions of an assessment are carefully con-

sidered, behavior analysts will sometimes find legitimate use for traditionally conceived devices, though probably not traditional use. Previous presentations of assessment have provided little or no analysis of the functions and their implications, and the present one is only a crude beginning.

Behavioral assessors have not completely shaken their trait thinking, as Cone and Hawkins (1977, pp. 382-388) point out, but they have contributed a wealth of assessment devices and used them for a wide variety of functions. Some of these functions had been largely neglected by traditional assessors.

The number of different assessment devices that will be needed for the various functions in clinical and educational environments will be vastly greater as the behavioral approach to assessment grows. The traditional description of "personality," "intelligence," and general academic achievement required few devices in comparison to a functional skills approach. Computerized clearinghouses for behavioral assessment devices may become necessary (Cone and Hawkins, 1977). No doubt greater standardization in using some devices will be necessary (Goldfried, 1977), but it may be possible to put such standardization into the device by providing adequate procedural description. Certainly normative standardization seems unnecessary for most behavioral assessment, but descriptive data on a particular population may often be valuable.

REFERENCE NOTE

1. Hartmann, D. P. Must the baby follow the bathwater? Psychometric principles—behavioral data. In D. H. Barlow (Chair), *Behavioral assessment: The relevance of traditional psychometric procedures*. Symposium presented at the meeting of the American Psychological Association, Washington, D.C., September 1976.

REFERENCES

- Abel, G. G., Blanchard, E. B., Barlow, D. H., and Mavissakalian, M. Identifying specific erotic cues in sexual deviations by audiotaped descriptions.

- Journal of Applied Behavior Analysis*, 1975, 8, 247-260.
- Adams, H. E., Doster, J. A., and Calhoun, K. S. A psychologically based system of response classification. In A. R. Ciminero, K. S. Calhoun, and H. E. Adams (Eds), *Handbook of behavioral assessment*. New York: Wiley, 1977.
- American Psychiatric Association. *Diagnostic and statistical manual of mental disorders, II*. Washington, D.C.: American Psychiatric Association, 1968.
- Anastasi, A. *Psychological testing* (3rd ed.), Toronto: Macmillan, 1968.
- Anastasi, A. *Psychological testing* (4th ed.), Toronto: Macmillan, 1978.
- Angle, H. V., Hay, L. R., Hay, W. M., and Ellinwood, E. H. Computer assisted behavioral assessment. In J. D. Cone and R. P. Hawkins (Eds), *Behavioral assessment: New directions in clinical psychology*. New York: Brunner/Mazel, 1977.
- Ayllon, T. and Michael, J. The psychiatric nurse as a behavioral engineer. *Journal of the Experimental Analysis of Behavior*, 1959, 2, 323-334.
- Baer, D. M., Wolf, M. M., and Risley, T. R. Some current dimensions of applied behavioral analysis. *Journal of Applied Behavior Analysis*, 1968, 1, 91-97.
- Barlow, D. H., Reynolds, E. G., and Agras, W. S. Gender identity change in a transsexual. *Archives of General Psychiatry*, 1973, 28, 569-579.
- Beck, A. T. *Depression: Causes and treatment*. Philadelphia: University of Pennsylvania Press, 1972.
- Bijou, S. W., Peterson, R. F., and Ault, M. A method to integrate descriptive and experimental field studies at the level of data and empirical concepts. *Journal of Applied Behavior Analysis*, 1968, 1, 175-191.
- Bijou, S. W. and Baer, D. M. *Behavior analysis of child development*. Englewood Cliffs, N.J.: Prentice-Hall, 1978.
- Bushell, D., Jr., Jackson, D. A., and Weis, L. C. Quality control in the behavior analysis approach to Project Follow Through. In W. S. Wood (Ed), *Issues in evaluating behavior modification*. Champaign, Ill.: Research Press, 1975.
- Busse, W. J. *A normative and validation study of the Behavior Problem Checklist*. Unpublished masters thesis, Western Michigan University, 1974.
- Carver, R. P. Two dimensions of tests: Psychometric and edumetric. *American Psychologist*, 1974, 19, 512-518.
- Cautela, J. R. and Upper, D. The process of individual behavior therapy. In M. Hersen, R. M. Eisler, and P. M. Miller (Eds), *Progress in behavior modification*, Vol. 1. New York: Academic Press, 1975.
- Chesney, M. A. and Tasto, D. L. The development of the menstrual symptom questionnaire. *Behaviour Research and Therapy*, 1975, 13, 237-244.
- Cone, J. D. The relevance of reliability and validity for behavioral assessment. *Behavior Therapy*, 1977, 8, 411-426.
- Cone, J. D. and Hawkins, R. P. (Eds), *Behavioral assessment: New directions in clinical psychology*. New York: Brunner/Mazel, 1977.
- Cowen, E. L., Dorr, D., Clarfield, S., Kreling, B., McWilliams, S. A., Pokracki, F., Pratt, M., Terrell, D., and Wilson, A. The AML: A quick-screening device for early identification of school maladaptation. *American Journal of Community Psychology*, 1973, 1, 12-35.
- Cronbach, L. J. *Essentials of psychological testing* (3rd ed.). New York: Harper & Row, 1970.
- Cronbach, L. J., Gleser, G. C., Nanda, H., and Rajaratnam, N. *The dependability of behavioral measures*. New York: Wiley, 1972.
- Daley, M. F. The "Reinforcement Menu"—Finding effective reinforcers. In E. J. Mash & L. G. Terdal (Eds), *Behavior-therapy assessment: Diagnosis, design, and evaluation*. New York: Springer, 1976.
- Eaton, M. D. and Lovitt, T. C. Achievement tests vs. direct and daily measurement. In G. Semb, D. R. Green, R. P. Hawkins, J. Michael, E. L. Phillips, J. A. Sherman, H. Sloane, and D. R. Thomas (Eds), *Behavior analysis and education—1972*. Lawrence: University of Kansas Center for Follow Through, 1972.
- Ferster, C. B. and DeMeyer, M. K. A method of experimental analysis of the behavior of autistic children. *American Journal of Orthopsychiatry*, 1962, 32, 89-98.
- Freedman, B. J. An analysis of social-behavioral skill deficits in delinquent and non-delinquent adolescent boys. Unpublished doctoral dissertation, University of Wisconsin, 1974.
- Geer, J. H. Sexual functioning: Some data and speculations on psychophysiological assessment. In J. D. Cone & R. P. Hawkins (Eds), *Behavioral assessment: New directions in clinical psychology*. New York: Brunner/Mazel, 1977.
- Geer, J. H., Morokoff, D., and Greenwood, P. Sexual arousal in women. The development of a measurement device for vaginal blood volume. *Archives of Sexual Behavior*, 1974, 3, 599-564.
- Gilbert, T. F. *Human competence: Engineering worthy performance*. New York: McGraw-Hill, 1978.
- Goldfried, M. R. Behavioral assessment in perspective. In J. D. Cone & R. P. Hawkins (Eds), *Behavioral assessment: New directions in clinical psychology*. New York: Brunner/Mazel, 1977.
- Goldfried, M. R. and D'Zurilla, T. J. A behavioral-analytic model for assessing competence. In C. D. Spielberger (Ed), *Current topics in clinical and community psychology*. New York: Academic Press, 1969.

- Goldfried, M. R. and Pomeranz, D. Role of assessment in behavior modification. *Psychological Reports*, 1968, **23**, 75-87.
- Goldsmith, J. B. and McFall, R. M. Development and evaluation of interpersonal skill-training programs for psychiatric inpatients. *Journal of Abnormal Psychology*, 1975, **84**, 51-58.
- Guidance Associates of Delaware, Inc. *Wide Range Achievement Test*. Wilmington: Guidance Assoc. of Delaware, 1978.
- Guthrie, J. T. and Tyler, J. Operational definitions of reading. In T. A. Brigham, R. Hawkins, J. W. Scott, and T. F. McLaughlin (Eds), *Behavior analysis in education: Self-control and reading*. Dubuque, Iowa: Kendall/Hunt, 1976.
- Hall, R. V. *Behavior modification, Vol. 1: The measurement of behavior*. Lawrence, Kans.: H & H Enterprises, 1971.
- Hansen, C. L. and Lovitt, T. C. Reading: Round one—matching the child to the book. In T. A. Brigham, R. P. Hawkins, J. W. Scott, and T. F. McLaughlin (Eds), *Behavior analysis in education: Self-control and reading*. Dubuque, Iowa: Kendall/Hunt, 1976.
- Harcourt Brace Jovanovich. *Skills Monitoring System*. New York: Harcourt Brace Jovanovich, 1975.
- Hart, B. M., Allen, K. E., Buell, J. S., Harris, R. F., and Wolf, M. M. Effects of social reinforcement on operant crying. *Journal of Experimental Child Psychology*, 1964, **1**, 145-153.
- Hawkins, R. P. Who decided that was the problem? Two stages of responsibility for applied behavior analysis. In W. S. Wood (Ed), *Issues in evaluating behavior modification*. Champaign, Ill.: Research Press, 1975.
- Hawkins, R. P., Axelrod, S., and Hall, R. V. *Teachers as behavior analysts: Precisely monitoring student performance*. In T. A. Brigham, R. P. Hawkins, J. W. Scott, and T. F. McLaughlin (Eds), *Behavior analysis in education: Self control and reading*. Dubuque, Iowa: Kendall/Hunt, 1976.
- Hawkins, R. P., Peterson, R. F., Schweid, E., and Bijou, S. W. Behavior therapy in the home: Amelioration of problem parent-child relations with the parent in a therapeutic role. *Journal of Experimental Child Psychology*, 1966, **4**, 99-107.
- Hersen, M. and Barlow, D. H. *Single case experimental designs: Strategies for studying behavior change*. New York: Pergamon, 1976.
- Hobbs, N. *The futures of children*. San Francisco: Jossey-Bass, 1975.
- Holland, C. An interview guide for behavioral counseling with parents. *Behavior Therapy*, 1970, **1**, 70-79.
- Houghton Mifflin. *Iowa Test of Basic Skills*. Boston: Houghton Mifflin, 1974.
- Jones, F. H. and Eimers, R. C. Role playing to train elementary teachers to use a classroom management "skill package." *Journal of Applied Behavior Analysis*, 1975, **8**, 421-433.
- Kanfer, F. H. and Saslow, G. Behavioral diagnosis. In C. M. Franks (Ed), *Behavior therapy: Appraisal and status*. New York: McGraw-Hill, 1969.
- Kazdin, A. E. Assessing the clinical or applied importance of behavior change through social validation. *Behavior Modification*, 1977, **1**, 427-452.
- Kent, R. N. and Foster, S. L. Direct observational procedures: Methodological issues in naturalistic settings. In A. R. Ciminero, K. S. Calhoun, and H. E. Adams (Eds), *Handbook of behavioral assessment*. New York: Wiley, 1977.
- Kubany, E. S. and Sloggett, B. A. A coding procedure for teachers. *Journal of Applied Behavior Analysis*, 1973, **6**, 339-344.
- Lindsley, O. R. Reduction in rate of vocal psychotic symptoms by differential positive reinforcement. *Journal of the Experimental Analysis of Behavior*, 1959, **2**, 269.
- Linehan, M. Issues in behavioral interviewing. In J. D. Cone & R. P. Hawkins (Eds), *Behavioral assessment: New directions in clinical psychology*. New York: Brunner/Mazel, 1977.
- Livingston, S. A. Psychometric techniques for criterion-referenced testing and behavioral assessment. In J. D. Cone & R. P. Hawkins (Eds), *Behavioral assessment: New directions in clinical psychology*. New York: Brunnel/Mazel, 1977.
- Lobitz, G. K. and Johnson, S. M. Normal versus deviant children: A multimethod comparison. *Journal of Abnormal Child Psychology*, 1975, **3**, 353-374.
- LoPiccolo, J. and Steger, J. C. The Sexual Interaction Inventory: A new instrument for assessment of sexual dysfunction. *Archives of Sexual Behavior*, 1974, **3**, 585-595.
- Mahoney, M. J. Some applied issues in self-monitoring. In J. D. Cone & R. P. Hawkins (Eds), *Behavioral assessment: New directions in clinical psychology*. New York: Brunner/Mazel, 1977.
- Mash, E. J. and Terdal, L. G. (Eds) *Behavior-therapy assessment*. New York: Springer, 1976.
- McFall, R. M. Analogue methods of behavioral assessment: Issues and prospects. In J. D. Cone & R. P. Hawkins (Eds), *Behavioral assessment: New directions in clinical psychology*. New York: Brunner/Mazel, 1977.
- McFall, R. M. and Lillesand, D. V. Behavior rehearsal with modeling and coaching in assertive training. *Journal of Abnormal Psychology*, 1971, **77**, 313-323.
- Mills, K. C., Sobell, M. B., and Schaefer, H. H. Training social drinking as an alternative to abstinence for alcoholics. *Behavior Therapy*, 1971, **2**, 18-27.
- Morganstern, K. P. Behavioral interviewing: The initial stage of assessment. In M. Hersen & A. S.

- Bellack (Eds), *Behavioral assessment: A practical handbook*. New York: Pergamon, 1976.
- Nay, W. R. Analogue measures. In A. R. Ciminero, K. S. Calhoun, & H. E. Adams (Eds), *Handbook of behavioral assessment*. New York: Wiley, 1977.
- Nelson, R. O. Methodological issues in assessment via self-monitoring. In J. D. Cone & R. P. Hawkins (Eds), *Behavioral assessment: New directions in clinical psychology*. New York: Brunner/Mazel, 1977.
- Nelson, R. O. and Bowles, P. E., Jr. The best of two worlds: Observations with norms. *Journal of School Psychology*, 1975, **13**, 3-9.
- Phillips, E. L., Phillips, E. A., Wolf, M. M., and Fixsen, D. L. Achievement Place: Development of the elected manager system. *Journal of Applied Behavior Analysis*, 1973, **6**, 541-561.
- Popovich, D. *A prescriptive behavioral checklist for the severely and profoundly retarded*. Baltimore: University Park Press, 1977.
- Psychological Corporation. *Mooney Problem Checklist*. New York: Psychological Corporation, 1950.
- Rathus, S. A. A 30-item schedule for assessing assertive behavior. *Behavior Therapy*, 1973, **4**, 398-406.
- Resnick, L. B., Wang, M. C., and Kaplan, J. Task analysis in curriculum design: A hierarchically sequenced introductory mathematics curriculum. *Journal of Applied Behavior Analysis*, 1973, **6**, 679-709.
- Ross, A. O. Psychological disorders of children: *A behavioral approach to theory, research, and therapy*. New York: McGraw-Hill, 1974.
- Scholastic Testing Service. *STS youth inventory*. Bensenville, Ill.: Scholastic Testing Service, 1971.
- Schwartz, R. M. and Gottman, J. M. Towards a task analysis of assertive behavior. *Journal of Consulting and Clinical Psychology*, 1976, **44**, 910-920.
- Schwarz, M. L. and Hawkins, R. P. Application of delayed reinforcement procedures to the behavior problems of an elementary school child. *Journal of Applied Behavior Analysis*, 1970, **3**, 85-96.
- Spitzer, R. L., Cohen, J., Fleiss, J. L., and Endicott, J. Quantification of agreement in psychiatric diagnosis. *Archives of General Psychiatry*, 1967, **17**, 83-87.
- Stuart, R. B. and Stuart, F. M. Prestructuring behavior therapy through precounseling assessment. In E. J. Mash & L. G. Terdal (Eds), *Behavior-therapy assessment: Diagnosis, design, and evaluation*. New York: Springer, 1976.
- Trap, J. J., Milner-Davis, P., Joseph, S., and Cooper, J. O. The effects of feedback and consequences on transitional cursive letter formation. *Journal of Applied Behavior Analysis*, 1978, **11**, 381-393.
- Walls, R. T., Zane, T., and Werner, T. J. *The Vocational Behavior Checklist* (Experimental Edition). Morgantown, W. Va.: Rehabilitation Research and Training Center (509 Allen Hall, West Virginia University), 1978.
- Walker, H. M. and Hops, H. Use of normative peer data as a standard for evaluating classroom treatment effects. *Journal of Applied Behavior Analysis*, 1976, **9**, 159-168.
- White, M. A. Natural rates of teacher approval and disapproval in the classroom. *Journal of Applied Behavior Analysis*, 1975, **8**, 367-372.
- Wildman, B. G. and Erickson, M. T. Methodological problems in behavioral observation. In J. D. Cone & R. P. Hawkins (Eds), *Behavioral assessment: New directions in clinical psychology*. New York: Brunner/Mazel, 1977.
- Williams, J. G., Barlow, D. H., and Agras, W. S. Behavioral measurement of severe depression. *Archives of General Psychiatry*, 1972, **27**, 330-333.
- Wolf, M. M. Social validity: The case for subjective measurement or how applied behavior analysis is finding its heart. *Journal of Applied Behavior Analysis*, 1978, **11**, 203-214.
- Yeaton, W. H. and Bailey, J. S. Teaching pedestrian safety skills to young children: An analysis and one-year follow-up. *Journal of Applied Behavior Analysis*, 1978, 315-329.

Received 6 October 1978.

(Final Acceptance 20 July 1979.)