THE NATURE OF BEHAVIORAL ASSESSMENT: A COMMENTARY

ROSEMERY O. NELSON AND STEVEN C. HAYES

THE UNIVERSITY OF NORTH CAROLINA AT GREENSBORO

This special issue of the *Journal of Applied Behavior Analysis* provides the reader with a sample of current work in behavioral assessment. The purpose of this paper is to present an overview of behavioral assessment and to place the other articles in context of this developing area.

The last five years have witnessed the emergence of behavioral assessment as a field of professional attention in its own right. Several recent developments have marked this emergence. A number of books on this topic have appeared: Ciminero, Calhoun, and Adams (1977); Cone and Hawkins (1977); Haynes (1978); Hersen and Bellack (1976); Keefe, Kopel, and Gordon (1978); and Mash and Terdal (1976). An annual review by Haynes and Wilson (*in press*) has begun. The first issues of two new journals are about to appear: *Behavioral Assessment* (Pergamon and the Association for Advancement of Behavior Therapy) and *Journal of Behavioral Assessment* (Plenum). Finally, the *Journal of Applied Behavior Analysis* is devoting this entire issue to behavioral assessment. In this article, we will briefly examine some of the main characteristics and issues in behavioral assessment (after Nelson and Hayes, 1979) and show how the articles presented here exemplify them.

WHAT IS BEHAVIORAL ASSESSMENT?

Behavioral assessment is the identification of meaningful response units and their controlling variables (both current environmental and organismic) for the purposes of understanding and altering human behavior. These activities

have been integral to behavior analysis and therapy since its beginning. Initially, however, much more emphasis was understandably placed on treatment, intervention, and independent variables than on design, methodology, and dependent variables. Only when evidence of treatment effectiveness began to accumulate were assessment questions meaningful: Are these appropriate behaviors to alter? Are these valid and reliable ways of measuring these behaviors? Can these changes be attributed to the intervention? Was this the most effective intervention to use?

Behavioral assessment and traditional assessment share some techniques, *e.g.*, interviewing, questionnaires, observations of behavior. The two approaches differ radically, however, in their assumptions and levels of inference. In traditional assessment, it is frequently assumed that behavior is a function of relatively stable intra-organismic (*e.g.*, intrapsychic) variables. Hence, behavior or appearance is interpreted as a *sign* of these underlying variables (Goldfried and Kent, 1972; Goodenough, 1949). Since the causes of behavior lie within the person, the assessment situation is thought to be of little relevance. Since these causes are stable, assessment techniques can seemingly be evaluated by

Requests for reprints should be addressed to Rosemary O. Nelson, Psychology Department, University of North Carolina at Greensboro, Greensboro, North Carolina 27412.

their psychometric properties. In contrast, behavioral assessment emphasizes environmental as well as organismic determinants of behavior. Thus, behavior is viewed as a *sample* of responding in a particular assessment situation (Goldfried and Kent, 1972; Goodenough, 1949). Without empirical justification, inferences are not made beyond the present behavior and situation to underlying causes, other responses, or different situations. Psychometric evaluation may be necessary, but it is not a sufficient measure of the quality of behavioral assessment. Functional utility must also be demonstrated: Does this behavioral assessment enhance our understanding of behavior and/or our ability to alter it?

GOALS OR FUNCTIONS OF BEHAVIORAL ASSESSMENT

The goals or functions of behavioral assessment may be categorized by several different schemata. One schema distinguishes between the assessment of individuals *versus* the assessment of programs. Each type of assessment may require particular assessment methods. For example, in this issue Hawkins (1979) describes the following stages in the assessment of individuals: (1) screening and general disposition; (2) definition and general quantification of the problem or achievement; (3) pinpointing and design of intervention; (4) monitoring of progress; and (5) follow-up. Historically, behavioral assessment has focused on these latter four stages. Two articles in this issue, however, provide good examples of the screening stage of behavioral assessment, in this case, screening children for possible problems of social withdrawal. Foster and Ritchey (1979) discuss the assets and limitations of sociometric measures (peer nomination and peer rating scales) in identifying socially competent and incompetent children. Greenwood, Walker, Todd, and Hops (1979) establish an empirical relationship between two screening measures, teacher judg-

ment (rankings and ratings) and peer judgment (sociometric ratings), and interactive behavior observed in the criterion situations of free play and structured tasks.

In addition to assessing individuals, behavior assessment must sometimes be undertaken for institutional or programmatic purposes. Hawkins (1979) describes several of these purposes: classification of individuals for administrative record keeping, program evaluation, and the collection of normative data. Hawkins (1979) notes several idiosyncratic classification systems developed by behavioral assessors (*e.g.*, Cautela and Upper, 1975). Alternatively, perhaps the improved reliability of DSM-III (American Psychiatric Association, 1978) will make it suitable for administrative and classificatory purposes, with an adjunct functional analysis performed by behavior assessors for intervention purposes. Program evaluation has a unique set of considerations and constraints. Some of these are nicely described in the present issue by Schnelle, Kirchner, Galbaugh, Domash, Carr, and Larson (1979) and by Filipczak, Archer, Neale, and Winett (1979). These articles point out the opportunity during exploratory stages for a high cost-benefit ratio; the need for the use of multiple dependent measures; the operational constraints imposed by existent systems and bureaucracies; and practical considerations which threaten the internal and external validity of experimental designs.

Besides the individual-programmatic distinction, a second schema to classify the functions or goals of behavioral assessment is descriptive *versus* experimental work (Bijou, Peterson, and Ault, 1968). Both functions are important, but at different stages in the development of behavioral assessment. Descriptive studies are static or structural in that they describe an existent state of affairs. In this issue, for example, Henson, Rubin, and Henson (1979) describe intra-individual stability in women's sexual arousal across recording sessions and across measures of arousal; Williams (1979) shows that happily

married and distressed couples may be differentiated on their temporal patterns of interaction as well as on the ratio of positive to negative interactions; and Prinz, Foster, Kent, and O'Leary (1979) find that distressed and nondistressed mother-adolescent dyads may be differentiated on the bases of maternal and adolescent reports of behavior, and of independent ratings of tape-recorded interactions. Such descriptive studies in behavioral assessment are often useful in describing behavioral patterns, developing assessment methodologies, and suggesting relationships which may be used to increase our clinical impact. These studies, therefore, are useful in the early stages of development in particular assessment areas by providing a solid base for further work. They do not, however, ask or answer *why* particular results were obtained; the controlling variables of which these results are a function are not identified. A complete understanding of behavior requires both descriptive and experimental findings. Examples of functional or experimental studies in this issue include DeProspero and Cohen's examination (1979) of the independent variables (pattern and degree of mean shift between experimental phases, slope and variability of data within phases) which determine judgments of significant effects in visual analyses of data in reversal designs; and Martinez-Diaz and Edelstein's examination (1979) of the effects of demand characteristics (pretreatment assessment *vs.* experiment; competent behavior *vs.* neutral) on social skills during analogue assessment.

Finally, a distinction can be made between nomothetic and idiographic conclusions. The results of both descriptive and experimental studies are nomothetic, whereas behavioral assessment is always conducted idiographically. Research provides general guidelines that are useful in the assessment of the individual client. Currently, our knowledge of how best to move from nomothetic to idiographic levels is limited. That, in itself, may be an important future direction in behavioral assessment research.

CURRENT TRENDS IN BEHAVIORAL ASSESSMENT

Identification of Target Behavior

Triple response system and response covariation. Most contemporary behaviorists consider all forms of organismic activity to be "behavior." Thus, behavioral assessment includes the measurement of overt motor, physiological-emotional, and cognitive-verbal behavior. All three response systems may provide important target behaviors. In this issue, for example, the motor behavior of children in a classroom setting was recorded by Kent, O'Leary, Dietz, and Diamant (1979); the physiological responses indicative of women's sexual arousal was measured by Henson *et al.* (1979); and the verbal responses of mother-adolescent dyads were assessed via questionnaires by Prinz *et al.* (1979). Although these three types of behavior may covary, it is not assumed that such covariation occurs (Lang, 1968). Previous studies have often found only moderate correlations among the three response systems (*e.g.*, Hartshorne and May, 1928). Thus, the relationship among measures taken from the three response systems is frequently examined empirically. Several examples are provided in this issue. Henson *et al.* (1979) examined the relationship between physiological measures of sexual arousal and subjective report of arousal; and Greenwood *et al.* (1979) examined the correspondence between children's actual observed rates of social interaction and teacher or peer judgments about rates of interactions. These types of studies are valuable in helping us to assess a fuller range of behavior in the applied setting.

Even within a particular response system, it cannot be assumed that different responses covary. Behavioral assessors have responded to this by taking multiple measures so as to ensure a broad assessment. In the assessment of smoking, for example, Frederiksen, Martin, and Webster (1979) argue that frequency and topography of smoking as well as substance smoked

must be recorded. Similarly, several different questionnaires were administered to the adolescent-mother dyads by Prinz *et al.* (1979); and academic achievement was measured by standardized achievement tests, completion of assignments, and class grades by Filipczak *et al.* (1979).

Behavioral assessment techniques. The assessment technique selected must suit the response system that is being measured and reveal its important dimensions. Hence, overt motor behavior is most frequently measured by direct observation by trained observers (*e.g.*, Kent *et al.*, 1979), by mediators, or by self-recorders (*e.g.*, Williams, 1979) in either naturalistic (*e.g.*, Greenwood, Greenwood *et al.*, 1979) or analogue situations (*e.g.*, Martinez-Diaz and Edelstein, 1979). Physiological-emotional responses are usually measured in analogue situations, due to instrumentation requirements. For example, Henson *et al.* (1979) used an erotic film while measuring women's sexual arousal, and Martinez-Diaz and Edelstein (1979) used a contrived conversation while measuring heart rate. Cognitive-verbal responses are measured *via* rating scales (*e.g.*, Williams, 1979), checklists (*e.g.*, Prinz *et al.*, 1979), questionnaires (*e.g.*, Martinez-Diaz and Edelstein, 1979), and academic tests (*e.g.*, Filipczak *et al.*, 1979).

In addition to direct assessment by the three response systems, sometimes the by-products of those responses are examined. For example, Schnelle *et al.* (1979) used archival records to determine the number of reported robberies, number of arrests, and amount of recovered stolen money. Frederiksen *et al.* (1979) provide another example by advocating the analysis of chemical by-products of smoking. As discussed by Kazdin (1979), such indirect measures have the advantage of being less obtrusive and consequently less reactive than more direct measures.

In addition to suiting the response system, the assessment technique should also suit the stage of assessment. Devices used during screening should be less costly in terms of time and money

than devices used to monitor a target behavior during intervention (Hawkins, 1979). This is one reason why Foster and Ritchey (1979) and Greenwood *et al.* (1979) theoretically and empirically examined the use of a relatively inexpensive screening device—peer judgments—to predict children's observed social withdrawal. If resources were no object it would be better to measure the behavior directly. Those in the actual applied environment, however, do have limited resources; and behavioral assessment is beginning to attempt to provide them the assessment tools they need.

Meaningful target behaviors. An alternative to selecting target behaviors on the basis of their face validity is to establish empirically the importance of particular behaviors. One strategy is the known groups method, that is, identifying two groups of subjects that differ on a relevant dimension and then seeking specific behaviors that differentiate the two groups. For example, in this issue, Williams (1979) used this method to identify behavioral patterns of social interaction which distinguished happy and distressed couples, while Prinz *et al.* (1979) reported that distressed and nondistressed adolescent-mother dyads differed on particular self-report measures and on ratings of tape-recorded interactions. These types of studies point to several possible targets for clinical intervention. The next step, however, should also be taken to demonstrate that beneficial results are produced by changing these targets.

Another empirical method of establishing the importance of particular target behaviors is to examine subjective ratings by others (Wolf, 1978). For example, the specific behaviors comprising desirable social interactions by children may be identified in relation to their ability to produce positive social ratings by peers and/or teachers (Foster and Ritchey, 1979; Greenwood *et al.*, 1979). It may be important, however, to show that changes in these behaviors result in other positive benefits beyond the verbal evaluation by others.

Yet another method is to relate particular dependent variables to a variable that is consensually acknowledged to be important. Thus Fredriksen *et al.* (1979) argue for the measurement of several aspects of smoking (frequency, topography, substance smoked) because of their relationship to the smoker's health. This also points to the importance of measuring the most meaningful *dimension* of a target behavior. Behavioral assessors may at times overemphasize rate or frequency due to the history of behavioral interventions.

Additional empirical means of establishing the importance of target behaviors, not represented in this issue, include: the collection of normative data (Kazdin, 1977; Nelson and Bowles, 1975), the use of task analysis and of developmental norms (Hawkins, 1975), and the establishment of regression equations (Cobb, 1972).

On a philosophical level, the specific behaviors selected for modification inherently involve a value judgment on the part of both client and therapist. Two articles in this issue are concerned that mediocrity may become the standard by which behavior analysts evaluate the need for and outcome of treatment. Foster and Ritchey (1979) argue that children's social competence is more than the absence of incompetence, stating that "socially competent behavior [are] those responses which within a given situation prove effective, or, in other words, maximize the probability of producing, maintaining, or enhancing positive effects for the interactor." Van Houten (1979) asserts that, in general, competency or excellence should be the standard of evaluation. He suggests empirical procedures by which to accomplish this goal: identify the specific behaviors of individuals with known competence, train individuals to different rates or levels and ascertain the effects on future learning or on competency ratings, and vary the parameters of the target behavior to determine which are most effective in achieving the relevant goal.

Identification of Controlling Variables

Besides identifying the behavior to be changed, a second goal of behavioral assessment is to identify the variables that control the occurrence of that behavior. In general, two classes of controlling variables are considered: current environmental variables (antecedent and consequent stimuli) and organismic variables (individual differences produced by physiology and by past learning). The relevant aspects of behavioral assessment have been neatly summarized by Goldfried and Sprafkin's acronym (1976) of SORC (stimulus-organism-response-consequence).

Organismic variables. Organismic variables include individual differences produced by past learning and by physiology. For example, in this issue, Henson *et al.* (1979) found large intersubject differences in measures of sexual arousal, and Martinez-Diaz and Edelstein (1979) resorted to an analysis of covariance because of baseline differences in subjects' heart rate. Many behavioral assessors have somewhat neglected organismic variables because they are frequently not amenable to changes or because they smack of traditional assessment. However, the assessment of organismic variables seem necessary for an adequate understanding and, ultimately, control of behavior. At times, organismic variables can be altered within an intervention program (*e.g.*, fatigue, hunger, unkempt appearance); or they may often be targeted for prevention-oriented programs.

Situation specificity. Most behavioral assessors hold an interactionist view, that behavior is the result of an interaction between the current situation and individual differences. The situational specificity of behavior has been well documented (Mischel, 1968). The central concern that situational specificity raises for behavioral assessment is the need to demonstrate that conclusions reached in the assessment situation can be generalized to the criterion "real-life" situation. If practical considerations preclude assess-

ment directly in the criterion situation, attempts are frequently made to reproduce aspects of the criterion situation within the assessment situation in an analogue or contrived manner. Henson *et al.*'s (1979) use of erotic films and Martinez-Diaz and Edelstein's (1979) use of an interaction with a female confederate provide examples. Ultimately, the validity of this method must still be demonstrated, however.

Given the situational specificity of behavior, many aspects of the assessment situation may affect the subject's behavior. Because of the possible reactivity of obtrusive assessment techniques, Kazdin (1979) advocates the use of unobtrusive measures. Martinez-Diaz and Edelstein (1979) found that instructions to act in a competent fashion *versus* neutral instructions did not affect heterosocial behavior, but portraying the measurement situation as pretreatment assessment *versus* an experiment did affect at least a few of the dependent variables. This points to the care with which behavior therapists must construct their assessment methods.

Not only is the behavior of the client but also the behavior of the assessor subject to situational variables. DeProspero and Cohen (1979) reported that the visual analyses of graphs depicting typical data from reversal designs were influenced by the pattern and degree of mean shift across experimental phases, and by the variability and slope of the data within experimental phase. In an investigation of the effects of observation media on data collection, Kent *et al.* (1979) found no differences in occurrence reliability among observations performed *in vivo*, through a one-way mirror, or from videotapes, and found differences in observed behavioral frequencies on only one of nine observational categories.

The importance of observed situational differences is often a matter of investigation. As will be discussed below, situational specificity need not indict the quality of behavioral assessment. However, only by understanding the variables producing this specificity can we have

confidence that our assessment results reflect meaningful information.

Evaluating the Quality of Behavioral Assessment

Psychometrics and generalizability theory. Classic psychometric theory is based on the assumption that the causes of behavior are relatively stable and intraorganismic. Because consistent responding is predicted, traditional assessment devices may be evaluated in terms of their psychometric properties. Consistency is expected across time (test-retest reliability), across items (inter-item consistency), across assessment situations and techniques (concurrent validity), and across assessment techniques and time (predictive validity). The various types of reliability and validity have been categorized more systematically within generalizability theory (Cronbach, Gleser, Nanda, and Rajaratnam, 1972), which relabels the types of reliability and validity into universes of generalization (*e.g.*, across scorer, item, time, setting, or method).

It has been suggested that psychometric or generalizability *procedures* be applied to the techniques of behavioral assessment, with recognition of the different *assumptions* between behavioral and traditional assessment (Cone, 1977; Jones, 1977). Although issues of generalizability are important, demonstrations of generalizability or the lack thereof cannot be used as the sole criterion of the quality of behavioral assessment for two reasons. First, inconsistencies in measurement may be produced by actual changes in behavior and not by an imprecise behavioral assessment technique (Nelson, Hay, and Hay, 1977). It is only when behavior is static or somehow "captured" (*e.g.*, videotaped) so that it cannot change that consistency would be expected from behavioral assessment devices. Second, the quality of behavioral assessment must ultimately be determined by its functional value in increasing our scientific understanding of behavior and the success of our therapeutic interventions (Nelson and Hayes, 1979). Static

or structural evaluation of psychometric properties may be a necessary but not a sufficient mark of quality.

The most frequent examination of reliability within behavioral assessment is interobserver agreement, that is, an evaluation of the consistency in recording among two or more observers. This procedure is theoretically defensible as a structural evaluation of behavioral observation because consistency is expected, since the observers are observing the same subject at the same time. Although reports of interobserver agreement have become routine, much controversy yet rages over the best means of calculating and of reporting interobserver agreement, as evidenced by articles in this issue authored by Birkimer and Brown (1979a, 1979b), by Kratochwill (1979), by Cone (1979), by Yelton (1979), by Hopkins (1979), by Hartmann and Gardner (1979), and by Hawkins and Fabry (1979). Similarly, DeProspero and Cohen's examination (1979) of the consistency of visual analyses of typical data from reversal designs relies on structural properties to evaluate the quality of visual evaluation of data. This may be justifiable because the judges examined the same graphs; consistency might well be expected, although it was not found.

Generalizability can be an interesting issue in its own right, without its being a reflection of the quality of an assessment device. For example, Kent *et al.*'s (1979) finding regarding the consistency in observational recordings despite alterations in the observational medium (*in vivo*, videotapes, one-way mirror) relate more to the use of assessment devices than to their quality *per se*. Henson *et al.* (1979) examined generalizability both across assessment techniques and time; there was high intrasubject consistency for most but not all subjects. This last study well exemplifies the separate issues of generalizability *versus* quality. Inconsistent responding was attributed not to poor assessment but rather to changing organismic variables, such as hormone level, or to chang-

ing situational variables, such as placement of the measurement device.

Conceptual validity. One function of behavioral assessment is to enhance our understanding of behavior and its controlling variables, even when response alteration is not desired or is not possible. Behavioral conceptualizations may be expanded by both descriptive studies which describe existent patterns and relationships, and by experimental studies which examine why those patterns and relationships occur (Bijou *et al.*, 1968). Several of the studies in this issue have clear conceptual implications.

Treatment validity. Another way that behavioral assessment may be evaluated is in its contribution to treatment success, termed evaluation of treatment validity (Angle, Note 1). Nomothetically, the nature of the target behavior may suggest several intervention strategies that have had a history of success. For example, with heterosocial difficulties, systematic desensitization, behavioral rehearsal, cognitive restructuring, and practice dating have been used. The treatment of choice in any particular case of heterosocial difficulties would depend on the response repertoire and controlling variables identified in the assessment. To demonstrate the treatment validity of the assessment used, it must be shown that treatment success was enhanced by this assessment.

Any experiment with internal validity tests treatment validity in a primitive fashion because such experiments demonstrate that changes in the dependent variable were most likely caused by manipulation of the independent variable. Thus conceptions of behavioral assessment generally encompass concerns about experimental design; program evaluation; statistical and visual analyses (*e.g.*, DeProspero and Cohen, 1979); monitoring of progress before, during, and after intervention; and social validity of outcome.

Experiments with internal validity may determine the efficacy of manipulating the chosen independent variable, but such experiments do

not generally determine the relative efficacy of having selected one target behavior over another or of manipulating one independent variable as opposed to another (Nelson and Hayes, 1979). For example, if two different assessment techniques led to the identification of two different target behaviors, the treatment validity of each could be assessed by comparing overall improvement when intervention is directed toward one target or the other. As another example, in treating a specific disorder the relative merits could be determined of performing idiographic functional analyses *versus* general implementation of a powerful treatment package. Similarly, the outcomes of differing treatments suggested by different assessment strategies could be compared (*e.g.*, by an alternating treatments design, Barlow and Hayes, 1979) to evaluate the treatment validity of these assessment strategies. It may be a mark of the infancy of the area that the behavioral assessment literature (and this issue) has paid little attention to treatment validity. Ultimately, however, the value of all of the assessment devices and methods described in this issue must be measured by the impact they have on our understanding and control of behavior.

FUTURE DIRECTIONS OF BEHAVIORAL ASSESSMENT

Expansion of Settings, Techniques, and Target Behaviors

Behavioral assessment is beginning to expand in settings beyond the clinic, home, and school to other settings, for example, the community (Schnelle *et al.*, 1979) and business (Komaki, Collins, and Thoene, *in press*). Techniques are expanding beyond observations in natural or contrived settings, physiological recordings, and questionnaires. New techniques are being developed, such as the Marital Satisfaction Time Lines (Williams, 1979). Other techniques are being borrowed from traditional assessment, such as peer ratings and nominations (Foster

and Ritchey, 1979; Greenwood *et al.*, 1979). Complicated target behaviors are being assessed, such as social competence and interpersonal interaction (Foster and Ritchey, 1979; Greenwood *et al.*, 1979; Martinez-Diaz and Edelman, 1979; Williams, 1979).

Multivariate Statistics

Increased attention to complex human behavior has brought the use of multiple dependent measures, often generated by multiple assessment techniques. This trend will create pressure for greater reliance on computers and complex statistical decision rules. In this issue, multivariate statistics were used by Martinez-Diaz and Edelman (1979) to examine the effects of their independent variables on a large number of dependent variables, and by Prinz *et al.* (1979) to determine which of their dependent variables best discriminated groups of distressed and non-distressed mother-adolescent dyads.

Generalizability and Psychometric Evaluation

Issues of reliability and validity, perhaps couched in terms of generalizability, are interesting issues, apart from their evaluation functions. The most important generalizability issue is under what conditions can generalization be assumed from the assessment situation to the criterion situation. Research on this question will undoubtedly expand.

Conceptual Validity

Descriptive studies presenting response covariations, normative data, intertechnique correlations, and known group differences will continue. More attention, however, will be given to experimental studies that investigate the *determinants* of these response occurrences.

Treatment Validity

Although the advantages of behavioral assessment and of functional analyses are frequently assumed, more evidence of their treatment validity is needed. What assessment procedures do,

in fact, lead to the most efficacious treatment outcome? Without this information, the promise of behavioral assessment cannot be fulfilled.

CONCLUSION

This special issue of the *Journal of Applied Behavior Analysis* provides the reader with a sample of current work in behavioral assessment. The purpose of this paper was to present an overview of behavioral assessment and to place the other articles in context of this developing area. From these articles, behavioral assessment can be seen to be a broad and exciting area, but one with enormous challenges still ahead.

REFERENCE NOTE

1. Angle, H. V. Personal communication. 1975.

REFERENCES

- American Psychiatric Association. Diagnostic and statistical manual of mental disorders (3rd ed.). Washington, D.C.: American Psychiatric Association, 1978.
- Barlow, D. H. and Hayes, S. C. Alternating treatments design: One strategy for comparing the effects of two treatments in a single subject. *Journal of Applied Behavior Analysis*, 1979, **12**, 199-210.
- Bijou, S. W., Peterson, R. F., and Ault, M. H. 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.
- Birkimer, J. C. and Brown, J. H. A graphical judgment aid which summarizes obtained and chance reliability data and helps assess the believability of experimental effects. *Journal of Applied Behavior Analysis*, 1979, **12**, 523-533. (a)
- Birkimer, J. C. and Brown, J. H. Back to basics: Percentage agreement measures are adequate but there are easier ways. *Journal of Applied Behavior Analysis*, 1979, **12**, 535-543. (b)
- 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, 1975.
- Ciminero, A. R., Calhoun, K. S., and Adams, H. E. (Eds) *Handbook of behavioral assessment*. New York: Wiley, 1977.
- Cobb, J. A. The relationship of discrete classroom behaviors to fourth-grade academic achievement. *Journal of Educational Psychology*, 1972, **63**, 74-80.
- Cone, J. D. The relevance of reliability and validity for behavioral assessment. *Behavior Therapy*, 1977, **8**, 411-426.
- Cone, J. D. Why the "I've got a better agreement measure" literature continues to grow: A commentary on two articles by Birkimer and Brown. *Journal of Applied Behavior Analysis*, 1979, **12**, 571.
- Cone, J. D. and Hawkins, R. P. (Eds) *Behavioral assessment: New directions in clinical psychology*. New York: Brunner/Mazel, 1977.
- Cronbach, L. J., Gleser, G. O., Nanda, H., and Rajaratnam, N. *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley, 1972.
- DeProspero, A. J. and Cohen, S. H. Inconsistent visual analyses of intrasubject data. *Journal of Applied Behavior Analysis*, 1979, **12**, 573-579.
- Filipczak, J., Archer, M. B., Neale, M. S., and Winett, R. A. Issues in multivariate assessment in a large-scale behavioral program. *Journal of Applied Behavior Analysis*, 1979, **12**, 593-613.
- Foster, S. L. and Ritchey, W. L. Issues in the assessment of social competence in children. *Journal of Applied Behavior Analysis*, 1979, **12**, 625-638.
- Frederiksen, L. W., Martin, J. E., and Webster, J. S. Assessment of smoking behavior. *Journal of Applied Behavior Analysis*, 1979, **12**, 653-664.
- Goldfried, M. R. and Kent, R. N. Traditional versus behavioral personality assessment: A comparison of methodological and theoretical assumptions. *Psychological Bulletin*, 1972, **77**, 409-420.
- Goldfried, M. R. and Sprafkin, J. N. Behavioral personality assessment. In J. T. Spence, R. C. Carson, and J. W. Thibaut (Eds), *Behavioral approaches to therapy*. Morristown, N.J.: General Learning Press, 1976.
- Goodenough, F. L. *Mental testing*. New York: Rinehart, 1949.
- Greenwood, C. R., Walker, H. M., Todd, N. M., and Hops, H. Selecting a cost-effective screening measure for the assessment of preschool social withdrawal. *Journal of Applied Behavior Analysis*, 1979, **12**, 639-652.
- Hartmann, D. P. and Gardner, W. On the not so recent invention of interobserver reliability statistics: A commentary on two articles by Birkimer and Brown. *Journal of Applied Behavior Analysis*, 1979, **12**, 559-560.
- Hartshorne, H. and May, M. A. *Studies in the nature of character. Vol. I. Studies in deceit*. New York: Macmillan, 1928.
- Hawkins, R. P. Who decided that was the problem? Two stages of responsibility for applied

- behavior analysts. In W. S. Wood (Ed), *Issues in evaluating behavior modification*. Champaign, Ill.: Research Press, 1975.
- Hawkins, R. P. The functions of assessment: Implications for selection and development of devices for assessing repertoires in clinical, educational, and other settings. *Journal of Applied Behavior Analysis*, 1979, **12**, 501-516.
- Hawkins, R. P. and Fabry, B. D. Applied behavior analysis and interobserver reliability: A commentary on two articles by Birkimer and Brown. *Journal of Applied Behavior Analysis*, 1979, **12**, 545-552.
- Haynes, S. N. *Principles of behavioral assessment*. New York: Gardner Press, 1978.
- Haynes, S. N. and Wilson, C. C. *Recent advances in behavioral assessment*. San Francisco: Jossey-Bass, in press.
- Henson, D. E., Rubin, H. B., and Henson, C. Analysis of the consistency of objective measures of sexual arousal in women. *Journal of Applied Behavior Analysis*, 1979, **12**, 701-711.
- Hersen, M. and Bellack, A. S. (Eds) *Behavioral assessment: A practical handbook*. New York: Pergamon, 1976.
- Hopkins, B. L. Proposed conventions for evaluating observer reliability: A commentary on two articles by Birkimer and Brown. *Journal of Applied Behavior Analysis*, 1979, **12**, 561-564.
- Jones, R. R. Conceptual vs. analytic uses of generalizability theory in behavioral assessment. In J. D. Cone and R. P. Hawkins (Eds), *Behavioral assessment: New directions in clinical psychology*. New York: Brunner/Mazel, 1977.
- Kazdin, A. E. Assessing the clinical or applied importance of behavior change through social validation. *Behavior Modification*, 1977, **1**, 427-452.
- Kazdin, A. E. Unobtrusive measures in behavioral assessment. *Journal of Applied Behavior Analysis*, 1979, **12**, 713-724.
- Keefe, F. J., Kopel, S. A., and Gordon, S. B. *A practical guide to behavioral assessment*. New York: Springer, 1978.
- Kent, R. N., O'Leary, K. D., Dietz, A., and Diamant, C. Comparison of observational recordings *in vivo*, *via mirror*, and *via television*. *Journal of Applied Behavior Analysis*, **12**, 517-522.
- Komaki, J., Collins, R. L., and Thoene, T. J. Behavioral measurement in business, industry, and government. *Behavioral Assessment*, in press.
- Kratochwill, T. R. Just because it's reliable doesn't mean it's believable: A commentary on two articles by Birkimer and Brown. *Journal of Applied Behavior Analysis*, 1979, **12**, 553-557.
- Lang, P. J. Fear reduction and fear behavior: Problems in treating a construct. In J. M. Schlien (Ed), *Research in psychotherapy*, Vol. 3. Washington, D.C.: American Psychological Association, 1968.
- Martinez-Diaz, J. A. and Edelstein, B. A. The effects of demand characteristics on the assessment of heterosocial competence. *Journal of Applied Behavior Analysis*, 1979, **12**, 679-689.
- Mash, E. J. and Terdal, L. G. (Eds) *Behavior therapy assessment*. New York: Springer, 1976.
- Mischel, W. *Personality and assessment*. New York: Wiley, 1968.
- Nelson, R. O. and Bowles, P. E. The best of two worlds—observations with norms. *Journal of School Psychology*, 1975, **13**, 3-9.
- Nelson, R. O., Hay, L. R., and Hay, W. M. Comments on Cone's "The relevance of reliability and validity for behavioral assessment." *Behavior Therapy*, 1977, **8**, 427-430.
- Nelson, R. O. and Hayes, S. C. Some current dimensions of behavioral assessment. *Behavioral Assessment*, 1979, **1**, 1-16.
- Prinz, R. J., Foster, S., Kent, R. N., and O'Leary, K. D. Multivariate assessment of conflict in distressed and nondistressed mother-adolescent dyads. *Journal of Applied Behavior Analysis*, 1979, **12**, 691-700.
- Schnelle, J. F., Kirchner, R. E., Galbaugh, F., Domash, M., Carr, A., and Larson, L. Program evaluation research: An experimental cost-effectiveness analysis of an armed robbery intervention program. *Journal of Applied Behavior Analysis*, 1979, **12**, 615-623.
- Van Houten, R. Social validation: The evaluation of standards of competency for target behaviors. *Journal of Applied Behavior Analysis*, 1979, **12**, 581-591.
- Williams, A. M. The quantity and quality of marital interaction related to marital satisfaction: A behavioral analysis. *Journal of Applied Behavior Analysis*, 1979, **12**, 665-678.
- 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.
- Yelton, A. R. Reliability in the context of the experiment: A commentary on two articles by Birkimer and Brown. *Journal of Applied Behavior Analysis*, 1979, **12**, 565-569.

Received 1 August 1979.

(Final Acceptance 22 August 1979.)