5-1-1981

Interobserver Reliability and Convergent Validity in Observational Assessment

Jeanne M. Gilmore
University of Nebraska at Omaha

Interobserver Reliability and Convergent Validity in Observational Assessment

A Field Project

Presented to the

Department of Psychology

and the

Faculty of the Graduate College

University of Nebraska

In Partial Fulfillment

of the Requirements for the Degree

Educational Specialist

University of Nebraska at Omaha

by

Jeanne M. Gilmore

May 1981

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

Field Project Acceptance

Accepted for the faculty of the Graduate College, University of Nebraska, in partial fulfillment of the requirements for the degree Educational Specialist, University of Nebraska at Omaha.

Committee

Name	Department
Richard L. Wikoff	Psychology
Karl K. Smith	Teacher Education
Norman A. Hamm	Psychology

Richard L. Wikoff
Chairman

4-22-81
Date

Acknowledgement

I would like to thank Dan Wright, for both introducing me to this project and for his help in organizing and dealing with many of the details involved. I am also grateful to Craig Edelbrock for his assistance, as well as my co-observer, Michael Reed, who was so generous with his time, advice, and computer expertise. I would like to thank my Committee Chairman, Dr. Richard Wikoff, and the members of my Committee, Dr. Harl Jarmin and Dr. Norman Hamm, for their support and guidance. For humorous support and many hours of tedious typing, I would like to thank my friend, Lynnie Marie Swingen.

My deepest appreciation is due my best friend, Tom Gilmore, whose confidence and constant encouragement have brought my academic career to this point. I would further like to acknowledge Jim Gilmore and Christie Gilmore, who have given so generously of their time and have always provided a bright spot in the many hours invested in projects such as this.

Table of Contents

Abstract

The utility of observational assessment instruments has been well supported in the literature. However, numerous problems arise as to the establishment of adequate reliability and validity for these instruments. This study was conducted with the purpose of researching an observational assessment instrument, the Child Behavior Checklist - Direct Observation Form, which attempts to meet these psychometric qualifications. To address the topic of reliability, interobserver ratings for behavior dimensions of problem behavior and on-task behavior for 25 males, ages 6-11, were calculated through the use of an intraclass generalizability coefficient. These coefficients were found to be significant for 34 of the 36 items submitted for analysis. To address the issue of validity, the Child Behavior Checklist - Direct Observation Form was compared with two independent measures, the Child Behavior Checklist - Teacher Report Form, and the Child Behavior Checklist - Parent Report Form, to determine the degree of correlation present and to consequently determine the degree of convergent validity obtained. Convergent validity was not established for the Child Behavior Checklist - Direct Observation Form, however, the research did indicate several important findings, along with implications for future research.

Interobserver Reliability and Convergent Validity in Observational Assessment

An observational assessment instrument is a device which is typically employed in a systematic manner to record events occurring in a naturalistic environment. The context of these events may range from motor activity to cognitive and phenomenological activities, which together form the pattern of interaction to be observed.

Lynch (1977) describes an observation instrument as usually consisting of several pages, which are marked by the observer in a systematic manner to encode and categorize behavioral events. The instrument is typically designed to encourage selectivity and discrimination in recording for the observer. Following encoding of the data, a summary is usually made in quantitative form, typically frequencies or proportions.

Lynch states that observation instruments differ mainly in focus, sampling technique, types of observer judgment required, and underlying theoretical or practical rationale. Focus refers to attention directed toward a particular person or persons, physical features of the environment, or psychological states inferred from observable situations. Sampling technique refers to the method of collecting data. Three common methods are category systems, sign systems, and rating scales. The types of observer judgment required relates to the objective vs. subjective and inferential judgments the observer is required to make. Theoretical rationale refers to the conceptual orientations within which the instrument was developed.

Numerous studies have supported the utility of observational assessment (Forness & Esveldt, 1975a, 1975b; Gitler & Gordon, 1979;

Haynes & Kerns, 1979; Hunter, 1977; Keller, 1980; Kent, O'Leary, Diament & Dietz, 1974, 1979; Lipinski & Nelson, 1974; Lynch, 1977; Medley & Mitzel, 1963; Sitko, 1977). The use of observational assessment has become increasingly important in pre-intervention diagnosis and in evaluation of treatment programs (Haynes & Kerns, 1979). As Lipinski and Nelson (1974) state, the utility of observation lies in enabling the observer to objectively evaluate the effects of treatment. Keller (1980) reports that it is useful in obtaining a more direct and closer relationship between assessment and intervention, with fewer inferences than would be incurred with the sole use of standardized tests.

The rationale for the use of observational assessment lies in the assumption that true representation is obtained by consideration of in situ behavior as well as the results obtained by the use of standardized test instruments. Sitko (1977) reports an increasing interest in viewing educational assessment and programming as a systematic process considering interactional variables in the environment. The classroom is regarded as a "behavioral setting", wherein behavior is explained through analyzation of the situational context of that environment. This situational context is considered in diagnosis and recommendations.

Sitko states that the traditional standardized tests used by school psychologists are not true representations of a child's performance. This is due to "inherent problems" of standardized instruments, the variability of performance, and the fact that results of standardized tests may not relate to instructional objectives devised for the classroom. The author feels that the role of the school psychologist as a systematic observer can be critical in the decision making process. The data obtained will hopefully provide objective feedback to teachers and

students concerning the impact of their interactions, as well as providing more complete information for use in the analysis and prediction of behavior. Finally, observational data can be instrumental in developing the psychologist's skills in discriminating, generating, and evaluating teacher-pupil performance, as well as it's potential use for accountability purposes.

Forness and Esveldt (1975a) consider observational assessment to be superior to more traditional screening techniques, such as teacher rating scales, group testing, or individual testing.

Teacher rating scales have been found to be susceptible to teacher bias. Silberman (1969) has illustrated that when teacher screening is used for early referral and identification, judgments may be significantly affected by bias. Forness and Esveldt (1975b) conducted a study which presented several cases where classroom observation data either suggested problems unnoticed by teachers of clarified dimensions of an existing problem. Problems with group testing are realized by the fact that it is ineffective with younger children who have not yet acquired reading skills. Individual testing is economically prohibitive, as well as being somewhat negligent of situational or motivational variables. (Forness & Esveldt, 1975a).

Perhaps one of the most inclusive statements regarding the rationale inherent in the use of observational assessment by school psychologists is given by Keller (1980):

> The current and potential utility of direct observation in the assessment process is tremendous. Its use is consistent with the multidimensional assessment specified in P.L. 94-142. Observational assessment has the advantages of conducting assess-

> ment in those settings where the problems of concern exist, of providing a strategy that is directly related to psychoeducational planning, consultation, and intervention, and (from the perspective of others in the planning-intervention process) of being a credible information source. Its future usefulness can be enhanced by further research on psychometric characteristics and on how most effectively to implement and use observational assessment. (p. 28)

The majority of the existing observational instruments are focused on the interests of researchers in education (Lynch, 1977). As stated above, consideration should be given to the underlying theoretical or practical rationale of the instrument used. Therefore, as Lynch proposes, if most of the instruments now in use reflect the conceptual orientations of education researchers, then perhaps there is a necessity for school psychologists to develop instruments more congruent with their own objectives.

Within this context, research findings will be presented regarding reliability and validity factors considered to be essential to the development of psychometrically sound observational assessment instruments.

Reliability

Reliability is defined by Herbert and Attridge (1975) as "the consistency of the instrument as a measuring device, its tendency to obtain the same results from similar events even though these events are separated in time or location, or have different participants or setting" (p. 6). Lynch (1977) states that reliability has been defined two different ways. The most common definition is that reliability is interobserver agreement in coding the same events, and secondly, a more

psychometric definition of reliability as "the extent to which a set of observations is representative of a wider universe of pertinent observations" (p. 14). Medley and Mitzel (1973) state that "a measure is reliable to the extent that the average difference between two measures independently obtained in the same classroom is smaller than the average difference between two measurements obtained in different classrooms" (p. 250). Lipinski and Nelson (1974) define reliability as "a high level of agreement between two or more observers who are simultaneously recording the same behavioral sequence, utilizing the same recording procedure (p. 343).

As seen by the variations in reliability definitions given above, there is a lack of consensus of opinion as to the criteria which establish reliability. To further complicate matters, other studies to be reviewed below state that interobserver agreement is a questionable indicator of reliability.

A critical review of the literature on observational assessment by Johnson and Bolstad (1973) pointed out that a good deal of the data gathered by this method was subject to confounding influences which led to questionable results. Specifically, factors such as reactivity, controversy concerning the use of interobserver agreement as a measure of reliability, observer drift, observer bias, and complexity of response codes, are described and discussed. These factors are considered to be important variables in establishing the reliability of an observational instrument, and will be discussed below.

Reactivity

Reactivity refers to the concern that the presence of an observer may have an influence on the behavior of the person being observed.

This factor is believed to influence the generalizability of the data, in that the observed subjects' behavior in the natural setting may not generalize to settings where there is no awareness of being observed. (Johnson & Bolstad, 1973; Keller, 1980)

As Keller (1980) points out, reactivity is a relevant, yet inadequately researched concern in observational assessment. The problems involved have been discussed in the literature (Haynes, 1978; Johnson & Bolstad, 1973; Kazdin, 1979; Keller, 1980; Kent & Foster, 1977; Lipinski & Nelson, 1974; Reid, 1970) and it has been substantiated that there is a lack of direct data available, and that the data which is available appears contradictory.

Although the issue of reactivity is questionable at present, Johnson and Bolstad (1973) have attempted to localize sources of interference which should be considered in assessing reactivity. These sources are: conspicuousness of the observer, individual differences of subjects, personal attributes of the observer, and the rationale for observation. In considering conspicuousness of the observer, it is felt that the observer should be as neutral a stimulus as possible. There should be little interaction between the observer and the subject. The authors feel that longer habituation periods are necessary when the observer is particularly distracting in order to achieve more stable data. Individual differences of subjects include personality variables (e.g., guardedness), age, and sex. Personal attributes of the observer which may affect reactivity include age, sex, socio-economic status, and professional status. The issue of rationale for observation relates to informing the subjects of the rationale involved, in order to alleviate any anxiety or guardedness which could promote reactivity.

Interobserver Agreement

Mascaro (1969) defines interobserver agreement as "the degree of agreement among independent observers." Interobserver agreement is the most frequently used index of reliability in observational assessment. As Nelson and Hayes (1979) state, "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."

Despite this rationale, a number of authors (Abikoff, Gittleman - Klein, & Klein, 1977; Flanders, 1966; Herbert & Attridge, 1975; Johnson & Bolstad, 1973; Kazdin, 1977; Lynch, 1977; McGow and Wardrop, 1972; Repp, Dietz, Boles, Dietz & Repp, 1976; Romanczyk, Kent, Diament, O'Leary, 1973) have disputed the use of interobserver agreement as an indicator of reliability. The major sources of controversy regarding this issue stem from: a discrimination in definition between observer agreement and observer accuracy; reactivity of the observer; and the technique or method chosen to calculate interobserver agreement.

As Kazin (1977) states, reliability is often considered in terms of interobserver agreement in that it is assumed that if a high level of agreement is obtained, then it is a fairly accurate reflection of the subjects' behavior. According to Bijow, Peterson and Ault (1968) and Johnson and Bolstad (1973), agreement and accuracy are separate indices. Accuracy is obtained by comparing the observer's data with a predetermined standard, whereas observer agreement is obtained by determining the extent to which two observers agree on scoring. In the latter situation, there is no criteria on which to base the assumption that one observer's data should serve as the standard. Johnson and Bolstad further state

that "it is quite possible to have perfect observer agreement or accuracy on a given behavioral score with absolutely no reliability or consistency of measurement in the traditional sense." Furthermore, although there is a relationship between agreement and accuracy, they may function independently (Kazdin, 1977). For example, an observer can observe accurately, but have low interobserver agreement. Conversely, two observers can obtain high agreement, but have little accuracy.

The issue of observer reactivity refers to a problem quite similar to reactivity of the observee, described above. In the case of the observer, however, it has been found (Reid, 1970; Romanczyk et al., 1973) that reliability coefficients are higher when the observers are aware that they are being assessed than when they are unaware. Additional studies (Kent, Kanowitz, O'Leary, & Cheiken, 1977; Kent, O'Leary, Diament, & Dietz, 1974; Taplin & Reid, 1973) have demonstrated that awareness by observers of having their reliability assessed influenced their data. In general, studies have shown that observers show substantially higher agreement and accuracy when they believed that their observations and data were being evaluated. These findings present implications for interpreting reliability data based on interobserver agreement. Possible solutions to the problem of observer awareness are offered by Kazdin (1977).

Although several studies (Bellack, Kliebard, Hyman & Smith, 1966; Smith & Meux 1962; Flanders, 1967) have dealt with reliability only in terms of interobserver agreement, there is a good deal of controversy in the literature regarding the appropriateness of this index of reliability (Herbert & Attridge, 1975; McGaw & Wardrop, 1972; Medley & Mitzel, 1963; Repp, et al. 1976). In addition, these studies have suggested that reliability measured in these terms can be further distorted by the

methods or techniques used to calculate observer agreement. Medley and Mitzel suggest that observer agreement is not an adequate measure of reliability, and use the term "reliability coefficient" to refer to the correlation of scores obtained by observers.

An additional problem in using observer agreement as an index of reliability is found in the use of the method chosen for calculating this agreement (Repp, Dietz, Boles, Dietz, Repp, 1976). These authors state that this factor can have an effect on scores reported by varying the means across responses on the same data from 64% to 94%. Numerous other authors have contributed to this controversy (Birkimer & Brown, 1979a, 1979b; Cone, 1979; Hartmann & Gardner, 1979; Hawkins & Fabry, 1979; Hopkins, 1979; Kratochivill, 1979; Yelton, 1979).

Herbert and Attridge (1975) also agree that interobserver agreement should not be considered a measure of reliability. However, they do not discount its' usefulness, stating:

> Coefficients or percentages of observer agreement should be described and discussed for system developers and users as indicators of the clarity of the structure, focus, and procedures of the system, and as measures of observer bias or the ambiguity of observed events. (p. 14)

Despite the controversy over the use of interobserver agreement as a measure of reliability, it appears to remain in prominent usage among authors of observational assessment systems. Perhaps as Herbert and Attridge suggest, manuals and reports pertinent to an observational assessment system must contain information regarding the reliability measures selected. With this data, the employer of the system can then decide on the reliability of the scores obtained and their potential

utility.

Observer Drift

Most commonly, observers receive training regarding proper administration of observational instruments. One purpose of this training is to ensure a uniform level of accuracy, facilitated by determining behavioral definitions which are expected to remain consistent throughout observations.

Several studies have investigated the existence of this consistency, and found that observers who compare results can inadvertently produce modifications in their data which may produce high interobserver agreement without accompanying accuracy. (Johnson & Bolstad, 1973; Kazdin, 1977; Keller, 1980; Kent, Kanowitz, O'Leary, Cheiken, 1977; Kent, O'Leary, Diament, Dietz, 1974; O'Leary & Kent, 1973; Reid, 1970; Taplin & Reid, 1973).

This tendency of observers to modify recordings of behavior after comparing agreements or discussing definitions with one another is labeled "observer drift." (O'Leary & Kent, 1973). As these authors state, drift may not be reflected in interobserver agreement. That is, with continued interaction, two observers may develop similar variations of the original response definitions and achieve high levels of interobserver agreement with a declining level of accuracy.

Several studies (Browning & Stover, 1971; Johnson & Bolstad, 1973); Keller, 1980) have investigated possible preventative steps which could be taken to attenuate or eliminate observer drift. Among these steps are continuous training, discussion of the coding system, accuracy feedback for observers, the use of multiple observers, a random change of pairs of observers over time, and a continuous check of interobserver

agreement and accuracy.

Observer Bias

Observer bias was described by Rosenthal (1966) "as the extent to which experimenter effect or error is assymmetrically distributed about the 'correct' or 'true' value". Observer bias occurs when the data obtained by observation is influenced by factors other than the occurrence of the behaviors being observed. The major factors reported in the literature which appear to contribute to observer bias are: knowledge of expected results, knowledge that reliability is being assessed (Lipinski & Nelson, 1974), and the extent of code specificity (Johnson & Bolstad, 1973: Skindrud, 1972). An additional source of bias may exist in the theoretical views held by observers which may affect their perception and recognition of behaviors (Kendell, 1968).

In considering the effects of observer expectations, several studies (Azrin, 1961; Kass & O'Leary, 1970; Kendall, 1968, Kent, 1972; Kent, et al., 1974; Rapp, 1966; Rosenthal, 1963; Rosenthal & Fode, 1963, Rosenthal & Jacobson, 1966; Scott, Burton & Yarrow, 1967.) have presented evidence suggesting that the observer's knowledge of the purpose of assessment are preconceptions about the subjects to be observed, could facilitate an unintended source of variance in the results obtained. A number of these studies have been subject to question due to methodology and failure to replicate similar results. However, the majority of studies considering the effects of observer expectations suggest that a more reliable account of observational data may be obtained if observers did not have access to information which may lead to confounding of results.

An additional source of bias may stem from the observer's knowledge of the fact that the data obtained is being assessed for reliability.

Reid (1970) demonstrated that when observers were aware that reliability was being assessed, they obtained median reliability of 75%; when informed that reliability was not being assessed, the median reliability decreased to 51%. Romancyzk, et al. (1973) obtained similar results. These studies suggest that when reliability is not being assessed, reliabilities would be lower, in addition to the possibility of additional bias caused by underestimating the frequency of target behaviors. It is suggested that reliability assessment perhaps induces the observer to be more cognizant of the situation (Lipinski & Nelson, 1974).

According to Johnson and Bolstad, the specificity of the coding involved in observational assessment may be an important consideration in the matter of observer bias. In this context, it appears that a specific vs. a global coding system is favorable. The authors feel that ambiguity of the codes may produce a greater possibility for bias. In contrast, a system with well-defined behavioral codes should restrict interpretive bias. A similar position is stated by Skindrud (1972) regarding specific and behaviorally defined coding systems.

In summary, the literature supports the need to analyze the reliability of ratings between independent observers. The factors described above have also been documented as major problems in interpreting these analyses. In response to these factors, a generalizability theory approach was developed (Cronbach, Ikeda, & Avner, 1964; Cronbach, Rajaratnam, & Gleser, 1963; Gleser, Cronbach, & Rajaratnam, 1965). In essence, this theory provides a comprehensive method for assessing sources of variability in observed scores by use of an analysis of variance model. This theory directs itself to differentiating the sources of error described above, and provides a method for estimating the contribution of

various sources to the variance in observed scores.

Based upon this theory, Berk (1979) presents an intraclass correlation-generalizability theory approach which more precisely deals with the complexity of determining interobserver reliability. This intraclass correlation-generalizability theory purports to be a more comprehensive and flexible model for dealing with problem factors in interobserver reliability than any previously reported index of interobserver reliability.

Validity

As Lipinski and Nelson (1974) point out, even if an observational assessment instrument were to overcome the methodological problems involved in establishing reliability (as discussed above), this does not demonstrate validity. Stated simply, high reliability does not imply high validity.

The literature available offers little information on the matter of establishing validity on observational instruments (Borich & Malitz, 1975; Haynes & Kerns, 1979; Herbert & Attridge, 1975; Johnson & Bolstad, 1973; Kazdin, 1979; Keller, 1980; Lynch, 1977), although a few (Borich & Malitz, 1975; Haynes & Kerns, 1979; Patterson, Reid, Jones, & Conger, 1975; Wahler, House, & Stambaugh, 1976) have addressed the issues of content, criterion-related, or construct validity of observational systems.

Kazdin (1979) feels that a partial explanation for the lack of validation data may lie in the fact that most observational measures are usually improvised for a single study. The measures are not standardized and their psychometric aspects are not considered to the extent of a more generalizable instrument. The result of this is that difficulties are incurred in interpreting the data obtained from these measures.

When applied to observational assessment instruments, "validity

refers to the degree to which the measures obtained by an instrument actually describe what they purport to describe" (Herbert & Attridge, 1975). The types of validity which receive primary attention in the literature on observational assessment are: construct; content; criterion-related; and to a lesser extent, discriminative validity.

According to Herbert and Attridge, construct validity "refers to the degree to which the theoretical claims and supports of the instrument are substantiated both logically and empirically." In this instance, the researcher is empirically or analytically testing hypotheses which are generated by the construct(s). The authors feel that when dealing with construct validity, it is essential to clarify the theoretical basis of the construct to be investigated, e.g., psychoanalytic theory, behavioristic theory, etc. This is assumed to be important in that the potential user should be able to assess the adequacy and accuracy of the instrument as it is utilized in observational assessment.

Content validity is considered in determining to what extent the instrument items actually measure the behaviors under observation. Within this context, Herbert and Attridge state that these instrument items must be as low in the degree of observer inference required as the complexity of behavior under observation will allow. Observer inference refers to the amount of observer judgment required to score a particular item. A consideration must be made of the amount of inference made between the actual occurrence of the behavior and the observer's recording of the behavior, as observational systems vary in the amount of inference required. That is, high inference instruments may require measurement of a series of events as opposed to a specific unit of behavior; may measure more global or ambiguous characteristics of behavior; may have

vaguely defined criteria for use; or may force the observer to classify events from memory (Herbert & Attridge, 1975).

Arguments exist on both sides regarding the appropriateness of high inference items vs. low inference items. A few studies (Campbell, 1961; Medley & Mitzel, 1963; Soar, 1971) conclude that high inference items offer potential for distortion. However, Gellert (1955) and Boyd and DeVault (1966) feel that low inference items share the same fault due to their selectivity and objectivity. In support of the latter position, Rosenshine and Furst (1973) state that "...some observational systems which distort reality appear to be more predictive of student achievement than the systems which more closely represent the actual events" (p. 136).

The matter of degree of inference required, as it relates to items constituting content validity, is debatable. Perhaps, as Herbert and Attridge suggest, the decision of degree of inference allowable should rest upon the degree of complexity of the behavior under observation.

Borich and Bauman (1972) define discriminative validity as "the correlation between different measures measuring the same trait exceeding (a) the correlations obtained between that trait and any other trait not having method in common and (b) the correlations between different traits which happen to employ the same method" (p. 1029). To establish discriminative validity, clear observational differences must be demonstrated between groups (e.g., problem vs. non-problem children).

With the exception of a few studies (Foster & Ritchey, 1979; Oden & Asher, 1977; Wodarski, 1977), little information was available in the literature regarding discriminative validity. In the studies referred to, discriminative validity was not established. These studies did

indicate two possible considerations to be made in attempting to establish discriminative validity: observational codes chosen should be sensitive enough to capture differences between populations; and the behaviors chosen for discrimination should be appropriate for that purpose. Clearly, the need for further research is indicated by the sparsity of current data available on the establishment of discriminative validity in observational assessment.

The final method of validation to be considered is criterion-related validity. This type of validation is particularly pertinent when development of an observation instrument is focused toward evaluation or selection purposes, and is defined as the relationship between scores obtained on the observation instrument and scores obtained on some other variable used as a criterion.

Criterion-related validity includes both convergent (or concurrent) and predictive validity. Although little criterion-related validity research with observational instruments has been undertaken (Keller, 1980), several articles have addressed issues of convergent and predictive validity (Herbert & Attridge, 1975; Johnson & Bolstad, 1973; Lynch, 1977) and one study investigated convergent validation of several observational systems (Borich & Malitz, 1975).

Convergent validity is defined by Borich and Malitz an a "confirmation of traits (or variables or categories) by independent measuring methods that requires significant correlation between two methods (or systems) measuring the same trait" (p. 426). Herbert and Attridge suggest that a comparison of the results from the observational instrument with results from a currently validated test or observation tool would be desirable. However, the practicality of this method of validation is impeded by the

lack of such instruments currently available. Campbell and Fiske (1959) further support the utility of convergent validity:

> Validation is typically convergent; a confirmation by independent measurement procedures. Independence of methods is a common denominator among the major types of validity (excepting content validity) insofar as they are able to be distinguished from reliability. Reliability is the agreement between two efforts to measure the same trait through maximally similar methods. Validity is represented in the agreement between two attempts to measure the same trait through maximally different methods. (p. 86)

Johnson and Bolstad (1973) stress the need for convergent validation, particularly in instruments encompassing a broad band of behavioral dimension (e.g., deviant vs. nondeviant behaviors).

Predictive validity is established when the results of scores obtained on a behavioral dimension correlate with a criterion established by a dissimilar measurement instrument (Johnson & Bolstad, 1973). There is little data available which affords evidence for the convergent or predictive validity of observational data. A few exceptions are found, however (D'Heurle, Mellinger, & Haggard, 1959; Hughes, 1968; Meyers, Attwell, & Orpet, 1968; Patterson & Reid, 1971), which demonstrate that observational data have a potential to provide evidence of predictive validity.

It appears evident that validation of observational assessment instruments has been difficult even when the effort had been undertaken to establish an index of validity. There appears to be a need for standardized observational instruments which meet adequate psychometric

standards. Furthermore, although it would be desirable to employ an observational instrument which could demonstrate an adequate level of several of these types of validity, the particular objectives of the school psychologist (evaluation and/or selection) emphasize the need for convergent validity.

Statement of the Problem

Numerous articles have supported the rationale underlying observational assessment and its utility in providing a more holistic approach to the evaluation of behavior. Despite this consensus of opinion regarding rationale and utility, a summary of the literature reveals the existence of considerable controversy over the issue of reliability. This controversy extends from establishing a common definition for reliability to agreement as to what criteria actually do establish reliability in observational assessment.

Although several articles have discussed validity, primarily in the traditional-theoretic sense, very few have approached this problem empirically.

At this point, it may be useful to first consider the matter of theoretical orientation discussed at the onset of this paper. In interpreting the results of research relevant to reliability and validity in observational assessment, it should be noted that the majority of the literature available is behavioristic in theory. Within this orientation, the basis for establishing reliability and validity stems from objective vs. subjective data, specific vs. global data, overt vs. covert behavior definitions, etc. Unfortunately, the literature available does not appear to offer theoretical and/or empirical data or opinions from other theoretical orientations. Therefore, it would seem prudent to consider

the data presented in reference to reliability and validity as having been derived from a primarily behavioristic base. This situation also presents implications for the need for research directed at developing observational systems which are not solely behavioristic in theory, but rather may provide essential information through more subjectively based data.

An additional implication to be derived from the literature reviewed is that the current lack of validity data is an issue which must be addressed if the potential utility of observational assessment is to be realized. More importantly, it is of little relevance if an instrument is reliable if it is not measuring what it purports to measure. Therefore, it appears that there is an imminent need for empirically-based research on the establishment of validity for observational assessment instruments.

Within this context, a sizeable proportion of the literature referred to criterion-related validity (specifically convergent validity) as a favorable method of validation for observation instruments focused toward purposes of evaluation or selection. As this focus would appear most pertinent for the school psychologist, the use of convergent validity as the method of validation seems appropriate.

As stated at the onset of this paper, there is a need for school psychologists to develop or utilize observational instruments which are focused on concerns pertinent to their particular assessment objectives. In order to facilitate this need, an observational instrument should be developed which is consistent with the assessment objectives commonly encountered by the school psychologist, i.e., evaluation and/or selection. In conjunction with this premise, it is felt that convergent validity should be selected as the method of validation to be employed with this

instrument.

For the purposes of establishing convergent validity, the observational assessment instrument selected for this study will be compared with two additional independent measures to specify the degree of correlation present. In addition, the results obtained through the use of this instrument will be analyzed to determine the degree of interobserver reliability obtained. It is anticipated that the procedure of comparing the observational instrument with standardized instruments will improve its psychometric qualities, as well as better qualifying the degree of convergent validity to be determined. It is also anticipated that the method selected to determine interobserver agreement will be useful in the detection of sources of error variance, as well as providing an index of reliability. In the event that these procedures produce significant results, substantial promise may be present for an observational instrument which will be of sound practical usage for the school psychologist.

Method

Subjects

Twenty-five male students between the ages of 6 and 11 were observed. The racial composition consisted of 23 white males, one black, and one of Asian descent. The mean age was 9.

Solicitations for subject referrals were sent to special services personnel (i.e., special education teachers and counselors) in a small middle to upper-middle-class suburban school district. The criteria for referral specified that the child be male, between the ages of 6 and 11, and has been considered to have exhibited at least moderate behavior problems. This age criterion was specified, as the instruments selected

for comparison in validation have based their normative data on this age group.

Instruments

To address the issue of employing an observational instrument which is based on more subjectively based data than the predominantly behaviorist-theory oriented instruments currently available, the Child Behavior Checklist-Direct Observation Form (Achenbach & Edelbrock, Note 1) will be utilized in the research to be presented here. Because of its recent development, this form has not previously been subject to research. Consequently, the research to be presented here will constitute the only empirical data available to date on this observational instrument.

For the purposes of determining the convergent validity of the Child Behavior Checklist-Direct Observation Form (CBCL-DOF), this instrument will be compared to two additional independent measures to determine the degree of correlation present. These additional measures are the Child Behavior Checklist-Parent Report Form (Achenbach & Edelbrock, Note 2) and the Child Behavior Checklist-Teacher Report Form (Achenbach & Edelbrock, Note 3). The Child Behavior Checklist-Parent Report Form (CBCL-PRF) was utilized to develop a child behavior profile designed to descriptively classify children for clinical and research purposes. The edition of this profile developed for boys 6-11 was based on the factor analysis of Checklests completed on 450 children referred for mental health services. The results of the profile analysis indicated significant discriminative validity ($p < .001$) on all behavior problem and social competence scales. Eight-day test-retest correlations averaged .89, and interparent correlations averaged .74 (Achenbach, 1978).

The Child Behavior Checklist-Teacher Report Form (CBCL-TRF) is in

the process of standardization at this time. Edelbrock (1979) presents a detailed rationale for the use of and progress made in classification of children through the use of parent and teacher forms, as well as their usefulness in school psychology.

Procedure

The procedure was intended to be as unobtrusive as possible and the subjects were not informed of the observations (parent consent forms were obtained for all subjects)

Two observers simultaneously observed individual subjects for six 10-minute periods at random intervals throughout a five-week period. In accordance with the standard procedure of the CBCL-DOF, each observer observed the child for 10 minutes. During each minute of this 10-minute span, the observers independently transcribed a narrative description of the behavior observed, noting the occurrence, duration, and intensity of any behavioral problems. At the end of each minute, the child's on-task behavior was noted for a five-second interval. If the child's behavior was considered to be on-task, a box designated for that specific minute was crossed out. If the child's behavior was not considered to be on-task, the box was not crossed out. At the end of the 10-minute period, the behavior checklist was completed. This checklist consists of 97 problem behavior items which may be rated in intensity on a scale from 0-3 index of intensity.

To obtain a representative sample of behavior, this procedure was followed for six 10-minute periods, with each 10-minute period occurring at random times on six different days. During the approximate five-week period in which the observation data was being collected, CBCL Parent and Teacher forms were distributed for completion and return.

Results

Interobserver Reliability

To determine the degree of interobserver reliability obtained, generalizability coefficients were derived from a one-way analysis of variance with blocking for child. Blocking was done to remove the variance associated with differences between children and to decrease the error variance.

The degree of interobserver reliability was calculated on two separate dimensions; agreement on occurrence and intensity of problem behaviors and agreement of on-task behaviors. Within these two dimensions, generalizability coefficients were obtained for two factors, utilizing formulae presented by Berk (1979). The first factor is the generalizability of a single observation (P_1^2), which estimates the average agreement between one observer and another. The second factor (P_2^2) is the generalizability of the average of $\underline{k}$ observations ($\underline{k}$ = number of observers), which estimates the average agreement between the single random sample of observers used in a study and a theoretical set of other random samples drawn from the same universe.

In addition, reliability coefficients were obtained for observer ratings through (a) analyzing ratings of each of six sessions individually and averaging the coefficients, and (b) analyzing total behavior problem scores (the sum of all problems observed in six sessions) across all sessions. The results of the analysis are summarized in Table 1 and indicate significant agreement between observers in all situations. In addition, Table 1 depicts two important findings; (a) generalizability coefficients calculated on observer ratings averaged from individual

Table 1

Reliability Coefficients of Observer Ratings

Agreement Dimensions	Average of Six Individual Sessions	Sum Across Six Sessions
Problem Behavior	$P_1^2 = .78$*	$P_1^2 = .86$*
	$P_2^2 = .87$*	$P_2^2 = .92$*
On-Task Behavior	$P_1^2 = .59$*	$P_1^2 = .71$*
	$P_2^2 = .74$*	$P_2^2 = .83$*

* $p < .01$

sessions were lower than those calculated on ratings summed across the total six sessions, and (b) generalizability coefficients for agreement of on-task behavior were lower than those for behavior problems.

Table II lists the reliability coefficients derived for each of the 36 items rated by the observers. The analysis included only 36 of the possible 97 behavior problems, as only these 36 items were rated with a frequency of 15% or more by the two observers. The results indicate that 34 of the items were significant at the .01 level.

Convergent Validity

To determine the degree of convergent validity of the Child Behavior Checklist - Direct Observation Form, the per-item ratings obtained by the two observers on this form were averaged and compared to per-item ratings submitted by teachers (Child Behavior Checklist - Teacher Report Form) and parents (Child Behavior Checklist - Parent Report Form).

Following the averaging of the observer ratings for each item, a frequency breakdown on responses was conducted. The frequency breakdown revealed 40 items from the CBCL-DOF wherein there was sufficient range in scores to calculate correlations. Of these 40 items, there were 36 items from the Teacher Report Form and 28 items from the Parent Report Form which corresponded. These corresponding items were analyzed to determine the degree of correlation present between Observer/Teacher ratings and Observer/Parent ratings.

The correlation coefficients obtained from comparison of Observer/Teacher ratings are displayed in Table III, together with the corresponding Checklist item on which ratings were obtained. The results indicate that only 11 of the 36 items (approximately 31%) attained a significant

Table 11

Interobserver Reliability on 36 Problem Behavior Items

Child Behavior Checklist - Direct Observation Form

CBCL-DOF Item No.	Correlation Coefficient
1	.90*
2	.91*
3	.65*
5	.91*
7	.81*
9	.86*
10	.86*
11	.83*
12	.96*
13	.60*
15	.73*
17	.88*
20	.61*
21	.73*
33	.84*
38	.88*
42	.52*
44	.75*
46	.64*
52	.87*
53	.80*
54	.82*
55	.62*
56	.65*
57	.81*
63	.77*
64	.88*
69	.89*
75	.84*
77	.65*
79	.71*
83	.86*
84	.05
88	.85*
94	.93*
97	.29

* $p < .01$

Table III

Correlation Coefficients for Corresponding

Observer/Teacher Ratings

Corresponding Item	Correlation Coefficient
Acts too young for age	.28
Makes odd noises	.20
Argues	.23
Bragging, boasting	.22
Doesn't concentrate or pay attention for long	.52**
Doesn't sit still, restless or hyperactive	.18
Clings to adults or too dependent	.26
Confused or seems to be in a fog	.46*
Cries	.42*
Fidgets	.31
Daydreams or gets lost in thoughts	.36
Demands or tries to get attention of staff	.14
Disobedient	.47*
Disturbs other children	.36
Doesn't seem to feel guilty after misbehaving	.21
Impulsive or acts without thinking	.08
Bites fingernails	.00
Nervous, highstrung or tense	.05
Nervous movements or twitching	.14
Picks nose, skin, or other parts of body	.20
Apathetic or unmotivated	.55**
Disrupts group	.08
Shows off or clowns	.57**
Shy or timid behavior	.12
Explosive behavior	.05
Demands must be met immediately, easily frustrated	.14
Inattentive, easily distracted	.56**
Stares blankly	.55**
Stubborn, sullen or irritable	.25
Sudden changes in mood or feelings	.39
Talks too much	.27
Underactive, slow moving, or lacks energy	.73**
Unusually loud	.23
Whining	.10
Clumsy, poor motor control	.45*
Doesn't get along with peers	.42*

* $p < .05$
** $p < .01$

level of correlation.

The correlation coefficients obtained from comparison of Observer/Parent ratings are displayed in Table IV, together with the corresponding Checklist items on which the ratings were obtained. These results indicate that only 4 of the 28 items (approximately 14%) attained a significant level of correlation.

Discussion

The results of this study relating to interobserver reliability indicate a positive direction for the technique of observational assessment. The generalizability coefficients obtained from observer ratings on items from the Child Behavior Checklist - Direct Observation Form were significant and demonstrate one of the few attempts to-date to develop a psychometrically-sound observational assessment technique.

The use of the generalizability coefficient as described above presents probable solutions to many of the problems which have complicated the determination of reliability in observational assessment tools. As Berk (1979) states, the major advantages of generalizability theory are that it: yields unbiased estimates of interobserver reliability; provides an analysis of reliability-related factors by partitioning variance components; provides reliability estimates of single observations and sets of observations; permits the choice to include or exclude observer bias as part of the error variance term; and can yield one or more reliability coefficients using the data from one analysis. In general terms, a generalizability analysis permits the isolation of many of the problem variables which affect the measurement of behavior by observational methods.

Table IV

Correlation Coefficients for Corresponding Observer/Parent Ratings

Corresponding Item	Correlation Coefficient
Acts too young for age	.33
Argues	.06
Bragging, boasting	.28
Doesn't concentrate or pay attention for long	.06
Doesn't sit still, restless, or hyperactive	.16
Clings to adults or too dependent	.18
Confused or seems to be in a fog	.21
Cries	.38
Daydreams or gets lost in thoughts	.29
Demands or tries to get attention of staff	.40*
Disobedient	.52**
Disturbs other children	.24
Doesn't seem to feel guilty after misbehaving	.04
Impulsive or acts without thinking	.19
Bites fingernails	.21
Nervous, highstrung or tense	.19
Nervous movements or twitching	.02
Picks nose, skin, or other parts of body	.22
Shows off or clowns	.24
Shy or timid behavior	.00
Stares blankly	.12
Stubborn, sullen or irritable	.40*
Sudden changes in mood or feelings	.05
Talks too much	.24
Underactive, slow moving, or lacks energy	.15
Unusually loud	.18
Whining	.05
Clumsy, poor motor control	.65**

* $p < .05$
** $p < .01$

The definition of reliability was stated by Herbert and Attridge (1975) as "the consistency of the instrument as a measuring device, its tendency to obtain the same results from similiar events even though these events are separated in time or location, or have different participants or settings" (p. 6). The purpose of establishing generalizability coefficients is to indicate the generalizability of a sample of observations to a specified universe of observations. Therefore, if a significant generalizability coefficient were obtained through the use of an observational instrument, then it would appear to be reliable, as the definition of reliability and the purposes of establishing a generalizability coefficient are indeed the same.

Although the generalizability coefficients indicate that the CBCL-DOF is a reliable instrument, close analysis of the results point out three pertinent considerations:

1. Generalizability coefficients calculated on observer ratings averaged from individual sessions were lower than those calculated on ratings summed across the total six sessions. This finding has important implications for the development of a reliable observational assessment instrument, in that it indicates that the additional amount of information obtainable from more than one observation appears to increase reliability, and therefore, multiple observations should be completed.

2. Generalizability coefficients for agreement of on-task behavior were lower than those for problem behavior. This unexpected finding of obtaining lower generalizability coefficients for the more objectively defined on-task behaviors, as opposed to the more subjectively defined problem behaviors, presents a possibility that the observer's stopwatches

were not accurately synchronized. In utilizing the CBCL-DOF and other observation instruments which employ fixed intervals within which to assess on-task behavior, it appears that attention must be given to the precision of the time measurements.

3. As indicated in Table 11, two of the 36 behavior rating items failed to reach significance in terms of interobserver reliability. Specifically, these two items were to be rated on intensity and occurrence of "impatience" and "other problems." A probable explanation for the lack of reliability on these two items could be found in the amount of observer judgment required in these relatively high inference items. It is supposed that these two items may attempt to measure behaviors which are too global or ambiguous, and that perhaps they should be revised to specify more clearly defined criteria for rating purposes.

In summary, significant generalizability coefficients were obtained within the dimensions of Problem Behavior and On-Task Behavior. In accordance with generalizability theory in general, these results suggest that the CBCL-DOF has applicability or generalizability to settings other than the one employed for this study. The establishment of this generalizability is a relevant factor for the school psychologist, who may find it necessary to conduct observational assessments in a variety of settings.

The results of the study relating to convergent validity indicated that the majority of corresponding items compared for correlation from Observer/Teacher forms and Observer/Parent forms did not attain significance. In view of the relatively few items which correlated at a significant level, the CBCL-DOF is not considered to have attained a sufficient

level of convergent validity with the two independent measures with which it was compared.

As Lipinski and Nelson (1974) have stated, even if an observational instrument were to overcome the methodological problems involved in establishing reliability, this does not demonstrate validity. That is, high reliability does not imply high validity. This statement appears quite relevant to the validity results of the CBCL-DOF, as the indicators of reliability are quite high, while validity indicators are low.

The results of the correlations between Observer/Teacher ratings indicated that approximately 31% of the items reached significance, whereas correlations between Observer/Parent ratings indicated that approximately 14% of the items reached significance. This finding of greater correlation between Observer/Teacher ratings is perhaps best explained by the fact that observers and teachers measure the child's behavior in the same setting, whereas there would seem to be allowance for a greater variety of behaviors in a home setting. Perhaps this finding indicates that observers should not limit their observations to the school setting, but rather employ multiple observations in multiple settings, including the home.

In discussing the items which succeeded or failed to demonstrate significant correlation between observers and teacher/parent ratings, the matter of degree of inference required per item should be considered. As discussed above, studies conducted on the appropriateness of high inference vs. low inference items present arguments on each side; that is, that high inference items offer potential for distortion, whereas low inference items share the same fault due to their selectivity and objectivity.

Upon examination of the significant vs. non-significant items listed in Tables III and IV, it appears that several high inference items reached significance, while several low inference items did not (e.g., "stubborn, sullen, or irritable" = $p < .05$; "bites fingernails" = not reaching significance). It also appears that the reverse is true in some instances (e.g., "clumsy, poor motor control = $p < .01$; shy or timid behavior = not reaching significance). Therefore, these mixed results suggest that perhaps the matter of low vs. high inference items is not as relevant as determining the level of content validity present in each item. The correlation coefficients derived from comparison of ratings on each item should indicate the usefulness of each item in measuring observed behavior; those items which demonstrate high correlation coefficients being retained for inclusion in the instrument. It appears that future research should be directed at establishing content validity for the items included in the Child Behavior Checklist series, and consequently re-submitting the CBCL-DOF for analysis in order to determine the level of convergent validity present.

Reference Notes

1. Achenbach, T.M. and Edelbrock, C. Child Behavior Checklist-Direct Observation Form. 1981. (Available from Thomas M. Achenbach, University of Vermont, Burlington, Vt. 05405; Craig Edelbrock, Boys Town Center for the Study of Youth Development, Boys Town, NE 68010.)

2. Achenbach, T.M. and Edelbrock, C. Child Behavior Checklist-Parent Report Form. (Available from addresses listed above.)

3. Achenbach, T.M. and Edelbrock, C. Child Behavior Checklist-Teacher Report Form. (Available from addresses listed above.)

References

Abikoff, H., Gittelman-Klein, R., & Klein, D. Validation of a classroom observation code for hyperactive children. Journal of Consulting and Clinical Psychology, 1977, 45, No. 5, 772-783.

Achenbach, T.M. The Child Behavior Profile: I, Boys aged 6-11. Journal of Consulting and Clinical Psychology, 1978, 46, 478-488.

Azrin, H.H., Holz, W., Ulrich, R., and Goldiamond, I. The control of the content of conversation through reinforcement. Journal of the Experimental Analysis of Behavior, 1951, 4, 25-30.

Bellack, A.A., Kiliebard, H.M., Hyman, R.T., & Smith F.L., Jr. The language of the classroom. New York: Teachers College Press, Columbia University, 1966.

Berk, R.A. Generalizability of behavioral observations: a clarification of interobserver agreement and interobserver reliability. American Journal of Mental Deficiency, 1979, 83, No. 5. 460-472.

Bijou, S.W., Peterson, R.E., 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 judgmental 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 Behavior Analysis, 1979, 12 535-543. (b)

Borich, G.D., & Bauman, P.M. Convergent and discriminant validation of the French and Guilford-Zimmerman spatial orientation and spatia visualization factors. Educational and Psychological Measurement, 1972, 32, 1029-1033.

Borich, G.D., & Malitz, D. Convergent and discriminative validation of three classroom observation systems: a proposed model. Journal of Educational Psychology, 1975,67, 426-431.

Boyd, R., & DeVault, M.V. Observation and recording of behavior. Review of Educational Research, 1966, 36, 529-551.

Browning, R.M. and Stover, D.O. Behavior modification in child treatment: An experimental and clinical approach. Chicago: Aldine-Atherton, Inc., 1971.

Campbell, D.T. The mutual methodological relevance of anthropology and psychology. In F.L.K. Hsu (Ed.), Psychological Anthropology. Homewood, Ill. Dorsey Publishing Co., 1961.

Campbell, D.T. and Fiske, D. Convergent and discriminative validation by the multi-trait, multi-method matrix. Psychological Bulletin, 1959, 56, 81-105.

Cone, J.D. The relevance of reliability and validity for behavioral assessment. Behavior Therapy, 1977, 8, 411-426.

Cronbach, I.J., Ikeda, M., & Avner, R.A. Intraclass correlation as an approximation to the coefficient of generalizability. Psychological Reports, 1964, 15, 727-736.

Cronbach, I.J., Rajaratnam, N., & Gleser, G.C. Theory of generalizability: A liberalization of reliability theory. The British Journal of Statistical Psychology, 1963, 16, 137-163.

D'Heurle, A., Mellinger, J.C., & Haggard, E.A. Personality, intellectual, and achievement patterns in gifted children. Psychological Monographs: General and Applied, 1959, 73, whole No. 483.

Edelbrock, C.S. Empirical classification of children's behavior disorders: progress based on parent and teacher ratings. School Psychology Digest, 1979, 8, 355-369.

Flanders, N.A. Interaction Analysis in the classroom: A manual for observers. (Rev. Ed.) Ann Arbor: University of Michigan, 1966.

Flanders, N.A. The problems of observer training and reliability. In Amidon, E. and Hough, J. (Eds.) Interaction Analysis: theory, research, and application. Massachusetts: Addison-Wesley, 1967.

Forness, S.R., & Esveldt, K.C. Classroom observation of Children with Learning and Behavior Problems. Journal of Learning Disabilities, 1975, 8, 49-52.

Forness, S.R., & Esveldt, K.C. Prediction of high-risk kindergarten children through classroom observation. Journal of Special Education, 1975, 9, 375-387.

Foster, S.L. & Ritchey, W.L. Issues in the assessment of social competence in children. Journal of Applied Behavior Analysis, 1979, 12, 625-638.

Gellert, E. Systematic observation: A method in child study. Harvard Educational Review, 1955, 25, 179-195.

Gitler, D. & Gordon, R. Observing and recording young handicapped children's behavior: A comparison among observational methodologies Exceptional Children, 1979, 46, 134-135.

Gleser, G.C., Cronbach, L.J., & Rajaratnam, N. Generalizability of scores influenced by multiple sources of variance. Psychometrika, 1965, 30, 395-418.

Hartmann, D.P. and Gardner, W. On the not so recent invention of inter-observer reliability statistics: A commentary on two articles by Birkimer and Brown. Journal of Applied Behavior Analysis, 1979, 12, 559-560.

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, 1978.

Haynes, S.N. & Kerns, R.D. Validation of a behavioral observation system. Journal of Consulting and Clinical Psychology, 1979, 47, 397-400.

Herbert, J., & Attridge, C. A guide for developers and users of observation systems and manuals. American Educational Research Journal, 1975, 12, 1-20.

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.

Hughes, L.D. A study of the relationship of coping strength to self-control, school achievement, and general anxiety level in sixth grade pupils. Dissertation Abstracts. 1968, 28, (A) (10), 4001.

Hunter, C.P. Classroom observation instruments and teacher inservice training by school psychologists School Psychology Monograph, 1977, Fall, 45-88.

Johnson, S.M. & Bolstad, O.D. Methodological issues in naturalistic observation: some problems and solutions for field research. In L.A. Hamerlynchk, Behavior Change: Methodology, concepts, and practice. Champaign, Ill., 1973.

Kass, R.E. and O'Leary, K.D. The effects of observer bias in field-experimental settings. Paper presented at a symposium entitled "Behavior Analysis in Education," University of Kansas, Lawrence, April, 1970.

Kazdin, A.E., Artifact, bias, and complexity of assessment: The ABC's of reliability. Journal of Applied Behavior Analysis, 1977, 10, 141-150.

Kazdin, A.E. Unobtrusive measures in behavioral assessment. Journal of Applied Behavior Analysis, 1979, 12, 713-724.

Keller, H.R., Issues in the use of observational assessment School Psychology Review, 1980, 9, 21-30.

Kendell, R.E., An important source of bias affecting ratings made by psychiatrists. Journal of Psychiatry, 1968,6, 135-141.

Kent, R. The human observer: An imperfect cumulative recorder. Paper presented at the Fourth Banff Conference on Behavior Modification, Banff, Alberta, Canada, March, 1972.

Kent, R.N., & Foster, S.L. Direct observational procedures: Methodological issues in naturalistic settings. In A.R. Ciminero, K.S. Calhown, & H.E. Adams (Eds.), Handbook of behavioral assessment. New York: Wiley, 1977.

Kent, R.N., Kanowitz, J., O'Leary, K.D., & Cheiken, M. Observer reliability as a function of circumstances of assessment. Journal of Applied Behavior Analysis, 1977, 9, 109-113.

Kent, R.N., O'Leary, K.D., Diament, C., & Dietz, A. Expectation biases in observational evaluation of therapeutic change. Journal of Consulting & Clinical Psychology, 1974, 42, 774-780.

Kent, R.N., O'Leary, K.D., Dietz, A., & Diament, C. Comparison of observational recordings in vivo, via mirror and via television. Journal of Applied Behavior Analysis, 1979, 12, 517-522.

Kratchowill, 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.

Lipinski, D. & Nelson, R. Problems in the use of naturalistic observation as a means of behavior assessment. Behavior Therapy, 1974, 5, 341-351.

Lynch, W.W. Guidelines to the use of classroom observation instruments by school psychologists. School Psychology Monograph, 1977, Spring, 1-22.

Mascaro, G.F. Some methodological aspects of systematic categorization of behavior. Perceptual and Motor Skills, 1969, 28, 779-784.

McGaw, B., Wardrop, J.L., & Bundy, M.A. Classroom observation schemes: where are the errors? American Educational Research Journal, 1972, 9, 13-27.

Medley, D.M., & Mitzel, H.E. Measuring classroom behavior by systematic observation. In N.L. Gage, (Ed.), Handbook of research on teaching, Chicago: Rand McNally, 1963.

Meyers, C.E., Attwell, A.A., and Orpet, R.E. Prediction of fifth grade achievement from kindergarten test and rating data. Educational and Psychological Measurement, 1968, 28, 457-463.

Nelson, R.O., & Hayes, S.C. The nature of behavioral assessment: a commentary. Journal of Applied Behavior Analysis, 1979, 12, 491-500.

Oden, S. & Asher, S.R. Coaching children in social skills for friendship making. Child Development, 1977, 48, 495-506.

O'Leary, K.D., & Kent, R.N. Behavior modification for social action: Research tactics and problems. In L.A. Hamerlynck, L.C. Handy, & J.E. Mash (Eds.), Behavior change: Methodology, concepts, and practice. Champaign, Ill.: Research Press, 1973.

Patterson, G.R., & Reid, J.B. Family intervention in the homes of aggressive boys: A replication. Paper presented at the American Psychological Association Convention, Washington, D.C., 1971.

Patterson, G.R., Reid, J.B., Jones, R.R., & Conger, R.E. A social learning approach to family intervention. Eugene, Ore.: Castalia, 1975.

Rapp, D.W. Detection of observer bias in the written record. Cited in R. Rosenthal, Experimenter effects in behavioral research. New York: Appleton-Century-Crofts, 1966.

Reid, J.B. Reliability assessment of observational data; a possible methodological problem. Child Development, 1970, 41, 1143-1150.

Repp, A.C., Dietz, E.D., Boles, S.M., Deitz, S.M., & Repp, C.F. Differences among common methods for calculating interobserver agreement. Journal of Applied Behavior Analysis, 1976, 9, 109-113.

Romanczyk, R.D., Kent, R.N.S. Diament, C., & O'Leary, K.D. Measuring the reliability of observational data: a reactive process. Journal of Applied Behavior Analysis, 1973, 6, 175-184.

Rosenshine, B., & Furst, N. The use of direct observation to study teaching. In R.M.W. Travers (Ed.), Second Handbook of research on teaching. Chicago: Rand McNally, 1973.

Rosenthal, R. On the social psychology of the psychological experiment: The experimenter's hypothesis as unintended determinant of experimental results. American Scientist, 1963, 51, 268-283.

Rosenthal, R. Experimenter effects in behavioral research. New York: Appleton-Century-Crofts, 1966.

Rosenthal, R. and Fode, K.L. The effect of experimenter bias on the performance of the albino rat. Behavior Science, 1963, 8, 183-189.

Rosenthal, R. and Jacobsen, L. Teacher's expectancies: Determinants of pupils IQ gains. Psychological Reports, 1966, 19, 115-118.

Scott, P., Burton, R.V., and Yarrow, M. Social reinforcement under natural conditions. Child Development, 1967, 38, 53-63.

Silberman, M. Behavioral expression of teachers' attitudes toward elementary school students. Journal of Educational Psychology, 1969, 60, 402-407.

Sitko, M.C. Utilizing systematic observation for decision making in school psychology. School Psychology Monograph, 1977, 3, 23-44.

Skindrud, K.D. Field evaluation of observer bias under covert and event monitoring of observer reliability: Two preliminary studies. Oregon Research Institute Monograph, 1972, 12, No. 7.

Smith, B.O., & Meux, M. A study of the logic of teaching Cooperative Research Project No. 258, U.S. Dept. of Health, Education, and Welfare: Office of Education, Univ. of Illinois, 1962.

Soar, R. Follow through classroom process measurement. Gainesville, Florida: Institute for Development of Human Resources, College of Education, University of Florida, 1971.

Taplin, P.S. & Reid, J.B. Effects of instructional set & experimenter influence on observer reliability. Child Development, 1973, 44, 547-554.

Wahler, R.G., House, A.E., & Stambaugh, E.E. Ecological assessment of child problem behavior: A clinical package for home, school, and institutional settings. New York: Pergamon Press, 1976.

Werry, J.S. & Quay, H.C. Observing the classroom behavior of elementary school children. Exceptional Children, 1969, 35, 461-472.

Wodarski, J.S. A comparison of behavioral consistency of anti-social and pre-social children in different contexts. Journal of Behavior Therapy and Experimental Psychiatry, 1977, 8, 275-280.

Yelton, A.R. Reliability in the context of the experiment: A reply to Birkimer and Brown. Journal of Applied Behavior Analysis, 1979, 12, 565-569.