THE RELIABILITY OF THE BEHAVIORAL INTERVIEW

by

William M. Hay

A Dissertation Submitted to
the Faculty of the Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Greensboro
1977

Approved by

Rosemary O. Nelson

Dissertation Adviser

APPROVAL PAGE

This dissertation has been approved by the following committee of the Faculty of the Graduate School at the University of North Carolina at Greensboro.

Dissertation Adviser Rosemary O. Nelson

Committee Members

November 9, 1977

Date of Acceptance by Committee

HAY, WILLIAM M. The Reliability of the Behavioral Interview. (1977) Directed by: Dr. Rosemery O. Nelson. Pp. 111.

The primary objective of the present study was to investigate the reliability of the behavioral interview. This objective was operationalized in three ways. First, a generalizability (G) study was performed to establish the generalizability across interviewers with respect to the number of areas identified as problems per client. Second, the agreement among the interviewers as to those specific areas which were identified as problems for a particular client and as to the specific problem-items within an identified area was determined. Third, the accuracy of interview data was measured by establishing the agreement between each human interviewer and a criterion (computer) interview.

Four interviewers conducted comprehensive behavioral interviews with the same four clients. These interviews were audio-taped and transcribed in order to provide a verbatim account of the content. Each interviewer dictated a summary following each interview. In addition, each client completed a standardized computer interview. Transcriptions of interviews, dictations, and computer interview printouts were coded independently by two raters for areas and items identified as problems and areas and items questioned.

The results of G study analyses of coded interview and dictation data indicated that interviewers did not identify significantly different numbers of problem areas. While these results suggested that it was possible to generalize across

interviewers with respect to the overall number of areas identified as problems for a client, inter-interviewer agreement on specific problem areas and items indicated low levels of reliability. Analyses of the agreement between each interviewer and the criterion (computer) interview also revealed a low level of agreement for areas and items identified as problems.

In summary, the results of the present study indicated low inter-interviewer agreement between the human interviewers themselves and the human interviewers and the criterion interview. Three factors, interviewer input and output differences and the consistency of client responses were examined to determine their influence on interview content. Although client responses across interviews were consistent, interviewer input and output differences seemed implicated as contributing sources to attenuated reliability. Standardization of interview procedures was proposed as one remedy for the reliability problem found in the present study. Computerization of standardized interviewing procedures was presented as a tool for increasing the potency of this remedy.

ACKNOWLEDGEMENT

I would like to thank the members of my dissertation committee for their encouragement and differential feedback throughout this research project. A special thanks goes to my committee chairperson, Dr. Rosemery O. Nelson, for her substantive and organizational inputs into this project. I also greatly appreciated the assistance of Dr. Hugh V. Angle both for supporting this project financially and for his role in the conceptualization of the research. In addition I wish to thank all the research interviewers and the staff of the Medication Evaluation Project for the exceptionally high standards of professional behavior they exhibited throughout the duration of this project.

I dedicate this dissertation to my parents who have been a consistent source of support. They are the two people who had confidence in me even when I was not sure I had confidence in myself.

TABLE OF CONTENTS

TABLE OF CONTENTS (Cont.)

LIST OF TABLES

LIST OF TABLES (Cont.)

CHAPTER I

INTRODUCTION

The Reliability of the Behavioral Interview

The interview has played a central role in virtually every form of psychotherapy. As a result of its importance to the therapeutic process, a voluminous amount of research on interviewing has been generated. Most of these research studies have examined the clinical interview in its traditional roles as a vehicle for psychotherapy and as a diagnostic instrument.

The interview has been conceptualized as an interpersonal interaction process in which the behavior of the participants is reciprocally determined. The predominant research strategy has been to reduce this interaction to an asymmetrical contingency: The behavior or characteristics of one of the participants is manipulated and the effect on the behavior of the other participant is observed (Heller, 1971). Indices of the effect of these manipulations have been in terms of such measures as productivity (overall verbal output) and fluency (e.g., Pope & Siegman, 1972), self-disclosure (reviewed by Cozby, 1973), client responsiveness (Pope, Nudler, Vandoroff & McGhee, 1974), and level of anxiety experienced by the interviewee (Dibner, 1958). In addition, a substantial amount of research has examined the reliability of psychiatric diagnoses and other clinical judgments which are based on

interview data. The low reliability found in a number of these studies (e.g., Ash, 1949; Eysenck, 1952; Ward, Beck, Mendelson, Mock & Erbough, 1962; Zigler & Phillips, 1961) has alerted clinicians to the potential problems inherent in the human interview process.

The results of these previous studies may have limited relevance to behavioral interviewing procedures. Behaviorally oriented approaches to assessment and treatment question the basic assumptions regarding personality structure that have guided the development and interpretation of previous assessment instruments (Goldfried & Sprafkin, 1974). Changes in the focus of assessment have already resulted in changes in the purpose, structure, development and evaluation of behavioral assessment procedures. As a result, while the process of interviewing in behavioral assessment (face-to-face interaction) has remained the same, the objective of the interview has changed in line with the changing focus of assessment.

A major objective of this introduction will be to describe the evolving role of the interview in the process of behavioral assessment and to examine its empirical status as an assessment instrument. Initially, the theoretical and practical differences between diagnostic and behavioral models of assessment will be discussed and a comprehensive behavioral-assessment model will be proposed. Presentation of this model will establish the importance and changed purpose of the interview in behavioral assessment. Two subsequent sections will review the empirical status of the behavioral interview and

outline potential sources of variance and error that may adversely affect the reliability of interview data. Finally, the appropriateness of utilizing traditional psychometric procedures to measure the reliability of the behavioral interview will be addressed, and the overall research design and hypotheses of the present study will be outlined.

Diagnostic and Behavioral Models of Assessment

Two approaches to the assessment of a patient or client are common in psychotherapy: The diagnostic and the behavioral. The two differ with respect to their objectives.

The outcome of the assessment process in the diagnositc model is the assignment of a person to a particular location along a continuum of personality dimensions or to a specific nosological category, such as the Kraepelinean system typically employed in psychiatric diagnosis. In theory such a diagnostic disposition dictates which treatment procedures are most appropriate for a particular patient. This approach represents an extrapolation of the disease model that guides the assessment of medical problems, in that similar symptoms are presumed to be the result of similar etiologies and thus responsive to similar treatments (Frank, 1975; Kanfer & Saslow, 1969). The procedures employed to reach a diagnosis are aimed at identifying the signs or behaviors that match the defining characteristics of the diagnostic lexicon.

The objective of the behavioral model of assessment is to identify the problem behaviors that are currently causing difficulties for an individual client. The problem behaviors

subsequently become the focus of treatment. The assessment procedures within this model are geared to the collection of verbal reports or representative samples of the client's functioning in a broad spectrum of life areas (Goldfried & Kent, 1972).

There are two major conceptual differences between the diagnostic and the behavioral approaches. The diagnostic approach is nomothetic: Patients assigned to the same category are presumed to have characteristics in common and consequently to be responsive to similar treatment regimens. Within this model the relationship between assessment information and treatment is indirect (McLean & Miles, 1974) in that nosological dispositions function as mediators in the selection of therapeutic procedures. In contrast the behavioral model represents an idiographic assessment approach in which the unique problems of the individual are of interest. Individualized assessment information feeds directly into the development of treatment plans that are, of necessity, uniquely tailored to amelioration of the individual client's problems (Kanfer & Saslow, 1969; Stuart, 1970).

Although the diagnostic model seems logical, the evidence has not supported its reliability or utility. In order for a nosological system to be useful, two criteria must be met. First, independent assessments of the same patient by various clinicians must result in the assignment of that patient to the same diagnostic category. Second, the assignment of a diagnostic label must have relevance for subsequent treatment decisions (Peterson, 1968).

Studies concerning the reliability of diagnosis have indicated poor interjudge reliability of assignment of patients to specific categories (Ash, 1949; Schmidt & Fonda, 1956; Ward, Beck, Mendelson, Mock, & Erbaugh, 1962). This attenuated reliability has been attributed to the lack of mutually exclusive categories (Bannister, Salmon, & Lieberman, 1964), resulting in an overlap of symptoms indicative of different diagnoses (Nathan, 1967; Wittenborn, Holzberg, & Simon, 1953).

Similarly, it has been demonstrated that assignment to a diagnostic category does not mandate the selection of particular treatment procedures (Bannister et al., 1964). Many investigators have noted that the proposed relationship between outcome and treatment does not exist in practice (Meehl, 1960; Peterson, 1968; Frank, 1975; Hayes-Roth, Longabaugh, & Ryback, 1973). Specific treatment techniques often appear to be more a function of therapist training than an outgrowth of assessment information (Goldfried & Pomeranz, 1968). Further, the availability of numerous therapeutic strategies for the treatment of people with a particular diagnosis suggests that factors other than diagnostic labels direct treatment decisions.

Even if a system could be designed that ensured reliable relationships among assessment information, diagnosis, and treatment, the utility of the diagnostic model would still be in question. The emphasis on identifying common characteristics rather than the unique life problems of the client directs therapeutic attention away from important aspects of the patient's life (Kanfer & Saslow, 1969), and may result in clinicians

attempting to remedy "inner symptoms" to the neglect of the specific life areas in which the client is experiencing problems (McPartland & Richart, 1966).

These concerns about the reliability and utility of diagnostic procedures have resulted in an increasing shift toward a behavioral approach to assessment (Bandura, 1969; Kanfer & Saslow, 1969).

A Comprehensive Behavioral-Assessment Model

A number of multifaceted behavioral assessment strategies have been proposed (e.g., Cautela & Upper, 1976; Goldfried & Pomeranz, 1968; Kanfer & Saslow, 1969; Peterson, 1968; Stuart, 1970; Thomas & Walters, 1973; Wolpe, 1969). In general, these strategies provide conceptual frameworks for the process of behavioral assessment and in some cases outline general guidelines for data collection (Kanfer & Saslow, 1969). The common objective of these strategies is the identification and functional analysis of problem behaviors so that a parsimonious and effective treatment program can be developed. The strategies differ in the scope of information (i.e., the comprehensiveness of the data base) that is deemed necessary to implement this objective.

Of the behavioral assessment strategies currently in use, the model proposed by Kanfer and his associates (Kanfer & Saslow, 1969; Kanfer & Grimm, 1977) represents the most comprehensive and detailed assessment format. The Kanfer and Saslow strategy specifies seven components or areas that should

be investigated during the behavioral assessment process: 1) initial analysis of the problem situation with the emphasis on identifying behavioral excesses and deficits; 2) clarification of the environmental parameters, antecedents, and consequences, currently maintaining problem behavior; 3) motivational analysis - reinforcement survey; 4) developmental analysis of biological, sociological, and behavioral changes affecting current behavior; 5) analysis of self-control repertoire; 6) analysis of social relationships; 7) analysis of the social-cultural-physical environment-- normative comparisons. This assessment strategy provides a broad data base for subsequent clinical decisions. Systematic and reliable procedures for the processing of this information toward the development of specific treatment targets, however, are not specified. Consequently, the clinician remains an "artist" in selecting behaviors for intervention (Dickson, 1975; Linehan, 1977).

The comprehensive behavioral-assessment model proposed in this paper incorporates the major components of previous assessment formats. The model differs from previous strategies in its emphasis on the identification of functional relationships within problem behavior as a procedure for integrating and processing assessment information. This comprehensive behavioral-assessment model views the individual as a system of behavior. The system is an exceedingly complex one in that it is probabilistic and involves a huge number of interactions. The individual's total system of behavior is further

differentiated into a variety of life areas, such as marriage, employment, and child management. Each of these life areas is viewed as a subsystem of the total system and is itself a complex system of behavior. The composite of life areas for each individual represents a different total system of behavior.

An individual's system of behavior is not static. Changes in the environment impose new demands on the system each day. In most instances the system adapts to the changing behavioral requirements of the environment. Occasionally, however, the system is confronted with an environmental demand to which it cannot successfully adjust.

When faced with a problem, the system may attempt to adjust, and the person may exhibit behaviors that provide immediate relief, but that have wide-range detrimental consequences for the remainder of the system. For example, to cope with stresses arising in the employment subsystem, a person may increase the rate of alcohol consumption. Although this behavior may provide temporary relief from the problem, its continuation may have negative ramifications for the system as a whole (e.g., marital difficulties, health problems, financial strains). When the person or a significant member of the environment detects that the system is not functioning effectively, the person may be referred for psychological evaluation. The task of assessment becomes the specification of those behaviors that are having detrimental effects on the system, in order that appropriate therapeutic action may be undertaken to improve the functioning of the system as a whole.

In making a comprehensive behavioral assessment the first task is the identification of the problem behaviors that are currently causing difficulties for the system. The problem behaviors may include not only overt motor behaviors but also the person's physiological and verbal responses. Verbal expressions of negative feelings or thoughts are not ignored, but are viewed as verbal behavior problems. Problem behaviors are defined with respect to their consequences and are described in terms of their frequency of occurrence in particular environmental settings.

Within a behavioral model the frequency of an inappropriate response rather than its nature is the primary determinant of whether a behavior is considered a problem (Ferster, Culbertson, & Boren, 1975). Almost any behavior is appropriate under certain environmental conditions. Almost everyone, for example, has consumed an alcoholic beverage or two at a cocktail party. What defines a person's drinking behavior as problematic is that he drinks to excess and/or in situations where its socially appropriate frequency is zero. Similarly, almost everyone has stated at one time or another that he feels unhappy or sad. When the frequency of this behavior increases markedly, however, the person is labeled depressed. Thus, problem behaviors within each life area are classified as behavioral excesses or deficits on the basis of their frequency of occurrence in particular situations (Kanfer & Saslow, 1969).

Once the problem behaviors have been identified, they are translated into therapeutic objectives for the system. These objectives are stated in terms of the projected frequency with which the behavior must occur in a given situation for the system to function more effectively. Since the system's problems and objectives are stated in terms of their present and desired frequencies, respectively, the current status of the system with respect to each objective can be quantitatively assessed.

Subsequent to the specification of the system's objectives, alternative treatment intervention strategies are considered for reduction of the difference between the present and the desired performance of the system. In some instances in which only a limited number of isolated and specific problem behaviors have been identified, the targets for modification are obvious. But people usually have many problems, extending over a wide range of life areas. Simultaneous treatment of each problem is neither practical nor feasible. The problem behaviors must be conceptualized systematically to facilitate the formulation of the most rational and economical hierarchy of treatment interventions to counter them.

In accordance with the system's approach to assessment, each problem behavior is considered with respect to its role in the total system of behaviors defining the individual client (Ryback, 1975; Ryback & Gardner, 1973). Before selecting target behaviors for treatment, one must determine functional relationships among the problem behaviors. It is important to

know whether a particular problem behavior is the result of a more fundamental behavioral excess or deficit.

Once the interrelationships among the problem behaviors have been specified, the positive and negative consequences of modifying each particular problem behavior on the client's other problem behaviors can be predicted. In addition, one must also consider the ramifications that a change in a particular type of behavior may have for the currently nonproblematic components of the system. The result of this interrelationship of problem behaviors to the system as a whole is the formulation of the most appropriate treatment intervention.

A hypothetical case presented by Goldfried and Pomeranz (1968) illustrates the importance of this stage of assessment for the selection of target behaviors for treatment intervention.

> Consider the case of a 50-year-old man who comes to therapy because he has difficulty in leaving his house. The situation has reached the point where merely contemplating getting out of bed results in such anxiety that most of his time is spent in a prone position and he therefore must be constantly looked after by his wife. Further questioning reveals that his most salient fear is having a heart attack which he states is the reason for remaining at home and in bed. Upon carrying the assessment further - this time evaluating the nature of his current life situation - it is found that this man has recently been promoted in his job to a position where he now has the responsibility for supervising a large staff. Prior to his promotion, he led a fairly normal life and his fears of having a heart attack were non-existent.
>
> Other assessment procedures reveal that the client has always had the tendency to become anxious in unfamiliar situations, and he is the type of person who would prefer to have other people look after him and care for him. Additionally, questioning his wife reveals that she does not find the current

> situation entirely noxious; rather, she feels important and needed now that she has to care for her husband, and she lavishes much attention and affection on him in his incapacitated state.

Examination of this case reveals the interrelationship between the client's presenting problem, inability to leave his house, and the more fundamental problem of anxiety concerning increased employment responsibilities. This anxiety may, in turn, be the result of a deficient repertoire of administrative behaviors. As Goldfried and Pomeranz (1968) note, the selection of the client's presenting problem for treatment would not have been the most appropriate system objective and most probably would have resulted in treatment failure. Further, they note that the problem could have been solved by suggesting that the man not accept the promotion, but this would have unfavorable ramifications for other components (i.e., financial situation, loss of status) of the system. The most appropriate intervention strategy would include an increase in the client's administrative repertoire with a concomitant decrease in anxiety concerning his administrative performance. Such a therapeutic plan would also have to consider the effect of changing the client's presenting problem behavior on his relationship with his wife.

In summary, a comprehensive behavioral assessment model has been described in which the patient or client is conceptualized as an exceedingly complex system of behavior. This model requires the collection of an extensive amount of information about each client's functioning in a wide range

of life areas. The chief means of gathering information within the model is the interview. During the initial phase of assessment, the interview is employed as a broad band procedure (comprehensive coverage) in the identification of the full range of problem behavior. As the scope of inquiry narrows, the interview complements other more specialized assessment instruments in the functional analysis of specific problem areas.

The objective of the interview within the comprehensive behavioral assessment model, therefore, differs from its objective within the diagnostic assessment model. The objective of the interview in the diagnostic approach to assessment is to gather the information necessary for the assignment of a psychiatric diagnosis. Once a diagnosis is arrived at assessment is complete. In contrast the objective of the interview in behavioral assessment is to identify the full range of a client's problem behavior and to aid in the specification of the environmental parameters effecting the occurrence of each problem behavior.

The Empirical Status of the Behavioral Interview

In spite of the importance and widespread use of the interview in behavioral assessment, there is an absence of research on the reliability and validity of interview data and on potential variables affecting the interview process. The lack of research on the behavioral interview is surprising

since behaviorists have repeatedly criticized other schools of psychology for their lack of experimental rigor. With respect to the issue of the interview, behaviorists may be equally at fault. The interview has not been included among those assessment procedures, for example, naturalistic observations (Johnson & Bolstad, 1973) and scales such as the Fear Survey Schedule (Dickson, 1975), which have undergone comparatively extensive experimental analyses concerning their psychometric properties. Two factors that may have contributed to this lack of research on the behavioral interview are the denigration of self-report data by behavioral practitioners and the development of narrow-band assessment devices.

Behaviorists have tended to denigrate self-report data in favor of the direct observation of motoric behavior (Mahoney, 1975). In some instances, however, where direct observation of motor behavior is prohibitive (e.g., the assessment of sexual behavior or covert behaviors), there may have to be an almost total dependence on verbal report as a method of assessing changes in other response modes. In other instances the client's verbal behavior may be the primary focus of the treatment program, such as in the treatment of delusional behavior.

A number of authors have commented on the disproportionate development and empirical investigation of behavioral treatments in comparison with advances in behavioral assessment procedures (Dickson, 1975; Linehan, 1977; Mash & Terdal,

1974). The proliferation of treatment techniques aimed at the modification of specific problem behaviors has fostered a narrow-band approach to behavioral assessment with a de-emphasis on more wide-band, comprehensive assessment instruments such as the interview. Narrow-band assessment instruments have been designed to supply information for problem areas where standard treatment interventions are available: sexual dysfunction (LoPiccolo & Steger, 1974); assertive behavior (Gambrill & Richey, 1975); and marital conflicts (Stuart & Stuart, 1975. Although these narrow-band assessment tools yield specific and useful information, they assume a priori that the clinician has already identified the important problem areas for assessment and modification.

The importance of employing wide-band procedures (i.e., more comprehensive coverage) such as the interview has been considered by Cronbach and Gleser (1965) in their discussion of decision theory. Essentially, the interview is utilized during the first stages in any sequential decision-making process of assessment "to identify questions that need to be considered and facts that need to be obtained" (Cronbach & Gleser, 1965, p. 146). As a first stage in behavioral assessment, the interview is primarily utilized as a wide-band information-gathering procedure. Information collected during interview sessions is crucial for establishing the client's problem behavior areas and as a source for formulating clinical hypotheses to be tested with more narrow-band assessment instruments.

The extensive range of information that can be gathered during an interview may improve every subsequent assessment decision that is to be made. In contrast, narrow-band procedures give specific information with respect to one decision, but provide little or no guidance for the remaining decisions (Cronbach & Gleser, 1965). The negative consequences that can accrue from improper decisions in the initial stages of assessment include poor treatment selection and planning, with resultant treatment failures (Lazarus, 1973).

The interview is therefore valuable as a wide-band assessment instrument. The interview should be considered as a viable measurement device subject to the same methodological problems (e.g., reliability) as other measurement procedures (Kahn & Cannel, 1957). The conceptualization of the behavioral interview as a measurement device clarifies the task at hand. As scientists our initial task is to establish the reliability (i.e., consistency or precision) of any measurement procedure we employ (Sidman, 1960).

Psychometric Evaluation of Behavioral Assessment Techniques

Traditional Concepts of Reliability.

In the traditional conceptualization of personality and behavior, which forms the basis for classical reliability theory, an individual's score on a test is assumed to be determined by his "true" score and "error" ($x=t+e$). The "true" score is considered to be the average score that an individual would obtain if an observation were repeated an infinite number of times. Error is considered

to be randomly distributed with a mean of 0: error is uncorrelated with the true score. Variance in observed scores therefore is a result of true score variance and error variance. Reliability coefficients reflect that proportion of the variance in observed scores that is nonerror variance or "true" score variance. This is expressed in a reliability coefficient or ratio of true score variance to the total observed scores variance (Cronbach, 1960; Wiggins, 1973). A perfectly reliable measurement instrument should yield correlations approaching unity even if data were collected in different situations. An individual's absolute score on a test may vary across situations or time but his rank order position on each dimension should theoretically remain invariant (Ekehammer, 1974).

As noted earlier, the behavioral conceptualization of personality considers behavior to be determined by the interaction of the individual with his environment. The individual's behavior is viewed as flexible, rather than stable and enduring, for flexibility is required to meet changing environmental demands. Consistencies in behavior are viewed to be a function of the similarity of consequences for a behavior across situations. Consequently, variability in behavior is not necessarily considered to reflect error.

Given these basic conceptual differences, it is evident that questions concerning the reliability of data should be different for traditional and behavioral assessment procedures. Although procedurally traditional measures of reliability (test-

retest, split-halves, and equivalent forms) can be adapted for the evaluation of behavior assessment procedures, it is questionable whether these measures provide the most appropriate evaluations of behavioral data (Cone, 1977; Nelson, Hay, & Hay, 1977; Wiggins, 1973).

The test-retest method of assessing reliability yields a coefficient of stability determined by correlating the scores from two observations of the same group of subjects separated by some specified period of time. In traditional personality theory, if the obtained correlation is high, the test is considered to be reliable; theoretically, individuals should maintain their rank order on each dimension over time. If a low correlation is obtained, the test is considered unreliable; the scores are affected by random error. Within behavioral personality theory, variablity in behavior across time does not reflect error. In fact, variability in behavior may provide information concerning external stimuli that influence the behavior of the individual.

The split-halves method of assessing reliability yields a coefficient of internal consistency that is computed by dividing test items in half and correlating scores from the two halves of a single test administration. In traditional personality theory, all test items are assumed to measure the same attribute of the individual. Low correlations signify that the test is unreliable because the items are not equivalent. In behavioral personality theory, test items are designed to sample the individual's responses to a wide variety of stimulus

situations. The behaviorist would not necessarily find the inter-item consistency of a test desirable because it is the differential responsiveness to test items that is useful in identifying the stimuli affecting the individual's behavior.

In the equivalent-forms method of assessing reliability a coefficient of equivalence is calculated by correlating the scores of a group of individuals on two equivalent measures of an attribute at the same time. Theoretically, it is assumed that an individual's attributes should affect both tests in the same way since the items on both tests are designed to measure the same attribute. A high correlation between two tests signifies that both are measuring the same attribute with a high degree of precision. To the behaviorist, the same correlation suggests that two tests may be presenting functionally similar stimuli to the individual.

To summarize, in traditional personality theory, the precision, stability and consistency of measurement are synonymous, referring to the reliability of the assessment procedure. The variance in observed scores across time, items, and situations is attributed to the error in measurement. Behavioral personality theory, on the other hand, does not predict inter-situational consistency of behavior. Thus, the stability, consistency and precision of measurement are separate issues for investigation.

Theory of Generalizability. In the early 1960's, Cronbach and his associates (Cronbach, Rajaratnam & Gleser, 1963) proposed an alternative conceptualization of the traditional concept of

reliability. Cronbach and his associates suggested that the question of reliability becomes the question of how well the data obtained by a measurement technique can be generalized to some broader class of observations or "universe." According to this view, the reliability of an assessment technique is always determined with respect to the universe to which the researcher wishes to generalize. For a given assessment method, therefore, there is no single reliability estimate but numerous reliabilities each relative to a specific assessment parameter. The researcher can specify the universe which is of interest and conduct a generalizability study (G study) to assess the procedure's reliability with respect to the specified universe. A G study is an analysis of variance from which estimates of the variance attributable to each "facet" (dimension) and their interactions are determined. Facets may include settings, observers, instruments, occasions or attributes (Wiggins, 1973).

Few behavioral researchers have conducted G studies to assess the reliability of their assessment procedures. An exception is Jones, Reid, and Patterson (1975) who conducted a G study to determine the reliability of their Behavior Coding System used in the observation of family interactions. The facets in their analysis were coders (regular coder/calibrating coder), occasions (days 1 and 2), subjects, and the interactions of these variables. Two coders recorded the behavior of 30 boys (two samples: 13 "problem" boys; 17 "normal" boys) on two different days of a 10-day observation period. An analysis of the components of variance for each of the samples of boys

revealed that almost all of the total variance was accounted for by subjects and the subjects x occasions interactions. Coders did not account for any significant segment of the total variance. Consequently, these investigators concluded that for a particular subject generalization across coders was justified.

The methodology (G study) recommended for determining the generalizability of data is not applicable in all cases. In the G study, scores obtained for a session or test administration (dependent variables) by different observers or in different situations (independent variables) are analyzed using an analysis of variance format to determine the percent of variability accounted for by the facet of interest to the researcher. To qualify for analysis using an analysis of variance model, certain assumptions concerning the data must be met. First, the data are assumed to be normally distributed with a normal distribution of errors: each population has a normal distribution of scores. Second, the model assumes the population error variances to be equal. Third, independence among the error component is required. Independence of observations will rarely be met when comparing measurements of the same individuals across trials or occasions. Although the first two assumptions can be violated to some extent, violation of the assumption of statistical independence of observations can cause serious errors in interpretation of results. Furthermore, the G study requires the selection of random samples from the universe of interest: in many research setting it may be very difficult for researchers to obtain samples selected at random.

In addition, the data must be in interval-scale measurements for statements about magnitudes or amounts to be meaningful. In many investigations, the researcher is interested in comparing nominally scaled measurements or exact point-to-point correspondences in recording. The analysis of variance model does not answer the question: "...how often do two observers watching one subject, and equipped with the same definitions of behavior see it occurring or not occurring at the same standard times" (Baer, 1977, p.118).

The Concept of Reliability in Behavioral Assessment. Under the rubric of behavioral assessment, the term reliability has been used to describe several different concepts of psychometric evaluation. Most commonly, the term reliability has meant the percent of interobserver agreement between observers who independently record the behavior of the same subject. Johnson and Bolstad (1973) point out, however, that two observers could conceivably obtain comparable behavioral codings of the same behavior interaction, resulting in a high level of interobserver agreement, yet both have recorded inaccurately. For example, a child may exhibit three instances of aggressive behavior during an observation interval and both observers may inaccurately record only two occurrences of the target behavior. To determine the accuracy of observational data, therefore, observer codings have been compared to a predetermined criterion coding or standard for the same behavioral interaction. Johnson and Bolstad (1973) refer to this criterion comparison as the establishment of observer accuracy. Thus, within a behavioral

framework, the establishment of reliability has been redefined as the assessment of interobserver agreement and interobserver accuracy.

Reliability of the Behavioral Interview

For the behavioral interview, no studies have examined the inter-interviewer agreement or accuracy of interview data (Morganste , 1976; Mash & Terdal, 1974). A major objective of the present study was to investigate the reliability of the behavioral interview. Reliability was defined in three ways: 1. Generalizability; 2. Inter-interviewer Agreement; 3. Interviewer accuracy.

Generalizability

A generalizability (G) study was conducted across subjects in order to obtain an estimate of the variance attributable to the facet of behavioral interviewers. The universe of interest was behavioral interviewers. Specifically, four behavioral interviewers conducted four comprehensive interviews, one interview with each of four clients. These interviews were tape-recorded and transcribed verbatim to provide a complete record of each interviewer-client interaction. Since verbatim transcriptions are not typically available, however, each interviewer was also asked to dictate the information obtained immediately following each interview. These dictations provided a more representative data base for assessing the reliability of the behavioral interview. Both the dictations and transcriptions were analyzed with respect to the number of problem areas identified for each client by each interviewer.

Inter-Interviewer Agreement

To determine the extent of agreement between the interviewers on the specific problem areas identified for the same clients, inter-interviewer agreement scores were computed from both dictation and transcription data. The computation of inter-interviewer agreement was analogous to the standard way in which inter-observer agreement is typically calculated: number of agreements divided by the number of agreements plus disagreements (Johnson & Bolstad, 1973). This formula was adapted, for the purposes of the present study, in order to determine the percent agreement for problem areas identified. In addition, within each problem area, the level of inter-interviewer agreement on the particular problem items identified was also calculated from the transcribed interview data. Problem items consisted of specific problem behaviors that were subsumed under a general problem area heading (e.g., Problem Area: Sex; Problem Item: Premature Ejaculation)

Interviewer Accuracy

In the present study, all the interviewers could conceivably agree on the number and/or specific problem areas identified for a particular client and all be inaccurate. To determine the accuracy of human interviews, the information reported in the dictations and transcriptions was compared to the information obtained by a computer interview of the same clients (computer criterion) with respect to the number of

problem areas identified. The Medication Evaluation and Resources Program at Duke University Medical Center has developed a computer-assisted comprehensive behavioral interview procedure. An interactive computer is employed to administer a broad spectrum screen for problem behaviors within 23 life areas. Questions are displayed by a CRT video terminal and the client types answers on a typewriter keyboard to an on-line computer. The computer interview possesses high content validity. Initial data have shown the computer to be superior to the human interviewer in the identification of problem behaviors. Twenty-eight married clients were seen prior to the transition to computer-assisted interviewing. Sexual difficulties were identified in 29 percent of these clients. The computer screen, however, revealed sexual difficulties for 86 percent of its clients (n=14). In addition, follow-up data for the last 25 clients seen by the program have not revealed a single problem behavior that was not identified by the computer interview.

Comparisons between the dictations, transcriptions, and computer information for each client were made to determine the percent of problem areas identified in which the computer and human interviewers agreed. In addition, these same comparisons to the computer criterion were made for problem items identified in the transcriptions.

Factors Affecting the Reliability of Interview Data

There are a number of variables or factors that potentially could have effects on the information gathered during a behavioral interview. Two sources of interviewer variance that may attenuate inter-interviewer agreement include factors of input and output variance. These same factors may be considered sources of input and output error hindering the accuracy of the behavioral interview. Other potential factors include the consistency of client responses and interviewer-interviewee biases.

Input variance refers to differences resulting from such variables as the number, type, and structure of interview questions. Input error includes the omission of critical questions, and to a lesser extent, the asking of irrelevant questions. Researchers have found a number of interviewer behaviors to differentially affect client statements: activity level of the interviewer with respect to the number of interviewer statements, (Heller, 1971); structure of the questions (Maccoby & Maccoby, 1957); clinical versus "street" language and even slight differences in the phrasings of questions (Cantril, 1944). Specific noncontent measures of interviewer behaviors have also been found to vary between interviewers (high inter-interviewer variability) but be reliable for the individual interviewer (low intra-interviewer variability). These variables include the frequency and duration of speech units and silences (Leonard & Bernstein, 1960; Matarazzo, Wein, & Saslow, 1965). Finally, an extensive literature

(e.g., Kanfer & Phillips, 1970) has demonstrated the shaping of verbal behavior resulting from the differential reinforcement (selective recording and verbal conditioning) of response content by the interviewer's behavior.

Output variance refers to interview data variations due to the selective recording of client responses by the interviewer Output error refers to the inaccurate (error of commission) and incomplete (error of omission) recording of client information. Typically, interviewers make notes of important client response during the interview and subsequently write or dictate a summary of the obtained client information immediately following the interview. Low inter-interviewer agreement in the recording of client responses has been found both when confederates were employed as interviewees and instructed to give the same answers to each interviewer (Smith & Hyman, 1950), and when numerous interviewers were asked to interview the same client (Guest, 1947). Symonds & Dietrich (1941), Corner (1942), Guest (1947) and Payne (1949) compared taped transcriptions to interviewer reports of client interview information. All reported a loss of information, with this loss increasing as a function of time between the actual interview and the writing of the report (Symonds and Dietrich, 1941). Further, these researchers reported that in addition to considerable losses in the quantity of interview content, marked distortions of client responses frequently occurred in reports written following

the interviews. In fact, Payne (1949) found as much as 25 percent of the statements attributed to respondents to be clearly incorrect.

In addition to these interviewer factors, variability in the information gathered by different interviewers from the same client may be attributable to inconsistencies in the client's responses. The extent to which variability in client response has been found to affect interview information ranges from approximately 5 percent in psychiatric interviews (Ward et al., 1962) to 75 percent on standard biographical questions (Bancroft, 1940; Hyman, 1944). The degree of consistency observed seems to be affected by the differential consequences for certain answers that are operating in each interview situation. Braginsky & Braginsky (1967), for example, found that patients tended to respond in ways that maximized their chances of accomplishing the implicit purpose of the interview. Specifically, if a psychiatric patient thought the purpose of the interview was to assess competence for discharge, an increase in bizarre behavior was observed. On the other hand, if that same patient thought the purpose of the interview was to assess competence for increased in-hospital benefits, verbal and motor behavior became increasingly appropriate.

Other factors, usually referred to as interviewer-interviewee biases, may influence the content of client information gathered by a human interviewer. These factors may be more appropriately termed stimulus control factors, as they imply that the interviewer and/or interviewee's behavior is being

controlled by inappropriate stimulus variables. Interviewer and client demographics such as age (Riesman & Ehrilich, 1961), sex (Benny, Reisman, & Star, 1956), social class (Lenski & Leggett, 1960) and racial background (Athey, Coleman, Reitman, & Tans, 1960; Katz, 1942) have been reported to affect the questions posed, answers given, and answers recorded during the interview process (Schwitzgibel & Kolb, 1975).

The extent to which the factors of output, input, client consistency and interviewer-interviewee biases affect the assessment information gathered during the behavioral interview is an unanswered question. The studies cited did not evaluate the affect of these factors on behavioral interviewing procedures. Rather these studies focused on the interview in its traditional roles as a diagnostic instrument or vehicle for psychotherapy. It was a primary task of the present study to examine the impact of three of these factors on the information gathered during behavioral interviews. Interviewer-interviewee biases were not investigated in the present study.

Input

The interview transcriptions of a given client were compared in terms of the number of problem areas sampled by each interviewer. Further, the transcriptions were compared to determine whether interviewers asked questions in the same areas and more specifically whether they sampled the same problem items within each area (Input Variance). The interview transcriptions were also compared to the computer criterion

to establish the comprehensiveness of the human interviewers (Input Error).

Output

Interviewer dictations were compared to the verbatim transcriptions of each interview. The loss of interview information was determined by comparing the number of problem areas identified as problems during the interview and included in the dictations to the total number of problem areas actually identified in the transcriptions (Omission Error). The accuracy of information in the dictations was also established by determining what percent of the areas reported as problems in the dictations were actually on the tape recorded transcriptions (Commission Error).

Consistency of Client Responses

The consistency of client responses across interviewers was determined in the present study by comparing client responses to those questions that were asked by multiple interviewers. In addition the client's responses to questions asked by both a human interviewer and the computer were compared.

Summary of Objectives

Objective 1 (Generalizability): To determine whether the four interviewers would differ from each other on the number of problem areas identified for the same clients in both their transcriptions and dictations of the interview information.

Objective 2 (Inter-Observer Agreement): To determine whether interviewers would identify different problem areas for the same clients.

Objective 3 (Inter-Observer Agreement): To determine whether interviewers would identify different problem items for the same client within each problem area.

Objective 4 (Inter-Interviewer Accuracy): To determine whether human interviewers would be less accurate than the computer and identify fewer problem areas, different problem areas, and fewer problem items than identified on the computer criterion for each client.

Objective 5 (Input Variance): a) To determine whether interviewers would ask questions concerning different numbers of areas; b) To determine whether interviewers ask questions about different areas; and c) To determine whether interviewers would ask questions about different items.

Objective 6 (Input Error): To determine whether interviewers would sample a limited number of problem areas or items within an area.

Objective 7 (Output Variance): To determine whether significant amounts of interview information would be lost as a result of the selective recording of client responses by various interviewers.

Objective 8 (Output Error): To determine whether interviewers would include a number of problem items in their dictations that were not actually sampled during interview sessions.

Objective 9 (Client Consistency): To determine whether client responses would be highly consistent across computer and human interviewer modalities.

CHAPTER II

METHOD

Setting

The present study was conducted in the facilities of the Medication, Evaluation, and Resources Program (MEARP) located in the Civitan Building, Duke University Medical Center. MEARP was supported by Public Health Service Grant 5H81DA 01665-02 from the National Institute of Drug Abuse (NIDA). MEARP was established as one of eleven pilot facilities funded by NIDA nationwide to study the use of psychoactive medication. Typically, clients were referred to MEARP from either local mental health centers or private physicians for psychological evaluation. All clients were seen free of charge and perfunctorily received a physical examination. A prospective client had to meet only two criteria to qualify for MEARP: 20 to 65 years of age and using a psychoactive medication.

Subjects

Client Selection Procedures. All clients referred to MEARP for psychological evaluation were asked to rate 23 life areas (Appendix A) on a 5 point rating scale ranging from (1) no difficulties to (5) very many difficulties. This initial rating scale was computerized so that clients viewed the questions as they were presented on a cathode ray video tube

terminal and responded by pressing the appropriate numbers on a typewriter keyboard. The first four clients who rated three or more of the 23 areas a (5), very many difficulties, were asked to participate in the study.

All clients asked to participate agreed to be subjects in the study. Prospective clients were told that the purpose of the study was to determine the relative effectiveness of the computer as an interviewer by comparing the quality of computerized and human interviews. Clients were also informed that the study would require them to participate in five interviews: one computer and four human interviews. Each client signed a consent form acknowledging his role in the research project. In addition, clients signed a consent form allowing audiotaping of their human interview sessions. Clients were identified by numbers throughout the study to insure the confidentiality of the information gathered.

Client Characteristics. Relevant demographic information for the clients who participated in the present study is presented in Table 1. (Table 1 and all subsequent tables are located in Appendix E.)

Client 1 had an extensive psychiatric history with numerous hospitalizations and was under treatment (antipsychotic and antidepressant medication) during the entire course of the study. Client 2 had been seen on an outpatient basis sporadically for a two-year period. Client 3 had a three-year history of outpatient treatment with one hospitalization following a drug overdose. Client 4 was an inpatient on a psychiatric unit throughout the course of the study.

Interviewers. Four advanced graduate students from the Department of Psychology at the University of North Carolina at Greensboro served as interviewers. A notice posted in the Greensboro Psychology Department advertised the opportunity to earn money for conducting behavioral interviews at the Duke University MEARP project. Only applicants who had a minimum of 1000 hours of clinical experience and who had completed both the Behavioral Assessment and Behavior Modification Theory and Practicum courses at the University of North Carolina at Greensboro were asked to apply. Interviewers were paid $30 a day plus travel and lunch expenses. All applicants were informed that a minimum of four interviews would be required. The first four students to sign the notice and who met the above criteria were included in the study. Each interviewer was given written instructions concerning how the behavioral interviews were to be conducted. These instructions specified that a comprehensive interview was desired in which the goal was to identify all problem behaviors (excesses and deficits) of the client (Appendix B). The interviewers were informed of the purpose of the research and each interviewer was asked to sign a research consent form allowing audio taping of their interview sessions.

Apparatus

Interview Rooms. Interviews were conducted in two eight-foot by ten-foot rooms furnished in modern decor. Each room was wired for audio taping of client sessions. The audio taping equipment consisted of two microphones on floor stands

which were placed in front of the interviewer and the interviewee. The microphones were connected to a Superscope recorder located in an adjacent room. A Cathode Ray Tube terminal, described below, was located on a table in each room.

Cathode Ray Tube Terminals. Two Applied Digital Data Systems, Inc. (ADDS) Console 580 terminals, with 11-inch diagonal terminal screens were used to display questions and accept client answers. The terminal displayed data in a format of 24 lines with 80 characters per line making a total of 1920 characters. The rate of data transmission was 960 characters per second. High legibility was achieved by displaying data as black characters on a white background. Clients typed responses on the console's typewriter keyboard on-line to the computer.

Computer. (1) Hardware: The computer was a PDP 11 mini-computer with 124,000 words of core memory and 6.6 million words of disc storage. It had 16 input-output lines (1/0 interface lines). (2) Software: The software was Digital Equipment Corporation's Resource Sharing Time System (RSTS/E) operating with Extended Basic Language. This system supported multiusers.

Computer Problem Behavior Interview (Interactive). The computer interview was organized in four sections: (a) Client Characteristics; (b) Problem Behaviors; (c) Drug History; (d) Motivation. Within the Problem Behavior section,

there were 23 problem areas (e.g., marriage, sex, employment, social isolation, assertion, sleep, tension). The number of specific questions in a problem area ranged from 20 to 80 items. The Problem Behavior Interview contained approximately 1540 question and answer items. The entire computer screen contained over 2000 items.

Approximately 97 percent of the questions were presented to the client in a multiple-choice format. The following are sample questions:

How strong is your fear, concern, or discomfort to people in authority?

( )

1. Extremely unpleasant
2. Very strong
3. Moderately strong
4. A little strong
5. None or minimal

Who in authority causes you to be fearful? (Answer 1 for yes, 2 for no, B for backup.)

( ) Spouse
( ) Employer
( ) Parents
( ) Other relatives
( ) Doctors, ministers, professional people
( ) Older people
( ) Other

The computer interview followed a linear path through a comprehensive survey of life areas: There was very little branching among items as a function of the client's previous answers.

Content validity was stressed in the computer interview. The initial item pool was gathered through various procedures. First, questionnaire surveys emphasizing behavioral rather than intrapsychic variables were reviewed for appropriate items. Second, clinical case histories were reviewed. In each case, the question was asked whether the computer screen would have revealed the problem described in the history. If not, then appropriate items were systematically added to the item pool.

The strategy of the Computer Problem Behavior Interview was to enumerate problem behaviors and not to provide a functional description of antecedent and consequent events. The product of the interview was a printout of relevant client information and of behavioral deficits and excesses.

Experimental Design

A four-by-four Latin square design with the independent variables of client and interviewer was employed. Each of four clients was interviewed by four different human interviewers. Each of the interviewers served as a first, second, third, and fourth human interviewer for a different client.

In addition, each client participated in a computer interview. In order to control for sequence effects, the order of participation in the human and computer interviews was counterbalanced. Thus, two clients participated in the human interviews prior to the computer interview and two clients participated in the computer interview prior to the human interviews.

Procedure

Each client participated in five interviews: one computer and four human interviews. Interviews were scheduled as closely in time as possible.

Computer Interview. The client was instructed in the use of the CRT terminal by the clinic coordinator. She demonstrated how to answer questions and advance to new question frames on the CRT typewriter keyboard. The client was told to work at her own pace and that clinic personnel were available to answer any questions.

Human Interviews. The human interviewers were given written instructions on how to conduct a comprehensive behavioral interview (Appendix B). Interviewers were also given a summary of the relevant demographic information of each client they interviewed. No time limit was placed on the duration of the interviews. The durations of the human interviews are presented in Table 2. All sessions were audio-tape recorded. In addition, interviewers were required to dictate the information obtained about the client at the end of each interview. Interview information and dictations were transcribed verbatim into a typed question-answer format. Typed interview and dictation transcriptions were compared to their corresponding audio-recordings by the program secretary to insure their comprehensiveness and accuracy.

After each interview, clients were asked to fill out a questionnaire rating their human or computer interviewer on a number of dimensions (Appendix C).

Raters and Rating Procedures

Raters. Two Master's level psychologists with behavioral training each independently coded the content of all transcribed human interviews, computer printouts, and dictations according to a behavioral classification system (Appendix D). Raters had extensive familiarity with the classification system. In addition, two one-hour sessions were held to clarify terminology and any areas of ambi uity in the classification system. The behavioral classification system represented a total listing of all the behavioral items sampled by the computer. Items were listed by life area (e.g., marriage, sex) with the specific content information gathered under each category detailed. For the purposes of this study, this classification system was considered to represent the content of a comprehensive behavioral interview.

Rating Procedures. (1) Human Interview Transcriptions: a. The raters indicated which of the 23 life areas the interviewer identified as a problem area. An area was considered as an identified problem area when two criteria were met: 1. the name of the area or any one of the classification system items within that area was mentioned; and, 2. the frequency, duration, or intensity of difficulty was interfering at least moderately with the interviewee's functioning. The same criteria were used by the raters in evaluating the dictation transcriptions and computer interviews. b. Within each identified problem life area, the raters designated those items specified as problems. c. Raters also indicated in

which of the 23 life areas the interviewer asked questions. An area was considered as inquired about if either the area itself was named or if the interviewer inquired about any item listed in the classification system under that area. d. Within each of the areas the interviewer inquired about, the raters indicated which particular items were sampled by the interviewer. (2) Dictations: a. The raters specified which of the 23 life areas were identified as problems. b. Within each identified problem life area, the raters determined which specific items were identified as problems. (3) Computer Interview: a. The raters coded printouts from the Computer Problem Behavior Screen for the life areas identified as problems. b. The raters also noted the particular items identified as problems in each problem life area.

Inter-Rater Agreement. Two raters independently coded the content of the human interviews, computer printouts, and dictations for areas and items queried and identified as problems. Inter-rater agreement of the raters' data was calculated by dividing the number of agreements by the number of agreements plus disagreements and multiplying by one hundred. The mean inter-rater agreement for problem areas identified was .90, .83, and .98 for transcriptions, dictations, and computer interviews, respectively. For problem items identified, the mean inter-rater agreement scores were .87, .78, and .93. For areas questioned, agreement scores were .88, .79, and .95. Inter-rater agreement scores for items questioned were .85, .75, and .91 respectively.

Only areas or items that both raters agreed were questioned or identified as problems were included in subsequent analyses.

CHAPTER III

RESULTS

The results of the present study are presented in four major sections. The first three sections summarize the results for each of the three methods of establishing reliability discussed earlier: the generalizability (G) study; the computation of inter-interviewer agreement; and the measurement of inter-interviewer accuracy. The fourth section reviews the findings for a number of interviewer and client factors that may affect the reliability of interview information.

Generalizability: Number of Problem Areas Identified

In order to determine the generalizability of data obtained from behavioral interviews across interviewers, two four (interviewers) x four (clients) repeated measures analyses of variance were calculated on the number of areas identified as a problem for each client as coded by the raters from the interview transcriptions and dictations, respectively (see Tables 3 and 4).

A significant main effect for clients was obtained for the number of areas identified as problems in the interview transcriptions ($F = 4.17$; $df = 3,9$; $p < .05$) indicating that the clients differed significantly with respect to the mean number of areas identified as problems across the

interviewers. The mean number of problem areas identified in the interview transcription for each of the four clients was 13.25, 13.00, 9.00 and 8.00, respectively. Newman-Keuls post hoc comparisons revealed no significant differences between the means. No other significant main effects or interactions were obtained in this analysis.

A significant main effect for clients was also obtained for the number of areas identified as problems in the interview dictations ($F = 10.63$; $df = 3,9$; $p < .01$) indicating that the clients also differed significantly with respect to the mean numbers of areas identified as problems by the interviewers on the dictations. The mean number of problem areas identified on the dictations for each client across interviewers were 9.25, 11.00, 9.75, and 9.00, respectively. Newman-Keuls post hoc comparisons indicated that the means did not differ significantly. No other significant main effects or interactions were revealed.

In summary, these analyses indicated that clients differed as to the number of areas identified as problems by the interviewers on both interview transcriptions and dictations as coded by the raters. The results failed to reveal a significant main effect for interviewers: interviewers were not found to identify significantly different numbers of problem areas on either the interview transcriptions or dictations. These results suggest that it is possible to generalize across interviewers in terms of the overall

number of problem areas identified for a client during an interview.

The quantitative nature of comparisons utilizing the generalizability (G) study methodology does not allow the estimation of inter-interviewer agreement on the specific areas and items identified as problems. Consequently, although it may be possible to generalize across interviewers with respect to the number of problem areas identified, inter-interviewer agreement on the specific areas and items identified as problems may be limited.

Inter-Interviewer Agreement-Specific Problem Areas and Items

Although interviewers were not found to identify different numbers of problem areas, additional analyses were computed to determine the percent agreement with respect to the specific areas identified as problems for each client. The following reliability (inter-interviewer agreement) coefficient was calculated using the data coded from interview transcriptions and dictations:

$$\frac{\text{Number of Areas Both Interviewers Indicated as Problems}}{\text{Total Number of Areas Both or Either Interviewer Indicated as Problems}}$$

Every interviewer was compared with every other interviewer: six comparisons per client or 24 total comparisons. The results of these comparisons for interview transcriptions are presented in Table 5. The inter-interviewer agreement for the identification of problem areas in the transcriptions

ranged between .25 and .76. The average inter-interviewer agreement across the 24 comparisons for the identification of areas as problems in the transcriptions was .55.

The results of these comparisons for dictations are presented in Table 6. Inter-interviewer agreement for the identification of problem areas on the dictations ranged between .22 and .80. The mean inter-interviewer agreement across the 24 comparisons was .48.

To determine the agreement between human interviewers on the number of items identified as problems within an area identified as a problem for a client by more than one interviewer, the following reliability coefficient was calculated on the coded data from interview transcriptions. Coded and transcription data as opposed to dictation data were utilized in this and subsequent estimations of agreement in order to maximize inter-interviewer agreement scores.

$$\frac{\text{Number of Items Identified as Problems Within an Area by Both Interviewers}}{\text{Total Number of Items Identified as Problems Within an Area by Both or Either Interviewer}}$$

There was a total of 25 potential problem areas for each client. The mean inter-interviewer agreement scores for problem items within each identified problem area are presented in Table 7. Since inter-interviewer agreement for problem items in each area was calculated only when more than one interviewer identified an area as a problem for a

particular client, the number of inter-interviewer agreement coefficients comprising the means presented in Table 7 varied per problem area from 2 to a maximum of 24 inter-interviewer agreement coefficients. Agreement scores per problem area ranged between .10 and 1.00. The average inter-interviewer agreement for items identified as problems per problem area across interviewers and clients was 0.40.

While the results of the previous section indicated that interviewers did not differ with respect to the overall number of problem areas identified for a client, the findings of the present section suggest that inter-interviewer agreement on specific problem areas and items was attenuated.

The next section investigated the accuracy of interview information. Interviewers could agree on the number and/or specific problem areas or items identified for a particular client and be inaccurate. In the present study, the results of the computer interview were considered to be an accurate criterion and the results obtained during the human interviews with each client were compared to the computer printout in order to determine interviewer accuracy.

Interviewer Accuracy: Comparisons Between Human Interviewers and a Computer Criterion

The accuracy of human interviewer information was determined by comparing coded human-interview transcriptions and dictations with the computer-interview criterion. The number of problem areas identified on coded transcription and

dictation data was compared to the number of problem areas identified from coded computer data for each client. In addition, the specific agreement between each interviewer and the computer in identifying specific problem areas and items was also calculated.

To determine the accuracy of the human interview transcriptions and dictations as compared to the computer criterion with respect to the number of areas identified as problems, two five (4 human interviewers + 1 computer interview) x four (clients) repeated measure analyses of variance were calculated on coded interview dictation and transcription data respectively. In these analyses, the computer was treated as a fifth interviewer.

The results of the analysis comparing coded interview transcription and computer interview data revealed a significant main effect for clients ($F=4.609$; $df=3,12$; $p < .05$) indicating that the clients differed significantly with respect to the number of areas identified as problems across the four human interviewers and the computer interview (see Table 8). The mean number of areas identified as problems across the five interviews for each of the four clients was 13.80, 13.40, 9.40 and 9.40 respectively. Newman-Keuls post hoc comparisons revealed no significant differences between means. No other significant main effects or interactions were obtained in this analysis.

The results of the analysis comparing coded dictation and computer interview data, however, did indicate a significant main effect for interviewers ($F=13.36$; $df=4,12$; $p < .01$) on the number of problem areas identified (see Table 9). The mean numbers of areas identified as problems by each human interviewer was 9.00, 6.75 and 8.75, and 8.00 respectively. The mean number of problem areas identified by the computer interview across clients was 14.25. Newman-Keuls post hoc comparisons revealed significant differences between the mean number of areas identified as problems by the computer interview across clients and the mean number of areas identified as problems from coded dictation data by each of the four human interviewers across clients ($p < .01$ in all comparisons). Human interviewers were not found to differ significantly from each other in the mean number of areas they identified as problems across clients from coded dictation data.

The results of this analysis comparing coded dictation and computer interview data also indicated a significant main effect for clients ($F=11.40$, $df=3,12$; $p < .01$) indicating that the clients differed significantly with respect to the mean number of areas identified as problems across the four human interviewers and the computer interview. The mean number of areas identified as problems across the five interviews for each of the four clients was 10.60, 12.00, 6.80 and 8.00 respectively. The results of the Newman-Keuls post hoc comparisons indicated that Client 2 (Mean 12.00)

differed significantly ($p < .01$) from Client 3 (Mean 6.80) and Client 4 (Mean 8.00) and that Client 1 (Mean 10.6) differed significantly ($p < .05$) from Client 3 (Mean 6.80) in the number of areas identified as problems across the four human interviewers and the computer interview.

In summary, these analyses indicated that clients differed as to the number of areas identified as problems (i.e., coded by raters from transcriptions, dictations and computer printouts) by the four human interviewers and the computer interview. In addition, it was determined that each of the human interviewers (coded dictation data) differed from the computer criterion with respect to mean number of areas identified as problems across clients. Human interviewers consistently identified fewer problems on the average than the computer criterion.

To determine the specific agreement between each interviewer and the computer in identifying problem areas, the following was computed on coded transcription data:

$$\frac{\text{Number of Areas Identified as Problems by Both the Computer and Human Interviewer}}{\text{Total Number of Areas Identified as Problems by Both or Either the Computer or Human Interviewer}}$$

Each interviewer was compared with the computer interview: four comparisons per client or 16 total comparisons. The results of these comparisons are presented in Table 10. Agreement scores for areas identified as problems ranged between .27 and .68 with a mean of .55 across all 16 comparisons.

To determine the agreement between each interviewer and the computer in identifying specific items as problems, the following was also computed on coded transcription data for all areas identified as problems by both the computer and human interviewer.

$$\frac{\text{Number of Items Identified as Problems Within a Problem Area by Both the Human Interviewer and Computer}}{\text{Total Number of Items Identified as Problems Within a Problem Area by Both or Either}}$$

The agreement for items identified as problems was averaged for each problem area across the four clients. These means are presented in Table 11. The number of agreement coefficients between the computer and human interviewers comprising the means in Table 11 varied per problem from 1 to a maximum of 16. Agreement scores per problem area ranged between .07 and .54. The average agreement for items identified as problems per problem area between the computer criterion and human interviewers across clients and interviewers was .24.

The results of this section suggest that not only is there limited agreement among interviewers as to the specific areas and items identified as problems for a particular client, but also the accuracy of human interview information is in question. When human interview information is compared to a standardized computer interview (criterion) there is limited agreement as to those specific areas and items which are identified as problems for a particular client.

Factors Affecting the Reliability of Interview Data

A number of factors have been mentioned that potentially could have effects on the information gathered during an interview. These factors have been previously defined to include two sources of interviewer variance and error, input and output and a third factor, the consistency of client responses.

Input. Input variance refers to differences in interview information resulting from such variables as the number and specific questions posed to a particular client by various interviewers. In the present study, input variance was assessed first by determining whether interviewers differed as to the number of areas they questioned during an interview.

In order to determine whether interviewers differed as to the number of areas they questioned during the interviews, a four (interviewers) x four (clients) analysis of variance was calculated using coded interview transcription data on the number of areas in which the interviewers asked a client at least one question. A significant main effect for interviewers ($F=4.02$; $df=3,9$; $p < .05$) was obtained indicating that the interviewers differed with respect to the mean number of problem areas investigated during the interviews across clients (Table 12). The mean number of problem areas inquired about across clients was 12.50, 12.25, 14.75, and 16.00 for each interviewer respectively. Newman-Keuls post hoc comparisons revealed a significant difference between Interviewer 2 (12.25) and Interviewer 4 (16.00). No other significant main effects or interactions were obtained.

Input variance was also assessed by comparing coded interview transcription data to determine whether interviewers asked questions in the same areas and whether they questioned the same problem items within each area.

Agreement scores comparing the specific areas questioned by the interviewers were calculated as follows:

$$\frac{\text{Number of Specific Areas About Which Both Interviewers Asked Questions}}{\text{Total Number of Specific Areas About Which Both or Either Asked Questions}}$$

Every interviewer was compared with every other interviewer: six comparisons per client or 24 total comparisons. The results of these comparisons for transcription data are presented in Table 13. Agreement scores for specific areas questioned ranged between .33 and .87 with a mean of .62 across all 24 comparisons.

The inter-interview agreement for specific items questioned within each problem area was computed using the data from interview transcriptions:

$$\frac{\text{Number of Specific Items Questioned by Both Interviewers}}{\text{Total Number of Specific Items Questioned By Both or Either Interviewer}}$$

There was a total of 25 potential problem areas for each client. For each area all interviewers who inquired about the area were compared with respect to the specific items they questioned. The agreement for specific items questioned by each pair of interviewers was averaged for each problem area

across the four clients. The number of agreement coefficients comprising the means presented in Table 14 varied per problem area from 2 to a maximum of 24 inter-interviewer agreement coefficients. Agreement scores per questioned area ranged between 0.00 and 0.74. The average inter-interviewer agreement for specific items questioned per questioned area across interviewers and clients was 0.29.

Input error was defined as the omission of critical questions. In the present study, input error was assessed by comparing coded human interviewer transcription data to the computer criterion to establish the comprehensiveness of the human interviewers in terms of the number of areas and items questioned.

To ascertain the comprehensiveness of the human interviewer, the percent of life areas (as specified by the computer standard) about which each interviewer asked at least one question was calculated:

$$100 \times \frac{\text{Number of Areas Questioned}}{\text{Total Number of Problem Areas (25)}}$$

The computer standard sampled a total of 25 potential problem areas. Table 15 shows the percent of the areas questioned by each interviewer during every client interview. Percentages ranged from 36 to 76 percent. The mean percent of areas questioned across interviewers and clients was 55 percent (16 comparisons).

A further determination of the comprehensiveness of the human interviewers was made by calculating percent of potential computer items questioned by each interviewer within each area across the four clients:

$$100 \times \frac{\text{Number of Items Questioned in an Area by Each Human Interview}}{\text{Number of Items Questioned by the Computer Interview}}$$

The results of these computations are summarized in Table 16. For each area, the percent of computer items questioned by each interviewer across the four clients is presented. In addition, the mean percent of potential computer items questions for each area across the four interviewers is presented. The mean percent of items questioned per area ranged from 0.00 to 13.5 percent. The average percent of items questioned across interviewers, clients and areas was 6.03 percent.

To summarize, interviewers in the present study varied significantly in the number of areas about which they asked client's questions and with respect to the specific areas and items questioned. In addition, human interviewers were shown to lack comprehensiveness in comparison to a computer criterion. These results suggest that input variability and error factors may account in part for variations in the information gathered from the same client by different interviewers.

Client Responses. Variability in the problems identified for the same client by different interviewers may also be attributable to inconsistencies in client responses. Clients

may be inconsistent in their responses to the same question posed by different interviewers. Such variation in client responses may affect the areas and items identified as problems for a particular client by each interviewer.

In the present study, the consistency of client responses across the four interviews was determined by comparing client answers to those questions which were posed to a client by more than one of the interviewers. Raters coded the transcriptions of the actual interviews for both questions posed by the interviewers and client responses to these questions. Thus, it was the raters' decision whether or not a client had indicated an item or area as a problem during each interview. In addition, the client responses to questions asked by both a human interviewer and the computer interview were compared.

In order to estimate the consistency of client responses to the same questions posed by different interviewers, the following agreement scores were calculated on the coded transcription data for each pair of interviewers for each of the four clients:

$$\frac{\text{Number of Questions Asked by Each Pair of Interviewers to Which Raters Indicated a Client Gave a Consistent Response (Problem or No Problem)}}{\text{Total Number of Items Questioned by Each Pair of Interviewers}}$$

For each client, comparisons were made of the client's responses when two interviewers questioned the same item. Thus, a maximum of six comparisons per item for each client were possible. Agreement scores with respect to client responses (problem or no problem) ranged between .67 and 1.00. The

average agreement score for client responses per pair of interviewers across clients was .86. These results, in addition to the mean agreement for client responses for each pair of interviewers, are presented in Table 17.

To determine the consistency of client responses to questions posed by both the computer and at least one human interviewer, the following agreement scores were computed on the coded transcription data and computer interview data for each of the four clients:

$$\frac{\text{Number of Questions Asked by the Computer and a Human Interviewer to which Raters Indicated a Client Gave a Consistent Response (Problem or No Problem}}{\text{Total Number of Items Questioned by Both the Computer and Human Interviewer}}$$

For each client, comparisons were made of the clients' responses (Problem or No Problem-as assessed by the Raters) when a human interviewer and the computer interview both had questioned the same item. The agreement scores for the consistency (Problem or No Problem) of client responses to questions posed by the computer and human interviewers is also presented for each client (Table 18). The average agreement score for the consistency of client responses across clients and interviewers to questions asked in both the computer and human interviews is .89.

To summarize, these results indicate that clients were relatively consistent in their responses to questions asked by either a pair of human interviewers or a human interviewer and the computer.

Output. Output referred to interview data variations resulting from the selective recording of client responses by the interviewer. In the present study, the impact of output variables was assessed by comparing interviewer dictations to the verbatim transcriptions of each interview. The loss of information that would result from incomplete (error of omission) recording of client information was determined by comparing the number of areas identified as problems during the interview (transcriptions) and included in the dictations to the total number of areas actually identified in the transcriptions. In addition, the accuracy of information (error of commission) in the dictations was established by determining what percent of the areas reported as problems in the dictations were actually indicated as a problem by the client on tape recorded transcriptions of the interviews.

Omission Errors. To establish the percent of information collected by an interviewer and subsequently not included in a dictation, the following was computed from coded transcription and dictation data:

$$1 - \frac{\text{Number of Areas Identified as Problems During Each Interview and Reported in the Dictation}}{\text{Total Number of Areas Identified as Problems During the Interview}}$$

There was a total of 16 human interviews and corresponding dictations. The results of these comparisons are presented in Table 19. Percentages of areas identified during the interview and not commented about on dictations ranged

from 0 to 55 percent. The average percent of areas identified as problems in interview transcriptions and not reported in dictations across interviewers and clients was 28 percent.

Commission Errors. To determine the percent accuracy of the information reported in dictations, the following was computed from interview transcription and dictation data:

$$1 - \frac{\text{Number of Problem Areas Reported in Each Dictation That Agreed With the Transcription of the Same Interview}}{\text{Total Number of Areas Reported in the Dictation}}$$

There was a total of 16 comparisons. The percent accuracy of problem areas reported in each interview dictation is presented in Table 20. The percentage of areas identified as problems in interview dictations that did not agree with areas identified in interview transcriptions ranged between 0 and 33 percent. The average percent of areas that were identified as problems in dictations that were not identified from interview transcriptions was 5 percent. Consequently, commission errors ranged between 0 and 33 percent with an average error of 5 percent across interviewers and clients.

These results indicate that output variables may have substantial effects on the information that is actually reported following an interview. The greatest impact of output factors was in the loss of interview information: on the average over 25 percent of the areas identified as problems were lost in dictation (omission error). Actual distortions of client responses (commission error), however, were minimal, on the average of 5 percent.

Client Ratings: Human and Computer Interviews

Following each interview, clients completed a questionnaire rating their human and computer interviewers on a number of dimensions (Appendix C).

Clients rated the human interviewers on the dimensions of Empathy, Genuineness, and Warmth. All human interviewers were rated either 4 or 5 on three dimensions. The clients indicated that they "strongly liked" participating in the computerized interview. Three of the four clients "somewhat preferred" being asked questions by the computer rather than by the human interviewers. One client "much preferred" being asked questions by the computer. All four clients "somewhat preferred" being asked personal questions by the computer rather than by the human interviewers.

CHAPTER IV

DISCUSSION

The primary objective of the present study was to investigate the reliability of the behavioral interview. This objective was operationalized in three ways. First, the reliability of the behavioral interview was examined quantitatively, utilizing the methodology of the Theory of Generalizability (Cronbach et al., 1970). A generalizability (G) study was performed to ascertain the generalizability across interviewers with respect to the number of problem areas identified per client. Second, the computation of inter-interviewer agreement allowed for finer, more qualitative comparisons. The agreement among the interviewers as to those specific areas which were identified as problems for a particular client and as to the specific problem-items within an identified area was determined. Finally, the accuracy of interview data was measured both quantitatively and qualitatively by establishing the agreement between each human interviewer and a standardized criterion interview (computer interview) for the number of areas and the specific areas and items identified as a problem for each client.

The results of the G study analyses indicated no main effects for interviewers: interviewers were not found to identify significantly different numbers of problem areas on

either the interview transcriptions or dictations. While these results suggest that it is possible to generalize across interviewers with respect to the overall number of problem areas identified for a client, inter-interviewer agreement on specific problem areas and items was attenuated. The average inter-interviewer agreement across interviewer pairs and clients as to those specific areas which were identified as problems for a client was .55 on interview transcriptions and .48 on interview dictations. The average inter-interviewer agreement per area for specific items coded by the raters as identified problems from the transcriptions was .40.

The finding of generalizability for numbers of problem areas and minimal agreement on specific areas underscores a limitation of the generalizability methodology. The analysis of variance or G study model allows comparisons between magnitudes or amounts (interval-scale measurements) but does not allow for the establishment of exact point-to-point correspondences in recording (Baer, 1977; Cone, 1977; Nelson, Hay & Hay 1977). In the present study it would have been misleading to rely soley on the assessment of reliability from the generalizability study. A similar problem has been recognized by those researchers employing naturalistic observation procedures: Two observers watching the behavior of one subject may agree as to the number of times a behavior occurred during a specified time interval (frequency) yet not have recorded these behaviors at the same time. Thus, two observers may have each recorded a behavior as occurring five times (100 % agreement for response

frequency) when in fact the behavior could have occurred 10 times and the specific inter-observer agreement could be 0 %.

Conceivably, interviewers could agree on the number and/or specific problem areas or items identified for a particular client and be inaccurate. In terms of the number of problem areas identified, the results of a five (4 human + 1 computer interviews) x four (clients) repeated measures analysis of variance indicated a significant main effect for interviewers when interview dictation data were compared to computer interview results. The human and computer interviewers differed significantly in the number of problem areas identified on the dictations, with the computer interview identifying a greater number of areas as problems. Comparisons of human interview transcriptions and computer data produced no significant differences. These data suggest that a loss of information occurred when interviewers dictated summaries of their interviews.

More qualitative analyses of the specific agreement between each interviewer and the computer in identifying problem areas and items clearly demonstrated the lack of concordance between human and computer interviews. The average agreement between an interviewer and the computer as to those specific areas which were identified as problems for a client was .55. In those areas which were identified as a problem by both the

computer and at least one interviewer for a given client, the average agreement per problem area between an interviewer and the computer as to which specific items were problems was .24. Consequently, not only is there limited qualitative agreement among interviewers as to the specific areas and items identified as problems for a particular client but also the accuracy of human interview information when compared to a standardized computer interview is in question.

In summary, the reliability of the behavioral interview was examined using the methodology of the Theory of Generalizability and two more commonly employed methods of assessing reliability, inter-agreement and accuracy. It was found that the number of problem areas identified was generalizable (reliable) across interviewers. These results are in some ways misleading, however, since measurement of inter-interviewer agreement on specific areas and items and accuracy compared to the computer interview indicated attenuated levels of reliability.

The findings in the present study of limited inter-interviewer agreement in the identification of problem areas and items parallels the findings of studies which have examined the inter-rater agreement in the assignment of psychiatric diagnosis. In a highly critical review, Ennis and Litwack (1974) have contended that the reliability of psychiatric interviews is so poor as to make questionable the admissibility of a psychiatric diagnosis as testimony in legal procedings. This review indicated that typically the rate of agreement among

interviewers when only the major diagnostic divisions of psychosis, neurosis, and character disorder were used to classify patients was approximately 70 percent. Agreement across specific diagnostic categories was found to be much lower. The average percentage of agreement for specific diagnoses was 54 percent, ranging from 33 to 61 percent. Commenting on these findings, Ennis and Litwack also note that the majority of the studies reviewed were carried out under reliability-maximizing conditions. In actual practice the rate of agreement regarding particular diagnoses may be between 32 and 42 percent (Beck, Ward, Mendelson, Marks, and Erbough, 1962).

This latter point is especially salient to the present study which was designed to discern the potential reliability of the behavioral interview under optimal conditions. Specifically, interviewers were given detailed instructions which defined their tasks as "comprehensive behavioral interviews with the goal of identifying all of the problem behaviors of each client" (Appendix B). The interviewers were also aware of the fact that each interview was audiotaped and that each client was being interviewed by three other interviewers. The awareness of reliability assessment has been found to increase inter-observer agreement during behavioral observations (Taplin and Reid, 1973; Romanczyk, Kent, Diament and O'Leary, 1973).

Consequently, the 55 percent agreement on specific areas identified as problems and lower agreement found for specific items identified as problems may reflect an optimal rather than

an actual rate of agreement. In any case, it is doubtful whether such agreement rates would be considered adequate for either research or clinical purposes. Proponents of behavioral assessment have frequently criticized traditional assessment procedures on the basis of poor reliability. The results of the present study suggest that some self-criticism may be in order. Thus, although the goals of traditional and behavioral interviews may be different, diagnosis versus problem identification, both assessment procedures appear to suffer from a lack of reliability.

In the present study a number of factors were examined which have been traditionally considered to influence reliability. These factors were investigated with respect to how they affected both inter-interviewer agreement (variability among interviewers) and interviewer accuracy (error differences between human interview and computer interview criterion).

The potential effects of interviewer input differences on the reliability of interview data were analyzed both quantitatively and qualitatively in the present study. At a quantitative level, results of an analysis of variance indicated that interviewers differed with respect to the mean number of problem areas inquired about across clients. More qualitative analyses of the agreement between pairs of interviewers on specific areas and specific items questioned revealed substantial interviewer input differences. Agreement scores for areas questioned ranged between .33 and .87, with an average of .62 questioned per interview across interviewers and clients. In

those areas which were questioned by two or more interviewers, the mean agreement as to problem items questioned per area was .29, ranging between .00 and .74. Thus, interviewers differed in both the number of areas they asked a client about and also the specific areas and items they questioned. Finally, when human and computer interviewers were compared, it was found that on the average human interviewers sampled 55 percent of the potential problem areas and approximately 6 percent of the potential problem items included in the computer interview. Thus, interviewers were found to be less comprehensive in their questioning than the computer.

To summarize, in this study, interviewers were found to vary in the number of areas they questioned, the specific areas and specific items they questioned, and in their overall comprehensiveness in comparison to a standard (computer) interview. These results suggest that in the present study variations in the specific areas and specific items identified as problems between the four interviewers may have resulted in part from interviewers asking questions about different areas of the client's life. In addition, the low accuracy scores obtained by the human interviewers as compared to the computer criterion may have resulted in part from the lack of comprehensiveness in questioning clients about areas and items.

A number of factors can be identified which may affect the questions posed by an interviewer during a clinical interview. The interviewer's training, for example, can be conceptualized as a shaping process which supplies the interviewer with a

specific question pool or repertoire; the interviewer is taught what to look for. For different therapeutic schools or orientations "what to look for" may differ and consequently affect the interviewer's questioning behavior. Others (Goldfried and Pomeranz, 1968) have commented on a similar process in clinical treatment where the selection of specific treatment techniques seems to be mostly a function of therapist training. One can also speculate that areas of interest to a particular interviewer are more likely to be questioned in hopes of eliciting a positive response (Raines & Rohrer, 1960). Finally, interviewer-interviewee biases (i.e., demographic variables) have been found to affect differentially the questions posed during the interview process (e.g., Schwitzgebel & Kolb, 1975).

One remedy to reduce problems of input variability has evolved within the diagnostic assessment model. A number of structured psychiatric interviews have been developed which bring a high degree of standardization to the diagnostic interview (e.g., Spitzer, Endicott, Fleiss, & Cohen, 1967; 1970). Typically, these structured interview formats require interviewers to assess a patient's functioning in a wide variety of predetermined areas with a heavy emphasis on detailed inquiries into specific syndromes of psychopathology. In general, the use of these standardized psychiatric interviews has improved substantially the level of inter-rater agreement in the assignment of diagnoses based on interview information (Linehan, 1975; Helzer et al., 1977). These instruments, however, were not

specifically designed to gather the type of detailed information required in a comprehensive behavioral interview.

Within the behavioral assessment model some preliminary attempts have been made towards developing standardized assessment instruments. The majority of these assessment instruments are narrow-band, designed to supply information only for specific areas in which commonly employed treatment interventions are available (e.g., sexual dysfunction; LoPiccolo & Steger, 1974). Only two comprehensive behavioral coding systems have been compiled which include extensive arrays of potential problem areas and behaviors (Cautela & Upper, 1975; Hay & Hay, see Appendix D). With further refinement these coding systems may provide a partial remedy for the reliability problems that the present study has identified in the behavioral interview (O'Farrell and Upper, 1977). The development of broad-band standardized behavioral assessment procedures, however, is more difficult than the development of standardized diagnostic interviews. Standardized diagnostic interviews deal with a finite number of possible diagnoses, whereas the range of potential problem behaviors is infinite. This infinite range of potential problem behaviors is attributable to the fact that within the behavioral assessment model a behavior is considered problematic on the basis of its frequency of occurrence in a specific situation rather than on the basis of its topography (Ferster, Culbertson, & Boren, 1975).

The potential effects of interviewer output differences on the reliability of interview data were also analyzed in the present study. The usual procedure during interviews is for interviewers to make notes of salient client comments and subsequently write or dictate a summary of their information following the interview. The results of the present study indicate taht this procedure can result in substantial losses of interview information. On the average, interviewers reported 72 percent of the areas raters coded as problems from the transcriptions in the dictations of the information they had obtained during their interviews. Consequently, over 25 percent of the interview content was lost in dictation (Omission Error). Previous research comparing taped transcriptions to interviewer reports of client interview information supports the findings of the present study (e.g., Symonds & Dietrich, 1941). In these earlier studies information losses were found to increase as a function of time between the actual interview and the writing of the report. Although temporal factors were controlled in the present study (interviewers dictated studies immediately following each interview), substantial omissions of interview content occurred.

The actual distortion of client responses in interviewer summaries, documented in earlier reports such as that reported by Payne (1949) who found as much as 25% of the statements attributed to respondents to be clearly incorrect, was not found in the present study in relation to problem areas. The average commission error across interviewers was 5 percent:

problem areas attributed to the client and reported in the dictation that were not actually mentioned during the interview. Thus, although considerable losses in the quantity of interview content actually dictated were noted in the present study, but marked distortions of interview information did not occur. These results suggest that in the present study variations in the specific areas identified between the four interviewers may have resulted in part from differences in the client responses that interviewers chose to write down during an interview and subsequently dictated.

Some of the same variables that were postulated as underlying interviewer input differences may affect interviewer output. Interviewer training for example, may teach the interviewer not only what to look for but also what to hear, remember, and subsequently record. Interviewer biases have also been found to affect differentially subsequent reports of interview content (e.g., Schwitzgebel & Kolb, 1975). A study by Smith and Hyman (1950) demonstrated the potential impact of interviewer biases on interview reports. A "planted" respondent gave the same answers to the questions asked by a series of interviewers. The interviewers' reports of the answers, however, were found to be quite discrepant. In addition, the interviewer is involved in a number of competing behaviors during an actual interview: processing client input, recording "relevant" information and formulating additional questions. These competing behaviors probably interact with the interviewers' past learning history, training, (e.g., selective memory) and biases, to affect the interview information eventually reported.

The results of the present study suggest methods for decreasing problems of output variability. In the present study, interviewers dictated summaries immediately following each interview. The fact that this procedure did not reduce substantially the information lost, however, underscores the importance of audiotaping procedures. To minimize information loss, audiotape procedures are an absolute necessity for research purposes and probably should be employed more widely in clinical settings. The rate of commission errors in the present study was markedly lower than in earlier studies. One explanation for this result is that interviewers in the present study were aware that the author would have access to the actual interview transcriptions. This awareness may have resulted in the interviewers being more cautious with respect to the content of the interview dictations. In research or clinical settings, therefore, it may be a useful practice to employ overt reliability assessment procedures, perhaps on a randomized basis.

In addition to the effects of interviewer input and output factors, inter-interviewer agreement and interviewer accuracy may be affected by inconsistencies in the client's responses. In the present study clients were consistent; the average inter-interviewer agreement with respect to client responses was .86. When the consistency of client responses between the computer and human interviewers was determined, the average agreement was .89. In previous research the

extent to which variability in client responses has affected interview information ranged from approximately 5 percent to 75 percent (Ward, et al., 1961; Bancroft, 1940). The high rate of client consistency noted in the present study may be attributable to the clients' awareness of the purpose of the study as well as their knowledge that the interviews were being recorded. In more naturalistic settings, client responses may be more susceptible to shaping by differential interviewer feedback for certain answers (Krasner, 1967).

The standardization of interviewing procedures has been offered as a partial remedy for some of the factors which may affect interview reliability. The computerization of the standardized interview may represent a method of reducing the influence of these factors even further and thereby may improve the reliability of interview data. The standardization provided by computerized interviewing controls for differences in interviewer input by ensuring even and consistent coverage of potential problem areas for all clients. In addition interview recording differences (output) are eliminated by the computerization of interviewing procedures-- client responses are recorded immediately and verbatim. Besides allowing for the assessment of client information without the confounding influences of sources of interviewer variability, direct client-computer interaction should limit the effects of interviewer-interviewee biases on interview content. Standardized computer interviews have the additional advantage of ensuring the comparability of interview procedures when repeated assessment of clients is desired.

Such reassessment could be utilized to monitor improvements in functioning due to treatments or to signal impending crises. The consistency of interview procedures provided by standarized computer interviews would also be extremely useful in studies such as the present study in which comparisons between clients are necessary.

While computerization of standardized interviews has a number of advantages, there are certain limitations in its utility as an assessment instrument that should be recognized. One obvious limitation is that not everyone can interact with computerized interviewing procedures. Sightless, illiterate, acutely psychotic clients, or clients with organic impairments are not usually appropriate for interactive computerized procedures. An additional limitation is the fact that the computer programs are constructed by humans. Consequently, structural and substantive biases may be written into standardized computer interviews. These biases, however, would not be expected to affect differentially the responses of individual clients. Finally, the computer interview lacks the flexibility of human interviewing procedures. The computer's total repertoire of questions is predetermined and it does not have the human interviewer's inherent ability to spontaneously branch and follow up important client responses. Likewise, the computer cannot benefit from the nonverbal communication of the client: the appearance of tears, increased symptoms of anxiety, or other affect changes which may accompany questioning in certain areas.

Although it is important for researchers and clinicians to be cognizant of the limitations of computerized interviewing procedures, these limitations do not denigrate the potential impact that automated interviewing procedures may eventually have on the assessment process. These limitations emphasize the fact that computer procedures should not be viewed as a replacement for the human interviewer or other sources of assessment information (e.g., standardized behavioral observations). Rather the computer interview represents a valuable assessment tool which should be used as an adjunct to other assessment procedures. The positive responses of the clients in the present study to the computer interview further supportthe computer's potential as an assessment procedure.

To summarize, the results of the present study have indicated low inter-interviewer agreement and accuracy in the identification of specific problem areas. Three factors, interviewer input and output differences and the consistency of client responses, were examined to determine their potential effects on interview content. The results of these examinations suggested that interviewer input and output differences may be implicated in the variations in problem identification found in the present study. The standardization of interview procedures was offered as one remedy for the reliability problems found in the present study. In addition, the computerization of standardized interviewing procedures was presented as a tool for increasing the potency of this remedy.

In generalizing from the results of the present study, it is important to consider the limitations of the data base from which these results were drawn. The sample interviewers employed in the present study was small and homogeneous with respect to graduate training in clinical psychology. To increase confidence in the results of the present study, the study should be replicated using other interviewers, clients, and clinical settings. Research should be done to test experimentally the methods of improving reliability outlined in the present study. Furthermore, the present study did not take into account the relative importance of the problems identified by the interviewers. Ratings of problem severity should be included in future research studies investigating the reliability of problem identification. Subsequent research should also examine the relationship between problem identification and treatment outcome. Although Lazarus (1973) has suggested that "faulty problem identification (inadequate assessment) is probably the greatest impediment to successful therapy," research is necessary to substantiate this relationship.

BIBLIOGRAPHY

Ash, P. The reliability of psychiatric diagnosis. Journal of Abnormal and Social Psychology, 1949, 44, 272-277.

Athey, K.R., Coleman, J.E., Reitman, A.P., & Tang, J. Two experiments showing the effect of the interviewer's racial background on responses to questionnaires concerning racial issues. Journal of Applied Psychology, 1960, 44, 244-246.

Baer, D.M. Perhaps it would be better not to know everything. Journal of Applied Behavior Analysis, 1977, 1, 91-97.

Bancroft, G. Consistency of information from records and interviews. Journal of American Statistical Association, 1940, 35, 377-381.

Bandura, A. Principles of behavior modification. New York: Holt, Rinehart and Winston, 1969.

Bannister, D., Salmon, P., & Lieberman, D.M. Diagnosis-treatment relationships in psychiatry: A statistical analysis. British Journal of Psychiatry, 1964, 110, 726-732.

Beck, A.T., Ward, C.H., Mendelson, M., Mock, J.E., & Erbough, J.K. Reliability of psychiatric diagnosis: A study of consistency of clinical judgements and ratings. American Journal of Psychiatry, 1962, 351, 352-355.

Benney, M., Riesman, D., & Star, S. Age and sex in the interview. American Journal of Sociology, 1956, 62, 143-152.

Braginsky, B.M., & Braginsky, D.D. Schizophrenic patients in the psychiatric interview: An experimental study of their effectiveness at manipulation. Journal of Consulting Psychology, 1967, 31, 543-547.

Cantril, H. Gauging public opinion. Princeton, N.J.: Princeton, 1944.

Cautela, J.R., & Upper, D. The behavioral inventory battery: The use and self-report measures in behavioral analysis and therapy. In M. Hersen and A.S. Bellack (Eds.), Behavioral assessment: A practical handbook. New York: Pergamon Press, 1976.

Cautela, J.R., & Upper, D. The process of individual behavior therapy. In M. Hersen, R.M. Eisler, & P.M. Miller (Eds.). Progress in behavior modification. New York: Academic Press, 1975.

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

Corner, B.J. Studies in phonographic recordings of verbal material. I. The use of phonographic recordings in counseling practice and research. Journal of Consulting Psychology, 1942, 6, 105-113.

Cozby, P.C. Self-disclosure, reciprocity and liking. Psychological Bulletin, 1973, 79, 73-91.

Cronbach, L.J. Essentials of psychological testing. New York: Harper & Row, 1970.

Cronbach, L.J., & Gleser, G.C. Psychological tests and personal decisions (2nd Edition). Urbana: University of Illinois Press, 1965.

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

Dibner, A.S. Ambiguity and anxiety. Journal of Abnormal and Social Psychology, 1958, 56, 165-174.

Dickson, C.R. Role of assessment in behavior therapy. In P. McReynolds (Ed.), Advances in psychological assessment. San Francisco: Jossey-Bass, 1975.

Ekehammer, B. Interactionism in personality from a historical perspective. Psychological Bulletin, 1974, 81, 1026-1048.

Ennis, B.J., & Litwack, T.R. Psychiatry and the presumption of expertise: Flipping coins in the courtroom. California Law Review, 1974, 62, 693-752.

Eysenck, H.J. The scientific study of personality. London: Routledge, 1952.

Ferster, C.B., Culbertson, S., & Boren, M.C.P. Behavior principles. Englewood Cliffs, N.J.: Prentice-Hall Inc., 1975.

Frank, G. Psychiatric diagnosis: A review of research. Oxford: Pergamon Press, 1975.

Gambrill, E.D., & Richey, C.A. An assertion inventory for use in assessment and research. Behavior Therapy, 1975, 6, 550-561.

Goldfried, M.R., & Kent, R.M. Traditional versus behavioral personality assessment: A comparison of methodological and theoretical assumptions. Psychological Bulletin, 1972, 77, 409-420.

Goldfried, M.R., & Pomeranz, D.M. Role of assessment in behavior modification. Psychological Reports, 1968, 23, 75-87.

Goldfried, M.R., & Sprafkin, J.N. Behavioral personality assessment. Morristown, N.J.: General Learning Press, 1974.

Guest, L. A study of interviewer competence. International Journal of Opinion Attitude Research, 1947, 1, 17-30.

Hayes-Roth, F., Longabaugh, R., & Ryback, R. Mental health: Systems and nonsystems. British Journal of Medical Psychology, 1973, 46, 317-340.

Heller, K. Laboratory interview research as an analogue to treatment. In A.E. Bergin and S.L. Garfield (Eds.), Handbook of psychotherapy and behavior changes. New York: John Wiley and Sons, 1971.

Helzer, J.E., Robins, L.N., Taiblison, M., Woodruff, R.A., Reich, T., & Wish, E.D. Reliability of psychiatric diagnosis. Archives of General Psychiatry, 1977, 34, 129-141.

Hyman, H. Do they tell the truth? Public Opinion Quarterly, 1944, 8, 557-559.

Johnson, S.M., & Bolstad, O.D. Methodological issues in naturalistic observation: Some problems and solutions for field research. In L.A. Hamerlynck, L.C. Handy, & E.J. Mash (Eds.), Behavior change: methodology, concepts and practice. Champaign, Illinois: Research Press, 1973.

Jones, R.R., Reid, J.B., & Patterson, G.R. Naturalistic observation in clinical assessment. In P. McReynolds (Ed.), Advances in psychological assessment. (Vol. 3.), San Francisco: Jossey-Bass, 1975.

Kahn, R.L. and Cannel, C.F. The dynamics of interviewing: theory, technique and cases. New York: Wiley, 1957.

Kanfer, F.H., & Grimm, L.G. Behavioral analysis: selecting target behaviors in the interview. Behavior Modification, 1977, 1, 7-28.

Kanfer, F.H., & Phillips, J.S. Learning foundations of behavior therapy. New York: John Wiley and Sons, 1970.

Kanfer, F.H., & Saslow, G. Behavioral diagnosis. In C.M. Franks (Ed.), Behavior therapy: appraisal and status. New York: McGraw Hill, 1969.

Katz, D. Do interviewers bias poll results? Public Opinion Quarterly, 1942, 6, 248-268.

Krasner, L. The therapist as a reinforcement machine. In Berenson & Carphoff (Eds.), Sources of gain in counseling and psychotherapy. New York: Holt, 1967.

Lang, P.J. Fear reduction and fear behavior: Problems in treating a construct. Research in Psychotherapy, 1968, 3, 90-102.

Lazarus, A.A. Multimodal behavior therapy: Treating the "Basic ID". Journal of Nervous and Mental Disease, 1973, 156, 404-411.

Lenski, G.E., & Leggett, J.C. Caste, class and difference in the research interview. American Journal of Sociology, 1960, 65, 463-467.

Leonard, H.L., & Bernstein, A. The anatomy of psychotherapy: Systems of communication and expectation. New York: Columbia University Press, 1960.

Lehan, M.M. Basic issues in behavioral interviewing. In J.D. Cone, & R.P. Hawkins (Eds.), Behavioral assessment: New directions in clinical psychology. New York: Brunner-Mazel, 1977, in press.

LoPiccolo, J., & Steger, J.C. The sexual interaction inventory: A new instrument for assessment of sexual dysfunctions. Archives of Sexual Behavior, 1974, 3, 585-595.

McLean, P.D., & Miles, J.E. Evaluation and the problem-oriented record in psychiatry. Archives of General Psychiatry, 1974, 31, 622-625.

McPartland, T.S., & Richart, R.H. Social and Clinical outcomes of psychiatric treatment. Archives of General Psychiatry, 1966, 14, 179-184.

Maccoby, E., & Maccoby, N. The interview: A tool of social science. In G. Lindzey (Eds.), Handbook of social psychology, Vol. 1. Cambridge, Mass.: Addison-Wesley, 1954.

Mahoney, M.J. Cognition and behavior modification. Cambridge, Massachusetts: Ballinger, 1974.

Mash, E.J., & Terdal, L.G. Behavior therapy assessment: Diagnosis, design and evaluation. Psychological Reports, 1974, 35, 587-601.

Matarazzo, J.D., Wiens, A.M., & Saslow, G. Studies in interview speech behavior. In L. Krasner and L.P. Ullman (Eds.), Research in behavior modification. New York: Holt, 1965.

Meehl, P.E. The cognitive activity of the clinician. American Psychologist, 1960, 15, 19-27.

Morganstern, K.P. Behavioral interviewing: The initial stages of assessment. In M. Hersen and A. Bellack (Eds.), Behavioral assessment: A practical handbook. New York: Pergamon, 1976.

Nathan, P.E. Cues, decisions, and diagnosis: A systems-analytic approach to the diagnosis of psychopathology. New York: Academic Press, 1967.

Nelson, R.O., Hay, L.R., & Hay, W.M. Comments on Cone's "The relevance of reliability and validity for behavioral assessment." Behavior Therapy, 1977, 8 427-430.

O'Farrell, T.J., & Upper, D. The interjudge reliability of Cautela and Upper's behavioral Coding system. Journal of Behavior Therapy and Experimental Psychiatry, 1977, 8, 39-43.

Payne, S.L. Interviewer memory faults. Public Opinion Quarterly, 1949, 13, 684-685.

Peterson, D.R. The clinical study of social behavior. New York: Appleton-Century-Crofts, 1968.

Pope, B., Nudler, S., Vonborff, M.R., & McGee, J.P. The experienced and professional interviewer versus the complete novice. Journal of Consulting and Clinical Psychology, 1974, 42, 680-690.

Pope, B., & Siegman, A.W. Relationship and verbal behavior in the initial interview. In A. Siegman & B. Pope (Eds.), Studies in Dyadic Communication. New York: Pergamon Press, 1972.

Raines, G.N., & Rohrer, S.H. The operational matrix of psychiatric practice, II. Variability in psychiatric impressions and the projection hypothesis. American Journal of Psychiatry, 1960, 117, 133-139.

Riesman, D., & Ehrlich, J. Age and authority in the interview. Public Opinion Quarterly, 1961, 25, 39-56.

Romanczyk, R.G., Kurt, R.M., Diament, C., & O'Leary, K.D. Measuring the reliability of observational data: A reactive process. Journal of Applied Behavior Analysis, 1973, 1, 175-184.

Ryback, R.S. The problem oriented record in psychiatry and mental health care. New York: Grune and Stratton, 1975.

Ryback, R.S., & Gardner, J.S. Problem formulation: The problem-oriented record. American Journal of Psychiatry, 1973, 130, 312-316.

Schmidt, H.O., & Fonda, C.P. The reliability of psychiatric diagnosis: A new look. Journal of Abnormal and Social Psychology, 1956, 52, 262-267.

Schwitzgebel, R.K., & Kolb, D.A. Changing human behavior. New York: McGraw-Hill, 1975.

Sidman, M. Tactics of scientific research. New York: Basic Books Inc., 1960.

Smith, H.L., & Hyman, H. The biasing effect of interviewer expectations on survey results. Public Opinion Quarterly, 1950, 14, 491-506.

Spitzer, R.L., Endicott, J., Fleiss, J.L., & Cohen, J. The psychiatric status schedule. Archives of General Psychiatry, 1970, 23, 41-55.

Spitzer, R.L., Fleiss, J.L., Endicott, J. & Cohen, J. Mental status schedule. Archives of General Psychiatry, 1967, 16, 479-493.

Stuart, R.B. Trick or treatment. Champaign, Illinois: Research Press, 1970.

Stuart, R.B., & Stuart, F. Marital precounseling inventory. Champaign, Illinois: Research Press, 1975.

Symonds, P.M., & Dietrich, D.H. The effect of variations in the time interval between an interview and its recording. Journal of Abnormal Social Psychology, 1941, 36, 593-598.

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

Thomas, E.J., & Walter, C.L. Guidelines for behavioral practice in the open community agency: Procedure and evaluation. Behavior Research and Therapy, 1973, 11, 193-205.

Ward, C.H., Beck, A.T., Mendelson, M., Mock, J.E., & Erbough, J.K. The psychiatric nomenclature: Reasons for diagnostic disagreement. Archives of General Psychiatry, 1962, 7, 60-67.

Wiggins, J.S. Personality and prediction: Principles of personality assessment. Massachusetts: Addison-Wesley Publishing Company, 1973.

Witenborg, J., Holzberg, J., & Simon, B. Symptom correlates for descriptive diagnosis. Genetic Psychology Monographs, 1953, 47, 237-301.

Wolpe, J. The practice of behavior therapy. New York: Pergamon Press, 1969.

Ziegler, E., & Phillips, L. Psychiatric diagnosis and symptomology. Journal of Abnormal and Social Psychology, 1961, 63, 69-75.

APPENDIX A

LIFE AREA PROBLEMS

In a particular life area you may be experiencing problems or difficulties which are more than normal or typical. Without identifying the specific stress or problems, we would like you to examine different categories of these life areas and select one of the responses by circling the number.

What is the extent of your problems in the area of:

MARRIAGE

RELATIVES

RAISING CHILDREN

EMPLOYMENT

PHOBIA - FEAR

SEXUAL MATTERS

MEDICATION - DRUGS

HOUSING OR LIVING ARRANGEMENTS

MONEY

ASSERTIVE (SHYNESS - TIMID)

LEGAL

ANGER - HOSTILITY

FRIENDSHIP - SOCIALABILITY

ALCOHOL

COMPULSIVENESS

LONELINESS - LITTLE ACTIVITY

RELIGION

MEDICAL

DEPRESSION

APPEARANCE

PAIN

Rating Scale:

1. No difficulties
2. Minimum difficulties
3. A few
4. Some difficulties
5. Very many difficulties

APPENDIX B

BEHAVIORAL INTERVIEWING INSTRUCTIONS

The purpose of assessment in behavior therapy is to identify target behaviors for modification and design appropriate treatment plans. You will be asked to conduct four comprehensive behavioral interviews with the goal of identifying all of the problem behaviors of each client. The following excerpt from Kanfer and Saslow "Behavioral Diagnosis" should be used as a guide to assist you in directing the interviews. Kanfer and Saslow suggest that the interviewer look for the behavioral excesses and deficits of the client with regard to their eventual place in the treatment procedures.

Behavioral Excesses: A class of related behaviors occurs and is described as problematic by the patient or an informant because of excess in 1) frequency 2) intensity 3) duration or 4) occurrence under conditions when its socially sanctioned frequency approaches zero. Compulsive handwashing, combativeness, prolonged excitement, and sexual exhibitionism are examples of behavioral excesses along one or another of these four dimensions. Less obvious, because they often do not constitute the major presenting complaint and appear only in the course of the behavioral analysis are examples of socially unacceptable solitary, affectionate, or other private behaviors. For instance, a housewife showing

APPENDIX B (Continued)

excessive solitary preoccupation can do so by excessive homemaking activities, 1) several hours a day 2) seven days weekly for most of the waking day 3) to the extent that phone calls or doorbells are unanswered and family needs are unattended. From this example, it is clear that both duration and intensity values of the behavior may jointly determine the characterization of the behavior as excessive.

Behavioral Deficits: A class of responses is described as problematic by someone because it fails to occur 1) with sufficient frequency 2) with adequate intensity 3) in appropriate form or 4) under socially expected conditions. Examples are: reduced social responsiveness (withdrawal), amnesias, fatigue syndromes, and restrictions in sexual or somatic function (e.g., impotence, writer's cramp). Other examples of behavioral deficits can be found in depressed patients who have no appropriate behavior in a new social environment, e.g., after changes from a rural to an urban area, from marital to single status, or from one socioeconomic level to another. "Inadequate" persons often are also found to have large gaps in their social or intellectual repertoires which prevent appropriate actions.

You will be allowed as much time as you feel is necessary to collect this information from the client. Feel free to take notes during the interviews: you will be asked to dictate the information that you have obtained immediately following each interview.

APPENDIX C

SURVEY QUESTIONNAIRE

We would like to get your impression of the human and computer interviews. Please complete the following items by circling the appropriate numbers.

Human Interviewer

Empathy: Empathy is the ability to perceive accurately what another person is experiencing. How empathetic to your problems was the person who interviewed you?

1	2	3	4	5
not empathetic		moderately empathetic		extremely empathetic

Genuineness: Genuiness is the ability of an individual to be freely and deeply himself. It is nonphoniness, nondefensiveness. How genuine was the interviewer in your opinion?

1	2	3	4	5
not genuine		moderately genuine		very genuine

Warmth: Warmth is evidenced by positive comments of concern and affection for the client, and by smiles and other nonverbal gestures of appreciation, including touching. How warm did the interviewer seem to you?

1	2	3	4	5
not warm		moderately warm		very warm

APPENDIX C (Continued)

Computer Interview

How would you rate your experience of participating in the computer interview?

1. Strong dislike
2. Moderate dislike
3. Indifferent
4. Moderate like
5. Strong like

Did you prefer being asked questions by the human interviewers or the computer?

1. Much preferred the human interview
2. Somewhat preferred the human interview
3. Indifferent
4. Somewhat preferred the computer
5. Much preferred the computer

In the interviews, did you prefer that personal and private questions be asked by the human interviewers or the computer?

1. Strongly preferred human interviewer
2. Somewhat preferred human interviewer
3. Either one
4. Somewhat preferred computer
5. Strongly preferred computer

APPENDIX D

EXAMPLE SECTION FROM BEHAVIORAL CLASSIFICATION SYSTEM APPEARANCE PROBLEMS

AP-1	Inadequate hygiene-Behaviors involving personal cleanliness which occur at such a low frequency as to result in detriment to health and/or disapproval from others.								
	AP-1a - Not washing or bathing regularly								
	AP-1b - Inadequate brushing of teeth								
	AP-1c - Inadequate cleaning or cutting of nails								
	AP-1d - Not using deodorant when body is offensive								
	AP-1e - Inadequate hygiene skills								
	AP-1f - Other (specify)								
AP-2	Inappropriate dress - Behaviors involving personal dress which may result in detriment to health and/or disapproval from others (e.g., nudity in cold weather).								
	AP-2a - Nudity								
	AP-2b - Inappropriate clothes for weather or for situation								
	AP-2c - Repeatedly wearing soiled clothes								
	AP-2d - Wearing excessive clothing								
	AP-2e - Wearing poorly fitting clothes								
	AP-2f - Other (specify)								

APPENDIX D (Continued)

AP-3 Concerns about appearance

AP-3a - Height								
AP-3b - Weight								
AP-3c - Age								
AP-3d - Dress								
AP-3e - Hygiene								
AP-3f - Body Shape								
AP-3g - Other (Specify)								

APPENDIX E

TABLE 1

Client Demographics

Client	Age	Sex	Marital Status	Education	Occupation	Referral Source
1	27	F	Married	2 years College	Unemployed Nurse	Staff Psychiatrist
2	34	F	Married	8th grade	Unemployed Secretary	Staff Psychiatrist
3	21	F	Married	high school	Unemployed Waitress	Outpatient Psych. Clinic
4	29	F	Married	high school	Farmer/ Housewife	Staff Psychiatrist

TABLE 2

Human Interview Durations (Minutes)

	Interviewer			
Client	1	2	3	4
1	176	125	125	166
2	103	141	133	163
3	60	86	89	77
4	110	100	90	95

TABLE 3

Generalizability-Number of Problem Areas Identified: Interviewers (4) x Clients (4) Repeated Measures Analysis of Variance on the Number of Areas Identified as a Problem For Each Client From Coded Interview Transcriptions

Source	df	MS	F
Clients	3	29.23	4.172*
Interviewers	3	5.895	0.841
Clients x Interviewers	9	7.007	

*$p < .05$

TABLE 4

Generalizability-Number of Problem Areas Identified: Interviewers (4) x Clients (4) Repeated Measures Analysis of Variance On The Number of Areas Identified as a Problem For Each Client From Coded Interview Dictations

Source	df	MS	F
Clients	3	24.89	10.64**
Interviewers	3	5.062	2.163
Clients x Interviewers	9	2.340	

**$p < .01$

TABLE 5

Inter-Interviewer Agreement for Specific Areas Identified as Problems: Transcriptions

	Interviewer Combinations					
Clients	1+2	1+3	1+4	2+3	2+4	3+4
1	.73	.75	.59	.71	.53	.50
2	.53	.47	.59	.47	.76	.74
3	.50	.38	.55	.57	.50	.50
4	.25	.40	.38	.54	.64	.55

Mean Inter-Interviewer Agreement for Specific Areas Identified as Problems from Coded Transcription Data (24 comparisons) = .55.

TABLE 6

Inter-Interviewer Agreement for Specific Areas Identified as Problems: Dictations

	Interviewer Combinations					
Clients	1+2	1+3	1+4	2+3	2+4	3+4
1	.70	.57	.38	.50	.33	.50
2	.43	.69	.60	.50	.50	.69
3	.42	.33	.42	.44	.80	.33
4	.33	.44	.38	.22	.33	.63

Mean Inter-Interviewer Agreement for Specific Areas Identified as Problems from Coded Dictation Data (24 comparisons) = .48.

TABLE 7

Mean Inter-Interviewer Agreement Scores Across Interviewers and Clients for Items Identified as Problems by Problem Area

Problem Area	
Addictions	.78
Appearance Problems	.13
Assertion-Anger	.27
Child Rearing	.22
Eating	.34
Emotional Behavior	.62
Employment	.41
Fears	.31
Imagery	1.00
Intellectual Performance	- *
Legal Problems-Anti-Social Behavior	- *
Marriage	.24
Money and Finances	.41
Obsessive Behaviors: Repetitive Tasks	.37
Obsessive Behaviors: Thoughts	.32
Organic Impairments Influenced by Psychological Factors	.38
Relatives (Family Relationship)	.34
Religion	.60
Self-Injurious Behavior	.67
Sex	.68
Sleep	.24
Social Interactions	.15
Socially Inappropriate Behaviors	- *
Tension Problems	.10
Verbal Behavior (Speech)	.17
Average Inter-interviewer agreement on items identified as problems per problem area across interviewers and clients	.40

*Areas in which it was not possible to calculate inter-interviewer agreement for problem items because only one or less of the interviewers identified the area as a problem for a given client.

TABLE 8

Interviewer Accuracy-Comparisons Between Human Interviewers And A Computer Criterion: Interviewers (4 Human-1 Computer) x Clients (four) Repeated Measures Analysis of Variance On The Number of Areas Identified As A Problem For Each Client From Coded Interview Transcription And Computer Interview Data

Source	df	MS	F
Clients	3	29.53	4.609*
Interviewers	4	13.88	2.165
Clients x Interviewers	12	6.408	

*$p < .05$

TABLE 9

Interviewer Accuracy-Comparisons Between Human Interviewers And Computer Criterion: Interviewers (4 Human-1 Computer) x Clients (four) Repeated Measures Analysis of Variance On The Number of Areas Identified As a Problem For Each Client From Coded Interview Dictation And Computer Interview Data

Source	df	MS	F
Clients	3	28.18	11.39**
Interviewers	4	33.07	13.36**
Clients x Interviewers	12	2.474	

**$p < .01$

TABLE 10

Agreement Scores Between the Computer and Transcribed Human Interviews for Specific Areas Identified as Problems

	Human Interviewer			
Client	1	2	3	4
1	.69	.56	.59	.61
2	.47	.56	.53	.68
3	.50	.57	.53	.64
4	.27	.63	.56	.58

Mean Agreement for Specific Areas Identified as Problems Between Coded Transcription and Computer Interview Data (16 comparisons) = .55.

TABLE 11

Mean Agreement Scores Between the Computer and Human Interviewer for Items Identified as Problems by Problem Area

Problem Area	
Addictions	.34
Appearance Problems	.19
Assertion-Anger	.18
Child Rearing	.25
Eating	.31
Emotional Behavior	.54
Employment	.46
Fears	.17
Imagery	- *
Intellectual Performance	- *
Legal Problems-Anti-Social Behavior	- *
Argue-Marriage	.18
Money and Finances	.25
Obsessive Behaviors: Repetitive Tasks	- *
Obsessive Behaviors: Thoughts	.23
Organic Impairments Influenced by Psychological Factors	.28
Relatives (Family Relationship)	.11
Religion	.07
Self-Injurious behavior	.25
Sex	.52
Sleep	.14
Social Interactions	.18
Socially Inappropriate Behaviors	- *
Tension Problems	.09
Verbal Behavior (Speech)	.08
Average AGreement for Items Identified as Problems per Problem Area Across Interviewers and Clients	.24

*Areas in which it was not possible to calculate agreement for problem items because the area was not identified as a problem for a given client by the computer and at least one interviewer.

TABLE 12

Input Variance - Comparisons Between Human Interviewers: Interviewers (4) x Clients (4) Repeated Measures Analysis of Variance On the Number of Areas Questioned For Each Client From Coded Interview Transcriptions

Source	df	MS	F
Clients	3	11.75	3.61
Interviewers	3	13.08	4.02*
Clients x Interviewers	9	3.25	

*p .05

TABLE 13

Inter-Interviewer Agreement for Problem Areas Questioned from Coded Interview Transcriptions

	Interviewer Combinations					
Clients	1+2	1+3	1+4	2+3	2+4	3+4
1	.73	.50	.68	.59	.68	.65
2	.47	.67	.60	.63	.53	.67
3	.69	.59	.59	.65	.87	.65
4	.33	.44	.62	.67	.64	.63

Mean Inter-Interviewer Agreement for Specific Areas Questioned from Coded Transcription Data (24 comparisons) = .62.

TABLE 14

Mean Inter-Interviewer Agreement Scores Across Interviewers and Clients for Specific Items Questioned by Problem Area

Problem Area	
Addictions	.37
Appearance Problems	.18
Assertion-Anger	.22
Child Rearing	.50
Eating	.19
Emotional Behavior	.39
Employment	.56
Fears	.10
Imagery	.83
Intellectual Performance	- *
Legal Problems-Anti-Social Behavior	- *
Marriage	.15
Money and Finances	.25
Obsessive Behaviors: Repetitive Tasks	.30
Obsessive Behaviors: Thoughts	.28
Organic Impairments Influenced by Psychological Factors	.30
Relatives (Family Relationship)	.15
Religion	.13
Self-Injurious Behavior	.74
Sex	.24
Sleep	.13
Social Interactions	.20
Socially Inappropriate Behaviors	- *
Tension Problems	.17
Verbal Behavior (Speech)	.00
Average Inter-Interviewer Agreement on Specific Items Questioned per Problem Area Across Interviewers and Clients	.29

*Areas in which it was not possible to calculate inter-interviewer agreement for problem items because only one or less of the interviewers questioned the area for a given client.

TABLE 15

Percent of 25 Potential Problem Areas Questioned by Each Interviewer for Each Client

	Interviewer			
Client	1	2	3	4
1	52	52	56	76
2	56	44	64	72
3	52	56	56	56
4	36	44	56	48

Mean Percent of 25 Potential Problem Areas Questioned Across Interviewers and Clients (16 comparisons) = .55.

TABLE 16

Percent of Potential Computer Items Questioned
by Each Interviewer Across Clients

Problem Area	Interviewer 1	2	3	4	Mean Percent	Number of Potential Items Per Area
Addictions	4	4	11	13	8	19
Appearance Problems	-*	1	2	2	1.7	21
Assertion-Anger	5	4	4	6	4.8	94
Child Rearing	-	-	1	2	1.5	12
Eating	2	5	15	8	7.8	10
Emotional Behavior	9	8	10	11	9.5	20
Employment	5	4	3	6	4.5	20
Fears	-	2	7	14	7.7	51
Imagery	2	2	6	2	3.0	13
Intellectual Performance	-	-	-	-	-	15
Legal Problems-Anti Social Behavior	-	-	-	-	-	13
Marriage	11	6	8	16	10.3	80
Money and Finances	2	1	-	2	1.7	21

TABLE 16 (Continued)

Problem Area	Interviewer 1	2	3	4	Mean Percent	Number of Potention Items Per Area
Obsessive Behaviors: Repetitive Tasks	4	2	13	4	5.8	13
Obsessive Behaviors: Thoughts	6	9	15	24	13.5	25
Organic Impairments Influenced by Psychological Factors	7	4	7	7	6.3	47
Relatives (Family Relation-Ships)	12	6	8	11	9.3	27
Religion	-	-	10	3	6.5	17
Self-Injurious Behavior	2	1	1	5	2.3	21
Sex	2	2	4	6	3.5	85
Sleep	3	3	6	14	6.5	20
Social Interactions	11	4	23	7	11.3	21
Socially Inappropriate Behaviors	-	-	-	-	-	10
Tension Problems	8	-	3	-	5.5	18
Verbal Behavior (Speech)	-	1	2	2	1.7	30
Average Percent of Items Questioned Per Area Across Interviewers and Clients					6.03	

*Areas in which the interviewer did not ask questions.

TABLE 17

Agreement Scores for Consistency of Client Responses to Specific Items Questioned By Human Interviewer Pairs

Clients	I_1+I_2	I_1+I_3	I_1+I_4	I_2+I_3	I_2+I_4	I_3+I_4	Mean Agreement Across Interviewer Pairs
1	.89	.69	.84	.92	.89	.67	.82
2	.75	.76	.71	.78	.63	.90	.76
3	.94	.75	.92	1.00	.92	1.00	.93
4	.84	1.00	.75	1.00	1.00	.84	.91

Average Agreement for the Consistency of Client Responses Across Clients and Interviewer Pairs = .86.

TABLE 18

Agreement Scores for Consistency of Client Responses to Questions Asked in Both Computer and Human Interviews

	Interviewers				
Client	1	2	3	4	Mean Agreement
1	.82	.91	1.00	1.00	.93
2	.79	.83	.70	.86	.84
3	.86	.86	.85	.86	.86
4	1.00	.83	.91	.96	.93

Average Agreement Score for the Consistency of Client Responses Across Clients and Interviewers to Questions Asked in Both the Human and Computer Interviews = .89.

TABLE 19

Percent of Areas Identified as Problems in Interview Transcriptions and Not Reported in Dictations

	Interviewers				
Clients	1	2	3	4	Mean Percent of Areas Identified as Problems and Reported in Dictations across Interviewers
1	27	36	31	36	32
2	0	39	25	24	22
3	14	55	26	43	35
4	0	46	30	14	22

Average Percent of Areas Identified as Problems across clients and Interviewers in Interview Transcriptions and Not Reported in Dictations = 28.

TABLE 20

Percent of Problem Areas Identified in Dictations That Were Not Also Identified in Transcriptions

Client	Interviewers 1	2	3	4	Mean Percent of Areas Identified as Problems in Dictations Not Identified in transcriptions
1	0	0	10	0	2
2	23	11	10	0	11
3	0	0	0	0	0
4	33	0	0	0	8

Average Percent of Areas Identified as Problems in Dictations and Not Identified in Interview Transcriptions Across Interviewers and Clients = 5.