HANDBOOK OF PSYCHOLOGICAL ASSESSMENT

THIRD EDITION

EDITED BY
GERALD GOLDSTEIN
MICHEL HERSEN

HANDBOOK OF PSYCHOLOGICAL ASSESSMENT

Recent Titles of Related Interest

Bellack/Hersen COMPREHENSIVE CLINICAL PSYCHOLOGY

Rachman THE BEST OF BEHAVIOUR RESEARCH AND THERAPY

Sanavio BEHAVIOR AND COGNITIVE THERAPY TODAY

Related Journals

(Free sample copies available upon request)

ADDICTIVE BEHAVIORS
AGGRESSION AND VIOLENT BEHAVIOR
BEHAVIOUR RESEARCH & THERAPY
CLINICAL PSYCHOLOGY REVIEW
JOURNAL OF ANXIETY DISORDERS
JOURNAL OF BEHAVIOR THERAPY AND
EXPERIMENTAL PSYCHIATRY
PERSONALITY AND INDIVIDUAL DIFFERENCES

HANDBOOK OF PSYCHOLOGICAL ASSESSMENT

Third Edition

Edited by
GERALD GOLDSTEIN
VA Pittsburgh Healthcare System

MICHEL HERSEN
School of Professional Psychology, Pacific University

PERGAMON
An imprint of Elsevier Science
Amsterdam • Lousanne • New York • Oxford • Shannon • Singapore • Tokyo

ELSEVIER SCIENCE Ltd
The Boulevard, Langford Lane
Kidlington, Oxford OX5 1GB, UK

3rd edition 2000

Library of Congress Cataloging in Publication Data

Handbook of psychological assessment / edited by Gerald Goldstein, Michele Hersen.--
3rd ed.
p.cm.
Includes bibliographical references and indexes.
ISBN 0-08-043645-5 (hardcover)
1. Psychometrics. 2. Psychological tests. I. Goldstein, Gerald, 1931-II. Hersen, Michel.

BF39 .H2645 2000
150'.28'7--dc21

00-051936

British Library Cataloguing in Publication Data
A catalogue record from the British Library has been applied for.

CONTENTS

Preface

Since the publication of the second edition of this *Handbook* in 1990 there have been many new developments in psychological assessment. The new version of the Wechsler intelligence scales (WAIS-III) was published recently. It is described in this book by its developers. Neuropsychological assessment research continues to appear at a rapid pace, and that growth is expressed in entirely new chapters in the areas of pediatric and specialized neuropsychological assessment. The area of assessment of the elderly has grown remarkably rapidly since the appearance of the second edition. We now include a chapter on this very important aspect of psychological assessment.

We have continued the practice of inviting new authors to write some of the chapters in order to provide different perspectives and theoretical frameworks from authorities in their areas. The authors who wrote chapters previously have revised their work in a manner that reflects the significant developments in their specialties over the past decade. We attempted as much as possible to preserve the book as a basic reference work, while also providing current information.

The editors would like to thank the new authors for offering their new perspectives and philosophies of assessment, and the authors who wrote chapters previously for their concientious and detailed updates. The senior author acknowledges the support of the Department of Veterans Affairs in the preparation of this work, and Allison Beers for significant assistance with the manuscript. Dr. Hersen acknolwedges the support of his Editorial Assistant, Carole Louderée, and of Maura Sullivan and Erika Qualls.

Gerald Goldstein
Pittsburgh, PA
Michel Hersen
Forest Grove, OR

This Page Intentionally Left Blank

PART I

INTRODUCTION

This Page Intentionally Left Blank

CHAPTER 1

HISTORICAL PERSPECTIVES

Gerald Goldstein
Michel Hersen

INTRODUCTION

"A test is a systematic procedure for comparing the behavior of two or more persons." This definition of a test, offered by Lee Cronbach many years ago (1949/1960) probably still epitomizes the major content of psychological assessment. The invention of psychological tests, then known as mental tests, is generally attributed to Galton (Boring, 1950) and occurred during the middle and late 19th century. Galton's work was largely concerned with differences between individuals, and his approach was essentially in opposition to the approaches of other psychologists of his time. Most psychologists then were primarily concerned with the exhaustive study of mental phenomena in a few participants, while Galton was more interested in somewhat less specific analyses of large numbers of people. Perhaps the first psychological test was the "Galton whistle," which evaluated high tone hearing. Galton also appeared to have believed in the statistical concept that held that errors of measurement in individuals could be cancelled out through the mass effect of large samples.

Obviously, psychologists have come a long way from the simple tests of Galton, Binet, and Munsterberg, and the technology of testing is now in the computer age, with almost science-fiction-like extensions, such as testing by satellite. Psychometrics is now an advanced branch of mathematical and statistical science, and the administration, scoring, and even interpretation of tests have become increasingly objectified and automated. While some greet the news with dread and others with enthusiasm, we may be rapidly approaching the day when most of all testing will be administered, scored, and interpreted by computer. Thus, the 19th-century image of the school teacher administering paper-and-pencil tests to the students in her classroom and grading them at home has changed to the extensive use of automated procedures administered to huge portions of the population by representatives of giant corporations. Testing appears to have become a part of western culture, and there are indeed very few people who enter educational, work, or clinical settings who do not take many tests during their lifetimes.

In recent years, there appears to have been a distinction made between testing and assessment, assessment being the broader concept. Psychologists do not just give tests now; they perform assessments. The title of this volume, the *Handbook of Psychological Assessment*, was chosen advisedly and is meant to convey the view that it is not simply a handbook of psychological testing, although testing will be covered in great detail. The term assessment implies that there are many ways of evaluating individual differences. Testing is one way, but there are also interviewing, observations of behavior in natural or structured settings, and recording of various physiological functions. Certain forms of interviewing and systematic observation of behavior are now known as behavioral assessments, as opposed to the psychometric assessment accomplished through the use of formal tests. Historically, interest

in these two forms of assessment has waxed and waned, and in what follows we will briefly try to trace these trends in various areas.

INTELLIGENCE TESTING

The testing of intelligence in school children was probably the first major occupation of clinical psychology. The Binet scales and their descendants continue to be used, along with the IQ concept associated with them. Later, primarily through the work of David Wechsler and associates (Wechsler, 1944), intelligence testing was extended to adults and the IQ concept was changed from the mental age system (Mental Age/Chronological Age x 100) to the notion of a deviation IQ based on established norms. While Wechsler was primarily concerned with the individual assessment of intelligence, many group-administered paper-and-pencil tests also emerged during the early years of the 20th century. Perhaps the old Army Alpha and Beta tests, developed for intellectual screening of inductees into the armed forces during the first world war, were the first examples of these instruments. Use of these tests progressed in parallel with developments in more theoretical research regarding the nature of intelligence. The English investigators Burt, Pearson, and Spearman and the Americans Thurstone and Guilford are widely known for their work in this area, particularly with factor analysis. The debate over whether intelligence is a general ability (g) or a series of specific abilities represents one of the classic controversies in psychology. A related controversy that is still very much with us (Jensen, 1983) has to do with whether intelligence is primarily inherited or acquired and with the corollary issue having to do with ethnic differences in intellectual ability.

Another highly significant aspect of intelligence testing has to do with its clinical utilization. The IQ now essentially defines the borders of mental retardation, and intelligence tests are extremely widely used to identify retarded children in educational settings (American Psychiatric Association [APA], *Diagnostic and statistical manual of mental disorders,* 4th ed. (DSM-IV, 1994). However, intelligence testing has gone far beyond the attempt to identify mentally retarded individuals and has become widely applied in the fields of psychopathology and neuropsychology. With regard to psychopathology, under the original impetus of David Rapaport and collaborators (Rapaport, 1945), the Wechsler scales became clinical instruments used in conjunction with other tests to evaluate patients with such conditions as schizophrenia and various stress-related disorders. In the field of neuropsychology, use of intelligence testing is perhaps best described by McFie's (1975) remark, "It is perhaps a matter of luck that many of the Wechsler subtests are neurologically relevant" (p. 14). In these applications, the intelligence test was basically used as an instrument with which the clinician could examine various cognitive processes, on the basis of which inferences could be made about the patient's clinical status.

In summary, the intelligence test has become a widely used assessment instrument in educational, industrial, military, and clinical settings. While in some applications the emphasis remains on the simple obtaining of a numerical IQ value, it would probably be fair to say that many, if not most, psychologists now use the intelligence test as a means of examining the individual's cognitive processes; of seeing how he or she goes about solving problems; of identifying those factors that may be interfering with adaptive thinking; of looking at various language and nonverbal abilities in brain-damaged patients; and of identifying patterns of abnormal thought processes seen in schizophrenic and autistic patients. Performance profiles and qualitative characteristics of individual responses to items appear to have become the major foci of interest, rather than the single IQ score. The recent appearance of the new Wechsler Adult Intelligence Scale (WAIS-III) (Wechsler, 1997) reflects the major impacts cognitive psychology and neuropsychology have had on the way in which intelligence is currently conceptualized.

PERSONALITY ASSESSMENT

Personality assessment has come to rival intelligence testing as a task performed by psychologists. However, while most psychologists would agree that an intelligence test is generally the best way to measure intelligence, no such consensus exists for personality evaluation. In long-term perspective, it would appear that two major philosophies and perhaps three assessment method's have emerged. The two philosophies can be traced back to Allport's (1937) distinction between nomothetic versus idiographic methodologies and Meehl's (1954) distinction between clinical and statistical or actuarial prediction. In essence, some psychologists

feel that personality assessments are best accomplished when they are highly individualized, while others have a preference for quantitative procedures based on group norms. The phrase "seer versus sign" has been used to epitomize this dispute. The three methods referred to are the interview, and projective and objective tests. Obviously, the first way psychologists and their predecessors found out about people was to talk to them, giving the interview historical precedence. But following a period when the use of the interview was eschewed by many psychologists, it has made a return. It would appear that the field is in a historical spiral, with various methods leaving and returning at different levels.

The interview began as a relatively unstructured conversation with the patient and perhaps an informant, with varying goals, including obtaining a history, assessing personality structure and dynamics, establishing a diagnosis, and many other matters. Numerous publications have been written about interviewing (e.g., Menninger, 1952), but in general they provided outlines and general guidelines as to what should be accomplished by the interview. However, model interviews were not provided. With or without this guidance, the interview was viewed by many as a subjective, unreliable procedure that could not be sufficiently validated. For example, the unreliability of psychiatric diagnosis based on studies of multiple interviewers had been well established (Zubin, 1967). More recently, however, several structured psychiatric interviews have appeared in which the specific content, if not specific items, has been presented, and for which very adequate reliability has been established. There are by now several such interviews available including the Schedule for Affective Disorders and Schizophrenia (SADS) (Spitzer & Endicott, 1977), the Renard Diagnostic Interview (Helzer, Robins, Croughan, & Welner, 1981), and the Structured Clinical Interview for DSM-III, DSM-III-R, or DSM-IV (SCID or SCID-R) (Spitzer & Williams, 1983) (now updated for DSM-IV). These interviews have been established in conjunction with objective diagnostic criteria including DSM-III itself, the Research Diagnostic Criteria (Spitzer, Endicott, & Robins, 1977), and the Feighner Criteria (Feighner, et al., 1972). These new procedures have apparently ushered in a "comeback" of the interview, and many psychiatrists and psychologists now prefer to use these procedures rather than either the objective- or projective-type psychological test.

Those advocating use of structured interviews point to the fact that in psychiatry, at least, tests must ultimately be validated against judgments made by psychiatrists. These judgments are generally based on interviews and observation, since there really are no biological or other objective markers of most forms of psychopathology. If that is indeed the case, there seems little point in administering elaborate and often lengthy tests when one can just as well use the criterion measure itself, the interview, rather than the test. There is no way that a test can be more valid than an interview if an interview is the validating criterion. Structured interviews have made a major impact on the scientific literature in psychopathology, and it is rare to find a recently written research report in which the diagnoses were not established by one of them. It would appear that we have come full cycle regarding this matter, and until objective markers of various forms of psychopathology are discovered, we will be relying primarily on the structured interviews for our diagnostic assessments.

Interviews such as the SCID or the Diagnostic Interview Schedule (DIS) type are relatively lengthy and comprehensive, but there are now several briefer, more specific interview or interview-like procedures. Within psychiatry, perhaps the most well-known procedure is the Brief Psychiatric Rating Scale (BPRS) (Overall & Gorham, 1962). The BPRS is a brief, structured, repeatable interview that has essentially become the standard instrument for assessment of change in patients, usually as a function of taking some form of psychotropic medication. In the specific area of depression, the Hamilton Depression Scale (Hamilton, 1960) plays a similar role. There are also several widely used interviews for patients with dementia, which generally combine a brief mental-status examination and some form of functional assessment, with particular reference to activities of daily living. The most popular of these scales are the Mini-Mental Status Examination of Folstein, Folstein, and McHugh (1975) and the Dementia Scale of Blessed, Tomlinson, and Roth (1968). Extensive validation studies have been conducted with these instruments, perhaps the most well-known study having to do with the correlation between scores on the Blessed, Tomlinson, and Roth scale used in patients while they are living and the senile plaque count determined on autopsy in patients with dementia. The obtained correlation of .7 quite impressively suggested that

the scale was a valid one for detection of dementia. In addition to these interviews and rating scales, numerous methods have been developed by nurses and psychiatric aids for assessment of psychopathology based on direct observation of ward behavior (Raskin, 1982). The most widely used of these rating scales are the Nurses' Observation Scale for Inpatient Evaluation (NOSIE-30) (Honigfeld & Klett, 1965) and the Ward Behavior Inventory (Burdock, Hardesty, Hakerem, Zubin, & Beck, 1968). These scales assess such behaviors as cooperativeness, appearance, communication, aggressive episodes, and related behaviors, and are based on direct observation rather than reference to medical records or the report of others. Scales of this type supplement the interview with information concerning social competence and capacity to carry out functional activities of daily living.

Again taking a long-term historical view, it is our impression that after many years of neglect by the field, the interview has made a successful return to the arena of psychological assessment; but interviews now used are quite different from the loosely organized, "freewheeling," conversation-like interviews of the past (Hersen & Van Hassett, 1998). First, their organization tends to be structured, and the interviewer is required to obtain certain items of information. It is generally felt that formulation of specifically-worded questions is counterproductive; rather, the interviewer, who should be an experienced clinician trained in the use of the procedure, should be able to formulate questions that will elicit the required information. Second, the interview procedure must meet psychometric standards of validity and reliability. Finally, while structured interviews tend to be atheoretical in orientation, they are based on contemporary scientific knowledge of psychopathology. Thus, for example, the information needed to establish a differential diagnosis within the general classification of mood disorders is derived from the scientific literature on depression and related mood disorders.

The rise of the interview appears to have occurred in parallel with the decline of projective techniques. Those of us in a chronological category that may be roughly described as middle-age may recall that our graduate training in clinical psychology probably included extensive course work and practicum experience involving the various projective techniques. Most clinical psychologists would probably agree that even though projective techniques are still used to some extent, the atmosphere of ferment and excitement concerning these procedures that existed during the 1940s and 1950s no longer seems to exist. Even though the Rorschach technique and Thematic Apperception Test (TAT) were the major procedures used during that era, a variety of other tests emerged quite rapidly: the projective use of human-figure drawings (Machover, 1949), the Szondi Test (Szondi, 1952), the Make-A-Picture-Story (MAPS) Test (Shneidman, 1952), the Four-Picture Test (VanLennep, 1951), the Sentence Completion Tests (e.g., Rohde, 1957), and the Holtzman Inkblot Test (Holtzman, 1958). The exciting work of Murray and his collaborators reported on in *Explorations in Personality* (Murray, 1938) had a major impact on the field and stimulated extensive utilization of the TAT. It would probably be fair to say that the sole survivor of this active movement is the Rorschach test. Many clinicians continue to use the Rorschach test, and the work of Exner and his collaborators has lent it increasing scientific respectability (see Chapter 17 in this volume).

There are undoubtedly many reasons for the decline in utilization of projective techniques, but in our view they can be summarized by the following points:

1. Increasing scientific sophistication created an atmosphere of skepticism concerning these instruments. Their validity and reliability were called into question by numerous studies (e.g., Swensen, 1957, 1968; Zubin, 1967), and a substantial segment of the professional community felt that the claims made for these procedures could not be substantiated.
2. Developments in alternative procedures, notably the MMPI and other objective tests, convinced many clinicians that the information previously gained from projective tests could be gained more efficiently and less expensively with objective methods. In particular, the voluminous Minnesota Multiphasic Personality Inventory (MMPI) research literature has demonstrated its usefulness in an extremely wide variety of clinical and research settings. When the MMPI and related objective techniques were pitted against projective techniques during the days of the "seer versus sign" controversy, it was generally demonstrated that sign was as good as or better than seer in most of the studies accomplished (Meehl, 1954).

3. In general, the projective techniques are not atheoretical and, in fact, are generally viewed as being associated with one or another branch of psychoanalytic theory. While psychoanalysis remains a strong and vigorous movement within psychology, there are numerous alternative theoretical systems at large, notably behaviorally and biologically oriented systems. As implied in the section of this chapter covering behavioral assessment, behaviorally oriented psychologists pose theoretical objections to projective techniques and make little use of them in their practices. Similarly, projective techniques tend not to receive high levels of acceptance in biologically-oriented psychiatry departments. In effect, then, utilization of projective techniques declined for scientific, practical, and philosophical reasons. However, the Rorschach test in particular continues to be productively used, primarily by psychodynamically oriented clinicians.

The early history of objective personality tests has been traced by Cronbach (1949, 1960). The beginnings apparently go back to Sir Francis Galton, who devised personality questionnaires during the latter part of the 19th century. We will not repeat that history here, but rather will focus on those procedures that survived into the contemporary era. In our view, there have been three such major survivors: a series of tests developed by Guilford and collaborators (Guilford & Zimmerman, 1949), a similar series developed by Cattell and collaborators (Cattell, Eber, & Tatsuoka, 1970), and the MMPI. In general, but certainly not in all cases, the Guilford and Cattell procedures are used for individuals functioning within the normal range, while the MMPI is more widely used in clinical populations. Thus, for example, Cattell's 16PF test may be used to screen job applicants, while the MMPI may be more typically used in psychiatric health-care facilities. Furthermore, the Guilford and Cattell tests are based on factor analysis and are trait-oriented, while the MMPI in its standard form does not make use of factor analytically derived scales and is more oriented toward psychiatric classification. Thus, the Guilford and Cattell scales contain measures of such traits as dominance or sociability, while most of the MMPI scales are named after psychiatric classifications such as paranoia or hypochondriasis.

Currently, most psychologists use one or more of these objective tests rather than interviews or projective tests in screening situations. For example, many thousands of patients admitted to psychiatric facilities operated by the Veterans Administration take the MMPI shortly after admission, while applicants for prison-guard jobs in the state of Pennsylvania take the Cattell 16PF. However, the MMPI in particular is commonly used as more than a screening instrument. It is frequently used as a part of an extensive diagnostic evaluation, as a method of evaluating treatment, and in numerous research applications. There is little question that it is the most widely used and extensively studied procedure in the objective personality-test area. Even though the 566 true-or-false items have remained the same since the initial development of the instrument, the test's applications in clinical interpretation have evolved dramatically over the years. We have gone from perhaps an overly naive dependence on single-scale evaluations and overly literal interpretation of the names of the scales (many of which are archaic psychiatric terms) to a sophisticated configural interpretation of profiles, much of which is based on empirical research (Gilberstadt & Duker, 1965; Marks, Seeman, & Haller, 1974). Correspondingly, the methods of administering, scoring, and interpreting the MMPI have kept pace with technological and scientific advances in the behavioral sciences. From beginning with sorting cards into piles, hand scoring, and subjective interpretation, the MMPI has gone to computerized administration and scoring, interpretation based, at least to some extent, on empirical findings, and computerized interpretation. As is well known, there are several companies that will provide computerized scoring and interpretations of the MMPI.

Since the appearance of the earlier editions of this handbook, there have been two major developments in the field of objective personality-assessment. First, Millon has produced a new series of tests called the Millon Clinical Multiaxial Inventory (Versions I and II), the Millon Adolescent Personality Inventory, and the Millon Behavioral Health Inventory (Millon, 1982; 1985). Second, the MMPI has been completely revised and restandardized, and is now known as the MMPI-2. Since the appearance of the second edition of this handbook, use of the MMPI-2 has been widely adopted. Chapter 16 in this volume describes these new developments in detail.

Even though we should anticipate continued spiraling of trends in personality assessment, it would appear that we have passed an era of projective techniques and are now living in a time of objective assessment, with an increasing interest in the structured interview. There also appears to be increasing concern with the scientific status of our assessment procedures. In recent years, there has been particular concern about reliability of diagnosis, especially since distressing findings appeared in the literature suggesting that psychiatric diagnoses were being made quite unreliably (Zubin, 1967). The issue of validity in personality assessment remains a difficult one for a number of reasons. First, if by personality assessment we mean prediction or classification of some psychiatric diagnostic category, we have the problem of there being essentially no known objective markers for the major forms of psychopathology. Therefore, we are left essentially with psychiatrists' judgments. The DSM system has greatly improved this situation by providing objective criteria for the various mental disorders, but the capacity of such instruments as the MMPI or Rorschach test to predict DSM diagnoses has not yet been evaluated and remains a research question for the future. Some scholars, however, even question the usefulness of taking that research course rather than developing increasingly reliable and valid structured interviews (Zubin, 1984). Similarly, there have been many reports of the failure of objective tests to predict such matters as success in an occupation or trustworthiness with regard to handling a weapon. For example, objective tests are no longer used to screen astronauts, since they were not successful in predicting who would be successful or unsuccessful (Cordes, 1983). There does, in fact, appear to be a movement within the general public and the profession toward discontinuation of use of personality-assessment procedures for decision-making in employment situations. We would note as another possibly significant trend, a movement toward direct observation of behavior in the form of behavioral assessment, as in the case of the development of the Autism Diagnostic Observation Schedule (ADOS) (Lord et al., 1989). The zeitgeist definitely is in opposition to procedures in which the intent is disguised. Burdock and Zubin (1985), for example, argue that, "nothing has as yet replaced behavior for evaluation of mental patients."

NEUROPSYCHOLOGICAL ASSESSMENT

Another area that has an interesting historical development is neuropsychological assessment. The term itself is a relatively new one and probably was made popular through the first edition of Lezak's (1976) book of that title. Neuropsychological assessment is of particular historical interest because it represents a confluence of two quite separate antecedents: central and eastern European behavioral neurology and American and English psychometrics. Neurologists, of course, have always been concerned with the behavioral manifestations of structural brain damage and the relationship between brain function and behavior. Broca's discovery of a speech center in the left frontal zone of the brain is often cited as the first scientific neuropsychological discovery because it delineated a relatively specific relationship between a behavioral function, that is, speech, and a correspondingly specific region of the brain (the third frontal convolution of the left hemisphere). Clinical psychologists developed an interest in this area when they were called upon to assess patients with known or suspected brain damage. The first approach to this diagnostic area involved utilization of the already existing psychological tests, and the old literature deals primarily with how tests such as the Wechsler scales, the Rorschach test, or the Bender-Gestalt test could be used to diagnose brain damage. More recently, special tests were devised specifically for assessment work with patients having known or suspected brain damage.

The merger between clinical psychology and behavioral neurology can be said to have occurred when the sophistication of neurologists working in the areas of brain function and brain disease was combined with the psychometric sophistication of clinical psychology. The wedding occurred when reliable, valid, and well-standardized measurement instruments began to be used to answer complex questions in neurological and differential neuropsychiatric diagnosis. Thus, clinicians who ultimately identified themselves as clinical neuropsychologists tended to be individuals who knew their psychometrics, but who also had extensive training and experience in neurological settings. Just as many clinical psychologists work with psychiatrists, many clinical neuropsychologists work with neurologists and neurosurgeons. This relationship culminated in the development of standard neuropsychological test batteries, notably the Halstead-Reitan (Reitan & Wolfson, 1993) and

Luria-Nebraska batteries (Golden, Hammeke, & Purisch, 1980; Golden, Purisch, & Hammeke, 1985), as well as in the capacity of many trained psychologists to perform individualized neuropsychological assessments of adults and children. Thus, within the history of psychological assessment, clinical neuropsychological evaluation has recently emerged as an independent discipline to be distinguished from general clinical psychology on the basis of the specific expertise members of that discipline have in the areas of brain-behavior relationships and diseases of the nervous system. In recent years, there have been expansions of both the standard batteries and the individual neuropsychological tests. An alternate form (Golden, et al., 1985), as well as a children's version (Golden, 1981), of the Luria-Nebraska Neuropsychological Battery are now available. Prominent among the newly published or revised individual tests are the series of tests described in detail by Arthur Benton and collaborators in *Contributions to Neuropsychological Assessment* (Benton, Hamsher, Varney & Spreen, 1983), the California Verbal Learning Test (Delis, Kramer, Kaplan & Ober, 1987), and the recently revised and thoroughly reworked Wechsler Memory Scale (WMS-III) (Wechsler, 1997).

BEHAVIORAL ASSESSMENT

Over the last several decades behavioral assessment has been one of the most exciting developments to emerge in the field of psychological evaluation (Bellack & Hersen, 1988, 1998). Although its seeds were planted long before behavior therapy became a popular therapeutic movement, it is with the advent of behavior therapy that the strategies of behavioral assessment began to flourish (cf. Hersen & Bellack, 1976, 1981). As has been noted elsewhere (Barlow & Hersen, 1984; Hersen & Barlow, 1976a, 1976b), behavioral assessment can be conceptualized as a reaction to a number of factors. Among these were (a) problems with unreliability and invalidity of aspects of the DSM-I and DSM-II diagnostic schemes, (b) concerns over the indirect relationship between what was evaluated in traditional testing (e.g., the projectives) and how it subsequently was used in treatment planning and application, (c) increasing acceptance of behavior therapy by the professional community as a viable series of therapeutic modalities, and (d) parallel developments in the field of diagnosis in general, involving greater precision and accountability (e.g., the problem-oriented record).

We will briefly consider each of the four factors in turn and see how they contributed historically to the development of behavioral assessment. To begin with, DSM-I and DSM-II have been the targets of considerable criticism from psychiatrists (Hines & Williams, 1975) and psychologists alike (Begelman, 1975). Indeed, Begelman (1975), in a more humorous vein, referred to the two systems as "twice-told tales." They were "twice told" in the sense that neither resulted in highly reliable classification schemes when patients were independently evaluated by separate psychiatric interviewers (cf. Ash, 1949; Sandifer, Pettus, & Quade, 1964). Problems were especially evident when attempts to obtain interrater reliability were made for the more minor diagnostic groupings of the DSM schemes. Frequently, clinical psychologists would be consulted to carry out their testing procedures to confirm or disconfirm psychiatrists' diagnostic impressions based on DSM-I and DSM-II. But in so doing, such psychologists, operating very much as x-ray technicians, were using procedures (objective and projective tests) that only had a tangential relationship to the psychiatric descriptors for each of the nosological groups of interest. Thus, over time, the futility of this kind of assessment strategy became increasingly apparent. Moreover, not only were there problems with the reliability for DSM-I and DSM-II, but empirical studies documented considerable problems as well with regard to external validity of the systems (Eisler & Polak, 1971: Nathan, Zare, Simpson, & Ardberg, 1969).

Probably more important than any of the above was the fact that the complicated psychological evaluation had a limited relationship to eventual treatment. At least in the psychiatric arena, the usual isomorphic relationship between assessment and treatment found in other branches of therapeutics did not seem to hold. The isolated and extended psychological examination frequently proved to be an empty academic exercise resulting in poetic jargon in the report that eventuated. Its practical utility was woefully limited. Treatment seemed to be unrelated to the findings in the reports.

All of the aforementioned resulted in attempts by clinical psychologists to measure the behaviors of interest in direct fashion. For example, if a patient presented with a particular phobia, the

objective of evaluation was not to assess the underlying "neurotic complex" or "alleged psychodynamics." Quite the contrary, the primary objective was to quantify in distance how close our patient could approach the phobic object (i.e., the behavioral approach task) and how his heart rate (physiological assessment) increased as he got closer. In addition, the patient's cognitions (self-report) were quantified by having him assess his level of fear (e.g., on a 1–10 point scale). Thus, the behavioral assessment triad, consisting of motoric, physiological, and self-report systems (Hersen, 1973), was established as the alternative to indirect measurement.

Commenting on the use of direct measurement, Hersen and Barlow (1976) argue that

> whereas in indirect measurement a particular response is interpreted in terms of a presumed underlying disposition, a response obtained through direct measurement is simply viewed as a sample of a large population of similar responses elicited under those particular stimulus conditions....Thus, it is hardly surprising that proponents of direct measurement favor the observation of individuals in their natural surroundings whenever possible. When such naturalistic observations are not feasible, analogue situations approximating naturalistic conditions may be developed to study the behavior in question (e.g., the use of a behavioral avoidance test to study the degree of fear of snakes). When neither of these two methods is available or possible, subjects' *self-reports are also used as independent criteria,* and, at times, may be operating under the control of totally different sets of contingencies than those governing motoric responses. (p. 116)

We have already referred to the tripartite system of direct measurement favored by the behaviorists. But it is in the realm of motoric behavior that behavior therapists have made the greatest contributions as well as being most innovative (see Foster, Bell-Dolan, & Burge, 1988; Hersen, 1988; Tryon, 1986). With increased acceptance of behavior therapy, practitioners of the strategies found their services required in a large variety of educational, rehabilitation, community medical, and psychiatric settings. Very often they were presented with extremely difficult educational, rehabilitation, and treatment cases, both from assessment and therapeutic perspectives. Many of the clients and patients requiring remediation exhibited behaviors that previously had not been measured in any direct fashion. Thus, there were few guidelines with regard to how the behavior might be observed, quantified, and coded. In many instances, "seat-of-the-pants" measurement systems were devised on-the-spot but with little regard for psychometric qualities cherished by traditional testers.

Consider the following example of a measurement strategy to quantify "spasmodic torticollis," a tic-like disorder (Bernhardt, Hersen, & Barlow, 1972):

> A Sony Video Recorder model AV-5000A, an MRI Keleket model VC-1 television camera, and a Conrac 14-inch television monitor were employed in recording torticollis. A Gra Lab sixty-minute Universal Timer was used to obtain percentage of torticollis....A lightolier lamp served as the source of negative feedback.Two to three daily ten-minute sessions were scheduled during the experiment in which the subject was videotaped while seated in a profile arrangement. A piece of clear plastic containing superimposed Chart-Pac taped horizontal lines (spaced one-quarter to one-half inch apart) was placed over the monitor. A shielded observer depressed a switch activating the timer whenever the subject's head was positioned at an angle where the nostril was above a horizontal line intersecting the external auditory meatus. This position was operationally defined as an example of torticollis, with percentage of torticollis per session serving as the experimental measure. Conversely, when the horizontal line intersected both the nostril and auditory meatus or when the subject's nostril was below the horizontal line he was considered to be holding his head in a normal position. (p. 295)

If one peruses through the pages of the *Journal of Applied Behavior Analysis, Behaviour Research and Therapy, Journal of Behavior Therapy and Experimental Psychiatry,* and *Behavior Modification,* particularly in the earlier issues, numerous examples of innovative behavioral measures and more comprehensive systems are to be found. Consistent with the idiographic approach, many of these apply only to the case in question, have some internal or face validity, but, of course, have little generality or external validity. (Further comment on this aspect of behavioral assessment is made in a subsequent section of this chapter.)

A final development that contributed to and coincided with the emergence of behavioral assessment was the problem-oriented record (POR). This was a system of recordkeeping first instituted on medical wards in general hospitals to sharpen and pinpoint diagnostic practices (cf. Weed, 1964, 1968, 1969). Later this system was transferred to psychiatric units (cf. Hayes-Roth, Longabaugh, & Ryback, 1972; Katz & Woolley, 1975; Klonoff & Cox, 1975; McLean & Miles,

1974; Scales & Johnson, 1975), with its relevance to behavioral assessment increasingly evident (Atkinson, 1973; Katz & Woolley, 1975). When applied to psychiatry, the POR can be divided into four sections: (a) database, (b) problem list, (c) treatment plan, and (d) follow-up data. There can be no doubt that this kind of record keeping promotes and enhances the relationship of assessment and treatment, essentially forcing the evaluator to crystallize his or her thinking about the diagnostic issues. In this regard, we previously have pointed out that

> Despite the fact that POR represents, for psychiatry, a vast improvement over the type of record-keeping and diagnostic practice previously followed, the level of precision in describing problem behaviors and treatments to be used remedially *does not* yet approach the kind of precision reached in the carefully conducted behavioral analysis. (Hersen, 1976, p. 15)

However, the POR certainly can be conceptualized as a major step in the right direction. In most psychiatric settings some type of POR (linking it to specific treatment plans) has been or is currently being used and, to a large extent, has further legitimized the tenets of behavioral assessment by clearly linking the problem list with specific treatment (cf. Longabaugh, Fowler, Stout, & Kriebel, 1983; Longabaugh, Stout, Kriebel, McCullough, & Bishop, 1986).

ASSESSMENT SCHEMES

Since 1968 a number of comprehensive assessment schemes have been developed to facilitate the process of behavioral assessment (Cautela, 1968; Kanfer & Saslow, 1969; Lazarus, 1973). Since a very detailed analysis of these schemes is much beyond the scope of this brief historical overview, we will only describe the outlines of each in order to illustrate how the behavioral assessor conceptualizes his or her cases. For example, Cautela (1968) depicted in his scheme the role of behavioral assessment during the various stages of treatment. Specifically, he delineated three stages.

In the *first stage* the clinician identifies maladaptive behaviors and those antecedent conditions maintaining them. This step is accomplished through interviews, observation, and self-report questionnaires. The *second stage* involves selection of the appropriate treatment strategies, evaluation of their efficacy, and the decision when to terminate their application. In the *third stage* a meticulous follow-up of treatment outcome is recommended. This is done by examining motoric, physiological, and cognitive functioning of the client, in addition to independent confirmation of the client's progress by friends, relatives, and employers.

A somewhat more complicated approach to initial evaluation was proposed by Kanfer and Saslow (1969), which involves some seven steps. The *first* involves a determination as to whether a given behavior represents an excess, deficit, or an asset. The *second* is a clarification of the problem and is based on the notion that in order to be maintained, maladjusted behavior requires continued support. *Third* is the motivational analysis in which reinforcing and aversive stimuli are identified. *Fourth* is the developmental analysis, focusing on biological, sociological, and behavioral changes. *Fifth* involves assessment of self-control and whether it can be used as a strategy during treatment. *Sixth* is the analysis of the client's interpersonal life, and *seventh* is the evaluation of the patient's socio-cultural-physical environment.

In their initial scheme, Kanfer and Saslow (1969) viewed the system, in complementary fashion, to the existing diagnostic approach (i.e., DSM-II). They did not construe it as supplanting DSM-II. But they did see their seven-part analysis as serving as a basis for arriving at decisions for precise behavioral interventions, thus yielding a more isomorphic relationship between assessment and treatment. Subsequently, Kanfer and Grimm (1977) have turned their attention to how the interview contributes to the overall behavioral assessment. In so doing, suggestions are made for organizing client complaints under five categories:

> (1) behavioral deficiencies, (2) behavioral excesses, (3) inappropriate environmental stimulus control, (4) inappropriate self-generated stimulus control, and (5) problematic reinforcement contingencies. (p. 7)

Yet another behavioral assessment scheme had been proposed by Lazarus (1973), with the somewhat humorous acronym of BASIC ID: B = behavior, A = affect, S = sensation, I = imagery, C = cognition, I = interpersonal relationship, and D = the need for pharmacological intervention (i.e., drugs) for some psychiatric

patients. The major issue underscored by this diagnostic scheme is that if any of the elements is overlooked, assessment will be incomplete, thus resulting in only a partially effective treatment. To be fully comprehensive, deficits or surpluses for each of the categories need to be identified so that specific treatments can be targeted for each. This, then, should ensure the linear relationship between assessment and treatment, ostensibly absent in the nonbehavioral assessment schemes.

Despite development of the aforementioned schemes and others not outlined here (e.g., Bornstein, Bornstein, & Dawson, 1984), there is little in the way of their formal evaluation in empirical fashion. Although these schemes certainly appear to have a good bit of face validity, few studies, if any, have been devoted to evaluating concurrent and predictive validity. This, of course, is in contrast to the considerable effort to validate the third edition of DSM (i.e., DSM-III, 1980; Hersen & Turner, 1984) and its revisions (i.e., DSM-III-R; 1987; DSM-IV, 1994).

In a somewhat different vein, Wolpe (1977) has expressed his concern about the manner in which behavioral assessment typically is being conducted. Indeed, he has referred to it as "The Achilles' Heel of Outcome Research in Behavior Therapy." He is especially concerned that too little attention has been devoted to evaluation of the antecedents of behaviors targeted for treatment, thus leading to a therapeutic approach that may be inappropriate. For example, in treating homosexuality, Wolpe (1977) rightly argues that

> It seems obvious that each factor found operative in a particular patient needs to be treated by a program appropriate to it. Failure is predictable when homosexuality that is exclusively based on approach conditioning to males is treated by desensitization to heterosexual themes, or if homosexuality based on timidity or on fear of females is treated by aversion therapy. To compare the effects of different treatments on assorted groupings of homosexuals is about as informative as to compare the effects of different antibiotics on tonsillitis without bacterial diagnosis. (p. 2)

The same analysis, of course, holds true for other disorders, such as depression (Wolpe, 1986) and phobia (Michelson 1984, 1986). Blanket treatment that does not take into account antecedents undoubtedly should fail (Wolpe & Wright, 1988). But here too, the necessary research findings to document this are as yet forthcoming (see White, Turner, & Turkat, 1983).

CHANGES IN BEHAVIORAL ASSESSMENT

Contrasted to the field of psychological assessment in general, behavioral assessment as a specialty has had a history of about four decades. However, in these three decades we have witnessed some remarkable changes in the thinking of behavioral assessors. Probably as a strong overt reaction to the problems perceived by behavioral assessors in traditional psychological evaluation, many of the sound psychometric features of that tradition were initially abandoned. Indeed, in some instances it appears that "the baby was thrown out with the bath water." As we already have noted, consistent with the idiographic approach to evaluation and treatment, little concern was accorded to traditional issues of reliability and validity. (The exception, of course, was the obsessive concern with high interrater reliability of observations of motoric behavior.) This was particularly the case for the numerous self-report inventories developed early on to be consistent with the motoric targets of treatment (e.g., some of the fear survey schedules).

There were many other aspects of traditional evaluation that also were given short shrift. Intelligence testing was eschewed, norms and developmental considerations were virtually ignored, and traditional psychiatric diagnosis was viewed as anathema to behavior therapy. However, since the late 1970s this "hard line" has been mollified. With publication of the second, third, and fourth editions of *Behavioral Assessment: A Practical Handbook* and emergence of two assessment journals (*Behavioral Assessment* and *Journal of Psychopathology and Behavioral Assessment*), greater attention to cherished psychometric principles has returned. For example, the external validity of role playing as an assessment strategy in the social skill areas has been evaluated by Bellack and his colleagues (cf. Bellack, Hersen, & Lamparski, 1979; Bellack, Hersen, & Turner, 1979; Bellack, Turner, Hersen, & Luber, 1980) instead of being taken on faith. Also in numerous overviews the relevance of the psychometric tradition to behavioral assessment has been articulated with considerable vigor (e.g., Adams & Turner, 1979; Cone, 1977, 1988; Haynes, 1978; Nelson & Hayes, 1979; Rosen, Sussman, Mueser, Lyons, & Davis, 1981). Looking at behavioral assessment today from a historical per-

spective, it certainly appears as though the "baby" is being returned from the discarded bath water.

Also, in recent years there have been several calls for a broadened conceptualization of behavioral assessment (e.g., Bellack & Hersen, 1998; Hersen, 1988; Hersen & Bellack, 1988; Hersen & Last, 1989); Hersen & Van Hassett, 1998). Such broadening has been most noticeable with respect to the use of intelligence tests in behavioral assessment (Nelson, 1980), the relevance of neuropsychological evaluation for behavioral assessment (Goldstein, 1979; Horton, 1988), the importance of developmental factors especially in child and adolescent behavioral assessment (Edelbrock, 1984; Harris & Ferrari, 1983; Hersen & Last, 1989), and the contribution that behavioral assessment can make to pinpointing of psychiatric diagnosis (Hersen, 1988; Tryon, 1986, 1998).

DSMS III, III-R, IV AND BEHAVIORAL ASSESSMENT

In the earlier days of behavioral assessment, traditional psychiatric diagnosis was, for the most part, eschewed. Behavioral assessors saw little relationship between what they were doing and the overall implicit goals of DSM-II. Moreover, as we have noted, categories subsumed under DSM-II had major problems with reliability and validity. So, consistent with cogent criticisms about the official diagnostic system, behavioral assessors tended to ignore it when possible. They continued to develop their strategies independently of DSM-II and the then emerging DSM-III. In fact, some (e.g., Adams, Doster, & Calhoun, 1977; Cautela, 1973) advocated totally new diagnostic formats altogether, but these never had a chance of being accepted by the general diagnostic community, given the political realities.

In spite of its problems and limitations, with the emergence of DSM-III (APA, 1980), behavioral therapists and researchers appeared to have retrenched and assumed a somewhat different posture (cf. Hersen, 1988; Hersen & Bellack; 1988; Hersen & Turner, 1984; Nelson, 1987). Such positions have been articulated by a number of prominent behavior therapists, such as Nathan (1981) and Kazdin (1983). But the issues concerning DSM-III and behavioral assessment are most clearly summarized by Taylor (1983), a behavioral psychiatrist:

> The new Diagnostic and Statistical Manual of the American Psychiatric Association is a major improvement in psychiatric diagnosis over previous classification systems. Where symptomatic diagnoses are useful, as in relating an individual's problem to the wealth of clinical and research data in abnormal psychology or in identifying conditions which require specific treatments, DSM-III represents the best available system. Many conceptual and practical problems remain with DSM-III; for instance, it retains a bias toward the medical model, it includes many conditions which should not fall into a psychiatric diagnostic system, and it includes descriptive axes which have not been adequately validated. Nevertheless, behavior therapists are well advised to become familiar with and use DSM-III as part of behavioral assessment. (p. 13)

We, of course, would argue that the same holds true for the DSM system. We are fully in accord with Taylor's comments and believe that if behavior therapists wish to impact on the accepted nosological system, they are urged to work from within rather than from without. In this connection, Tryon (1986) has presented the field with a marvelous outline for how motoric measurements in both children and adults will enable the DSM categories to gain greater precision. He clearly shows how many of the diagnostic categories (e.g., depression; attention deficit-hyperactivity disorders) have motoric referents that could be evaluated by behavioral assessors. However, much work of a normative nature (to determine lower and upper limits of normality) will be required before any impact on the DSM system will be felt (Tryon, 1989). We believe that such evaluation represents an enormous challenge to behavioral assessors that could result in a lasting contribution to the diagnostic arena.

SUMMARY

We have provided a brief historical overview of several major areas in psychological evaluation: intellectual, personality, neuropsychological, and behavioral assessment. Some of these areas have lengthy histories, and others are relatively young. However, it seems clear that the tools used by psychologists as recently as 25 years ago are generally different from those used now. Behavioral assessment techniques, structured psychiatric interviews, and standard, comprehensive neuropsychological test batteries are all relatively new. Furthermore, the computer is making significant inroads into the assessment field, with on-line testing, scoring, and

interpretation a reality in some cases. Serious efforts have been made in recent years to link assessment more closely to treatment and other practical concerns. We may also note a trend away from indirect methods to direct acquisition of information and observation. The structured interview is an example of the former approach, and many behavioral assessment techniques would exemplify the latter one. Similarly, while neuropsychological assessment is still heavily dependent on the use of formal tests, there is increasing interest in the use of those tests in rehabilitation planning and in the association between neuropsychological test results and functional activities of daily living. We also note a corresponding decrease in interest in such matters as brain localization, particularly since the CT scan, MRI, and related brain-imaging procedures have solved much of that problem. We would prognosticate that psychological assessment will be increasingly concerned with automation, the direct observation of behavior, and the practical application of assessment results.

REFERENCES

Adams, H. E., Doster, J. A., & Calhoun, K. S. (1977). A psychologically-based system of response classification. In A. R. Ciminero, K. S. Calhoun, & H. E. Adams (Eds.), *Handbook of behavioral assessment* New York: Wiley.

Adams, H. E., & Turner, S. M. (1979). Editorial. *Journal of Behavioral Assessment, l,* 1–2.

Allport, G. W. (1937). *Personality: A psychological interpretation.* New York: Holt.

American Psychiatric Association (1980). *Diagnostic and statistical manual of mental disorders* (3rd ed.). Washington, DC: Author.

American Psychiatric Association (1987). *Diagnostic and statistical manual of mental disorders* (3rd ed. Rev.) Washington, DC: Author.

American Psychiatric Association (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.) Washington, DC: Author.

Ash, P. (1949). The reliability of psychiatric diagnosis. *Journal of Abnormal and Social Psychology, 44,* 272–276.

Atkinson, C. (1973). *Data collection and program evaluation using the problem-oriented medical record.* Miami, FL: Association for Advancement of Behavior Therapy.

Barlow, D. H., & Hersen, M. (1984). *Single-case experimental designs: Strategies for studying behavior change* (2nd ed.). New York: Pergamon Press.

Begelman, D. A. (1975). Ethical and legal issues in behavior modification. In M. Hersen, R. M. Eisler, & P. M. Miller (Eds.), *Progress in behavior modification* (Vol. 1.) New York: Academic Press.

Bellack, A. S., & Hersen, M. (1988). Future directions. In A. S. Bellack & M. Hersen (Eds.), *Behavioral assessment: A practical handbook* (3rd ed.). New York: Pergamon Press.

Bellack, A. S. & Hersen, M. (Eds.). (1998). *Behavioral assessment: A practical handbook* (4th ed.). Needham Heights, MA: Allyn & Bacon.

Bellack, A. S., Hersen, M., & Lamparski, D. (1979). Role-playing tests for assessing social skills: Are they valid? Are they useful? *Journal of Consulting and Clinical Psychology, 47,* 335–342.

Bellack, A. S., Hersen, M., & Turner, S. M. (1979). Relationship of role playing and knowledge of appropriate behavior to assertion in the natural environment. *Journal of Consulting and Clinical Psychology, 47,* 679–685.

Bellack, A. S., Turner, S. M., Hersen, M., & Luber, R. (1980). Effects of stress and retesting on role-playing tests of social skill. *Journal of Behavioral Assessment, 2,* 99–104.

Benton, A. L., Hamsher, K. deS., Varney, N. R., & Spreen, O. (1983). *Contributions to neuropsychological assessment: A clinical manual.* New York: Oxford University Press.

Bernhardt, A. J., Hersen, M., & Barlow, D. H. (1972). Measurement and modification of spasmodic torticollis: An experiment analysis. *Behavior Therapy, 3,* 294–297.

Blessed, G., Tomlinson, B. E., & Roth, M. (1968). The association between quantitative measures of dementia and of senile change in the cerebral grey matter of elderly subjects. *British Journal of Psychiatry, 114,* 797–811.

Boring, E. G. (1950). *A history of experimental psychology.* New York: Appleton-Century-Crofts.

Bornstein, P. H., Bornstein, M. T., & Dawson, B. (1984). Integrated assessment and treatment. In T. H. Ollendick, & M. Hersen (Eds.), *Child behavioral assessment: Principles and procedures.* New York: Pergamon Press.

Burdock, E. I., Hardesty, A. S., Hakerem, G., Zubin, J., & Beck, Y. M. (1968). *Ward behavior inventory.* New York: Springer.

Burdock, E I., & Zubin, J. (1985). Objective evaluation in psychiatry. *Psychiatric reference and record book* (2nd ed). New York: Roerig Laboratories, Inc.

Cattell, R. B., Eber, H. W., & Tatsuoka, M. M. (1970). *Handbook for the sixteen personality factor questionnaire.* (Technical Report), Champaign, IL: Institute for Personality and Ability Testing.

Cautela, J. R. (1968). Behavior therapy and the need for behavior assessment. *Psychotherapy: Theory, Research and Practice, 5,* 175–179.

Cautela, J. R. (1973, September). *A behavioral coding system.* Presidential address presented at the seventh annual meeting of the Association for Advancement of Behavioral Therapy, Miami, FL.

Cone, J. D. (1977). The relevance of reliability and validity for behavioral assessment. *Behavioral Therapy, 8,* 411–426.

Cone, J. D. (1988). Psychometric considerations and the multiple models of behavioral assessment. In A. S. Bellack & M. Hersen (Eds.), *Behavioral assessment: A practical handbook* (3rd ed.). New York: Pergamon Press.

Cordes, C. (1983). Mullane: Tests are grounded. *APA Monitor, 14,* 24.

Cronbach, L. J. (1960). *Essentials of psychological testing.* (2nd ed.) New York: Harper & Brothers. Original work published 1949

Delis, D. C., Kramer, J. H., Kaplan, E., & Other, B. A., (1987). *CVLT: California verbal learning test: Research Edition.* [Manual]. San Antonio, TX: The Psychological Corporation.

Edelbrock, C. (1984). Diagnostic issues. In T. H. Ollendick & M. Hersen (Eds.), *Child behavioral assessment: Principles and procedures* (pp. 30–37). New York: Pergamon Press.

Eisler, R. M., & Polak, P. R. (1971). Social stress and psychiatric disorder. *Journal of Nervous and Mental Disease, 153,* 227–233.

Foster, S. L., Bell-Dolan, D. J., & Burge, D. A. (1988). Behavioral observation. In A. S. Bellack & M. Hersen (Eds.), *Behavioral assessment: A practical handbook* New York: Pergamon Press.

Feighner, J., Robins, E., Guze, S., Woodruff, R., Winokur, G., & Munoz, R. (1972). Diagnostic criteria for use in psychiatric research. *Archives of General Psychiatry, 26,* 57–63.

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). Mini-mental state. A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research, 12,* 189–198.

Foster, S. L., Bell-Dolan, D. J., & Burge, D. A. (1988). Behavioral observation. In A. s. Bellack & M. Hersen (Eds.), *Behavioral assessment: A practical handbook.* New York: Pergamon Press.

Gilberstadt, H., & Duker, J. (1965). *A handbook for clinical and actuarial MMPI interpretation.* Philadelphia: Saunders.

Golden, C. J., Hammeke, T. A., & Purisch, A. D. (1980). *The Luria-Nebraska Battery manual.* Los Angeles: Western Psychological Services.

Golden, C. J., Purisch, A. D., & Hammeke, T. A. (1985). *The Luria-Nebraska Battery: Forms I and II.* Los Angeles: Western Psychological Services.

Golden, G. (1981). The Luria-Nebraska children's battery: Theory and formulation. In G. W. Hynd & J. E. Obrzut (Eds.), *Neuropsychological assessment and the school-aged child: Issues and procedures* New York: Grune & Stratton.

Goldstein, G. (1979). Methodological and theoretical issues in neuropsychological assessment. *Journal of Behavioral Assessment, 1,* 23–41.

Guilford, J. P., & Zimmerman, W. (1949). *Guilford-Zimmerman Temperament survey.* Los Angeles: Western Psychological Services.

Hamilton, M. (1960). A rating scale for depression. *Journal of Neurology, Neurosurgery and Psychiatry, 23,* 56–62.

Harris, S. L., & Ferrari, M. (1983). Developmental factors in child behavior therapy. *Behavior Therapy, 14,* 54–72.

Hayes-Roth, F., Longabaugh, R., & Ryback, R. (1972). The problem-oriented medical record and psychiatry. *British Journal of Psychiatry, 121,* 27–34.

Haynes, S. N. (1978). *Principles of behavioral assessment.* New York: Gardner Press.

Helzer, J., Robins, L., Croughan, J., & Welner, A. (1981). Renard Diagnostic Interview. *Archives of General Psychiatry, 38,* 393–398.

Hersen, M. (1973). Self-assessment and fear. *Behavior Therapy, 4,* 241–257.

Hersen, M. (1976). Historical perspectives in behavioral assessment. In M. Hersen & A. S. Bellack (Eds.), *Behavioral assessment designs: Strategies for studying behavior change.* New York: Pergamon Press.

Hersen, M. (Ed.). (1988). Behavioral assessment and psychiatric diagnosis. *Behavioral Assessment, 10,* 107–121.

Hersen, M., & Barlow, D. H. (1976a). *Single-case experimental designs: Strategies for studying behavior change.* New York: Pergamon Press.

Hersen, M., & Barlow, D. H. (1976b). *Single-case experimental designs: Strategies for studying behavior change.* New York: Pergamon Press.

Hersen, M., & Bellack, A. S. (Eds.) (1976). *Behavioral assessment: A practical handbook* (1st ed.). New York: Pergamon Press.

Hersen, M., & Bellack, A. S. (1981). *Behavioral assessment: A practical handbook* (2nd ed.). New York: Pergamon Press.

Hersen, M., & Bellack, A. S. (1988). DSM-III and behavioral assessment. In A. S. Bellack & M. Hersen (Eds.), *Behavioral assessment: A practical handbook* (3rd edition). New York: Pergamon Press.

Hersen, M., & Last, C. G. (1989). Psychiatric diagnosis and behavioral assessment in children. In M. Hersen, & C. G. Last (Eds.), *Handbook of child psychiatric diagnosis.* New York: John Wiley & Sons.

Hersen, M., & Turner, S. M. (1984). DSM-III and behavior therapy. In S. M. Turner & M. Hersen (Eds.), *Adult psychopathology: A behavioral perspective.* New York: Wiley.

Hersen, M. & Van Hassett, V. 13. (Eds.). (1998). *Basic interviewing: A practical guide for counselors and clinicians.* Mahwah, NJ.

Hines, F. R., & Williams, R. B. (1975). Dimensional diagnosis and the medical students' grasp of psychiatry. *Archives of General Psychiatry, 32,* 525–528.

Holtzman, W. H. (1958). *The Holtzman inkblot technique.* New York: Psychological Corporation.

Honigfeld, G., & Klett, C. (1965). The Nurse's Observation Scale for Impatient Evaluation (NOSIE): A new scale for measuring improvement in schizophrenia. *Journal of Clinical Psychology, 21,* 65–71.

Horton, A. M. (1988). Use of neuropsychological testing in determining effectiveness of ritalin therapy in an DDRT patient. *Behavior Therapist, 11,* 114–118.

Jensen, A. R. (1983, August). *Nature of the white-black differences on various psychometric tests.* Invited address presented at the meeting of the American Psychological Association Convention, Anaheim, CA.

Kanfer, F. H., & Grimm, L. G. (1977). Behavior analysis: Selecting target behaviors in the interview. *Behavior Modification, l,* 7–28.

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

Katz, R. C., & Woolley, F. R. (1975). Improving patients' records through problem orientation. *Behavior Therapy, 6,* 119–124.

Kazdin, A. E. (1983). Psychiatric diagnosis, dimensions of dysfunction, and child behavior therapy. *Behavior Therapy, 14,* 73–99.

Klonoff, H., & Cox, B. (1975). A problem-oriented system approach to analysis of treatment outcome. *American Journal of Psychiatry, 132,* 841–846.

Lazarus, A. A. (1973). Multimodal behavior therapy: Treating the "basic id." *Journal of Nervous and Mental disease, 156,* 404–411.

Lezak, M. (1976). *Neuropsychological Assessment.* New York: Oxford University Press.

Longabaugh, R., Fowler, D. R., Stout, R., & Kriebel, G. (1983). Validation of a problem-focused nomenclature. *Archives of General Psychiatry, 40,* 453–461.

Longabaugh, R., Stout, R., Kriebel, G. M., McCullough, L., & Bishop, D. (1986). DSM-III and clinically identified problems as a guide to treatment. *Archives of General Psychiatry, 43,* 1097–1103.

Lord, C., Rutter, M., Goode, S., Heemsbergen, J., Jordan, H., Mawhood, L., & Schopler, E. (1989). Autism diagnostic observation schedule: A standardized observation of communicative and social behavior. *Journal of Autism and Developmental Disorders, 19,* 185–212.

Machover, K. (1949). *Personality projection in the drawing of the human figure: A method of personality investigation.* Springfield, IL: Charles Thomas.

Marks, P. A., Seeman, W., & Haller, D. L. (1974). *The actuarial use of the MMPI with adolescents and adults.* Baltimore: Williams & Wilkins.

McFie, J. (1975). *Assessment of organic intellectual impairment.* London: Academic Press.

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

Meehl, P. E. (1954). *Clinical vs. statistical prediction.* Minneapolis, MN: University of Minnesota Press.

Menninger, K. A. (1952). *A manual for psychiatric case study.* New York: Grune & Stratton.

Michelson, L. (1984). The role of individual differences, response profiles, and treatment consonance in anxiety disorders. *Journal of Behavioral Assessment, 6,* 349–367.

Michelson, L. (1986). Treatment consonance and response profiles in agoraphobia: Behavioral and physiological treatments. *Behaviour Research and Therapy, 24,* 263–275.

Millon T. (1982). *Millon Clinical Multiaxial Inventory* (3rd ed.). Minneapolis, MN: National Computer Systems.

Millon T. (1985). The MCMI provides a good assessment of DSM-III disorders: The MCMI-II will prove even better. *Journal of Personality Assessment, 49,* 379–391.

Murray, H. A. (1938). *Explorations in personality.* New York: Oxford University Press.

Nathan, P. E. (1981). Symptomatic diagnosis and behavioral assessment: A synthesis. In D. H. Barlow (Ed.), *Behavioral assessment of adult disorders.* New York: Guilford.

Nathan, P. E., Zare, N. C., Simpson, H. F., & Ardberg, M. M. (1969). A systems analytic model of diagnosis: I. The diagnostic validity of abnormal psychomotor behavior. *Journal of Clinical Psychology, 25,* 3–9.

Nelson, R. O. (1979). DSM-III and behavioral assessment. In M. Hersen & C. G. Last (Eds.), *Issues in diagnostic research.* New York: Plenum Press.

Nelson, R. O. (1980). The use of intelligence tests within behavioral assessment. *Behavioral Assessment, 2,* 417–423.

Nelson, R. O. (1987). DSM-III and behavioral assessment. In C. G. Last & M. Hersen (Eds.), *Issues in diagnostic research.* New York: Plenum Press.

Nelson, R. O., & Hayes, S. C. (1979). Some current dimensions on behavioral assessment. *Behavioral Assessment, 1*, 1–16.

Overall, J. E., & Gorham, J. R. (1962). The brief psychiatric rating scale. *Psychological Reports, 10*, 799–812.

Rapaport, D. (1945). *Diagnostic psychological testing*. Chicago: Year Book Publications.

Raskin, A. (1982). Assessment of psychopathology by the nurse or psychiatric aide. In E. I. Burdock, A. Sudilovsky, & S. Gershon (Eds.), *The behavior of psychiatric patients: Quantitative techniques for evaluation*. New York: Marcel Dekker.

Reitan, R. M., & Wolfson, D. (1993). *The Halstead-Reitan Neuropsychological Test Battery: Theory and clinical interpretation* (2nd. ed.). Tucson, AZ: Neuropsychology Press.

Robins, S. L., Helzer, J., Croughan, N. A., & Ratcliff, K. (1981). National Institute of Mental Health Diagnostic Interview Schedule. *Archives of General Psychiatry, 38*, 381–389.

Rohde, A. R. (1957). *The sentence completion method*. New York: Ronald Press.

Rosen, A. J., Sussman, S., Mueser, K. T., Lyons, J. S., & Davis, J. M. (1981). Behavioral assessment of psychiatric inpatients and normal controls across different environmental contexts. *Journal of Behavioral Assessment, 3*, 25–36.

Sandifer, M. G., Jr., Pettus, C., & Quade, D. (1964). A study of psychiatric diagnosis. *Journal of Nervous and Mental Disease, 139*, 350–-356.

Scales, E. J., & Johnson, M. S. (1975). A psychiatric POMR for use by a multidisciplinary team. *Hospital and Community Psychiatry, 26*, 371–373.

Shneidman, E. S. (1952). *Make a picture test*. New York: The Psychological Corporation.

Spitzer, R. L., & Endicott, J. (1977). *Schedule for affective disorders and schizophrenia*. (Technical Report). New York: New York State Psychiatric Institute, Biometrics Research Department.

Spitzer, R. L., Endicott, J., & Robins, E. (1977). *Research diagnostic Criteria (RDC) for a selected group of functional disorders*. Bethesda, MD: National Institute of Mental Health.

Spitzer, R. L., & Williams, J. B. W. (1983). *Instruction manual for the structured clinical interview for DSM-III (SCID)*. New York: New York State Psychiatric Institute, Biometrics Research Department.

Swensen, C. H. (1957). Empirical evaluations of human figure drawings, 1957–1966. *Psychological Bulletin, 54*, 431–466.

Swensen, C. H. (1968). Empirical evaluations of human figure drawings. *Psychological Bulletin, 20*, 20–44.

Szondi, L. (1952). *Experimental diagnostics of drives*. New York: Grune & Stratton.

Taylor, C. B. (1983). DSM-III and behavioral assessment. *Behavioral Assessment, 5*, 5–14.

Tryon, W. W. (1986). Motor activity measurements and DSM-III. In M. Hersen (Ed.), *Innovations in child behavior therapy*. New York: Springer.

Tryon, W. W. (1989). Behavioral assessment and psychiatric diagnosis. In M. Hersen (Ed.), *Innovations in child behavior therapy*. New York: Springer.

Tyron, W. W. (1998). In A. S. Bellack & M. Hersen (Eds.), *Behavioral assessment: A practical handbook* (4th ed.). Needham Heights, MA: Allyn & Bacon.

VanLennep, D. J. (1951). The four-picture test. In H. H. Anderson & G. L. Anderson (Eds.), *An introduction to projective techniques*. New York: Prentice-Hall.

Wechsler, D. (1944). *The measurement of adult intelligence*. Baltimore, MD: Williams & Wilkins.

Wechsler, D. (1997). *WAIS-III administration and scoring manual*. San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1997). *Wechsler Memory Scale III (WMS-III)*. San Antonio, TX: The Psychological Corporation.

Weed, L. L. (1964). Medical records, patient care, and medical education. *Irish Journal of Medical Sciences, 6*, 271–282.

Weed, L. L. (1968). Medical records that guide and teach. *New England Journal of Medicine, 278*, 593–600.

Weed, L. L. (1969). *Medical records, medical education, and patient care*. Cleveland, OH: Case Western Reserve University Press.

White, D. K., Turner, L. B., & Turkat, I. D. (1983). The etiology of behavior: Survey data on behavior-therapists' contributions. The Behavior Therapist, *6*, 59–60.

Wolpe, J. (1977). Inadequate behavior analysis: The Achilles heel of outcome research in behavior therapy. *Journal of Behavior Therapy and Experimental Psychiatry, 8*, 1–3.

Wolpe, J. (1986). The positive diagnosis of neurotic depression as an etiological Category. *Comprehensive Psychiatry, 27*, 449–460.

Wolpe, J., & Wright, R. (1988). The neglect of data gathering instruments in behavior therapy practice. *Journal of Behavior Therapy and Experimental Psychiatry, 19*, 5–9.

Zubin, J. (1967). Classification of the behavior disorders. *Annual Review of Psychology, 18*, 373–406. Palo Alto, CA: Annual Review, Inc.

Zubin, J. (1984). Inkblots do not a test make. *\Contemporary Psychology, 29*, 153–154.

This Page Intentionally Left Blank

PART II

PSYCHOMETRIC FOUNDATIONS

This Page Intentionally Left Blank

CHAPTER 2

DEVELOPMENT OF A SCIENTIFIC TEST: A PRACTICAL GUIDE

Michael C. Ramsay and Cecil R. Reynolds

INTRODUCTION

Most authors portray test construction as a matter of carefully composing groups of items, administering them to a representative sample of people, and analyzing the responses using established statistical techniques. Many writers (e.g., Allen & Yen, 1979; Anstey, 1966; Kline, 1986; Robertson, 1990) lay out steps for the prospective test developer to follow. They often look something like this (adapted from Allen & Yen, 1979):

1. Develop a plan to cover the desired content.
2. Design items that fit the plan.
3. Conduct a trial administration.
4. Analyze the results, and modify the test if needed.
5. Administer the test again.
6. Repeat as necessary, beginning with Step 2 or 4.

The ensuing pages will, to a large extent, cajole and bludgeon the reader to follow this same time-tested trail to reliability and validity. Recent trends in test development, however, suggest that constructing a test requires more than close attention to the test itself. As growing numbers of psychologists and psychometricians tiptoe across the lines that once divided them, test developers are increasingly turning to mainstream research to test, and even to provide a basis for, the measures they develop.

The viewpoint that a good test should have a basis in empirical research, not theory alone (e.g., Reynolds & Bigler, 1995a, 1995b; Reynolds & Kamphaus, 1992a, 1992b), has gained in popularity. At a minimum, most test constructors would agree that a well-designed experiment can help explain the characteristics measured by a test (Embretson, 1985). Inevitably, such constructs as aptitude and schizophrenia do not respond readily to laboratory controls. Furthermore, obvious ethical constraints prevent scientists from manipulating certain variables, such as suicidal tendencies, and from holding others constant, such as learning rate. These considerations should restrain the influence of experimentalism on testing. Still, the foundations of testing have subtly shifted, and to some degree, the content of this chapter reflects this shift.

Theory, too, has played an important role in psychological test development. Before devising the groundbreaking Metrical Scale of Intelligence with Theodore Simon, Alfred Binet spent many years building and refining a concept of intelligence (Anastasi, 1986; Binet & Simon, 1916/1980). Soon after Binet and Simon released their scale, Charles Spearman (1927; 1923/1973; Spearman & Jones, 1950) presented a model of intelligence derived from correlational studies. Spearman (1904) posited four kinds of intelligence: present efficiency, native capacity, common sense, and the impression made on other people. Modern psychologists might recast these constructs loosely as achieve-

ment, ability, practical intelligence, and social intelligence. Thus, Spearman's model begins to resemble recent theories that include practical, tacit, and social intelligence (Gardner, 1983; Sternberg, 1985, 1990; Sternberg & Wagner, 1986). Notably, however, Spearman's model appears to omit creativity. Other contributors to the theory and modeling of mental ability include Guilford (1967, 1977), Cattell (1971), Thurstone (1938; Thurstone & Thurstone, 1941), Luria (1962/1980, 1966, 1972), Das, Kirby, and Jarman, (1979), and Kaufman and Kaufman (1983a, 1983b, 1983c). Personality testing, too, has a diverse theoretical base. Personality theories linked with testing include the *big five* personality factors (Costa & McCrae, 1992a, 1992b; John, 1990; Norman, 1963), trait or disposition theory (Mischel, 1990; Zeidner, 1995), Guilford's 14 personality dimensions (Guilford & Zimmerman, 1956) and Murray's manifest need system (Anastasi, 1988; Murray, 1938). The role of theory in test development becomes important in construct definition, Step 2 below.

FIRST GLIMPSES: TERMS, DEFINITIONS, AND CONCEPTS

Reynolds (1986) defines *measurement* as a set of rules for assigning numbers to objects, events, or actions. Reynolds goes on to define a psychological test as an instrument for applying these rules to behavior, whether overt or covert. The rules of measurement are the standardized procedures by which a measurement is taken so that it is reproducible. To determine the length of a rod, by which land was once measured, the king of England decreed that in each village, 10 men emerging from church—some tall, some short, some portly, some lean—would be taken and stood side by side. The distance from the beginning to the end of the line of men was the measure of a rod. Using the rod, a villager could measure a tract of land repeatedly and obtain the same result. Hence, the measurement was reproducible.

Assigning numbers according to rules is a part of everyone's life. We measure such varied dimensions as the height of a child, the force of an earthquake, and the earned-run average of a baseball player. Some characteristics we measure informally. For example, a shopper might measure a bagful of grapes by dropping it onto a grocery store scale. Other characteristics call for moderate formality, as when a nurse weighs a patient as part of an intake procedure. Finally, some measured properties demand elaborate techniques to ensure the best possible estimates. Psychological characteristics fall into this category. In every case, however, a person must follow certain rules to obtain a reproducible measurement.

In standardized testing, the manual provides the examiner with a set of rules for measuring performance: *Begin with item 5 for children ages 6 through 9; Stop after 4 consecutive failures; If an examinee gives an incomplete response, say, "Tell me more about it;" Allow 20 seconds for each item;* and so on. This reproducible set of procedures represents the rules of measurement that the examiner uses for a particular test.

Many instruments appear in popular publications accompanied by claims that readers can use them to assess their personality, their intelligence, their compatibility with their mates, and so forth. These informal inventories present qualitative descriptors such as, *If you scored between 20 and 25, you are highly self-aware.* These inventories offer no means, no standard scores, and no evidence of reliability, validity, or norms. (The ensuing pages explicate such concepts as reliability and validity). Accordingly, these scales do not constitute meaningful tests. Fortunately, such dime-store exercises are becoming relatively rare. However, Eyde and Primoff (1992) call attention to a new complication. Software billed as tests, but lacking scientific documentation, is becoming available to consumers.

Along different lines, courts sometimes treat so-called anatomically correct dolls as if they yielded authoritative results. Yet efforts to develop sound instruments based on these dolls are largely in the exploratory stages, and empirical support for the dolls remains equivocal, at best (Bauer, 1994; DeLoache, 1995; Skinner, Giles, & Berry, 1994). Additionally, anatomic dolls vary so widely in size, design, detail, proportion, and dress that if psychologists established the validity of one model, they would have little basis for applying the results to the others.

For the present, other introductory considerations merit attention. Most important, any test constructor should adhere to high ethical and technical standards. Rudner (1996) provides a brief outline for anyone in a position to evaluate a test. Additionally, educators and psychologists have worked together to produce *Standards for Educational and Psychological Testing* (American Educational Research Association, American

Psychological Association, & National Council on Measurement in Education, 1985). The *Standards*, under revision at the time of this writing, include a section on test construction and evaluation. *Guidelines for Computer-Based Tests and Interpretations* (American Psychological Association, 1986) have become available as well (See also Committee on Professional Standards & Committee on Psychological Tests and Assessment, 1986; Most & Zeidner, 1995). In a different vein, the steps involved in developing a test for commercial publication may also interest readers. Robertson (1990, 1992) describes this process.

STEP 1. REVIEWING THE LITERATURE

Every scientific study starts with an idea. So, too, does a scientific test. Many works on test construction create the impression that as soon as an idea springs to mind, the test developer should go to work busily charting the areas that the inchoate test should cover. Not at all! An aspiring test designer should turn first to the research literature to see how researchers are handling the construct in question, and how scientists or other professionals have measured this construct in the past. Both researchers and clinicians can guard against duplicating earlier efforts, and sidestep the rabbit trails that diverted their predecessors, by assessing the state of the field. PsycINFO (Psychological Abstracts Information Services) (APA, 1986) and other databases can speed the search process appreciably.

A review of the current literature can aid in constructing a test for local use, as well as for publication. The review can suggest improvements, or even uncover an existing instrument that may serve the test constructor's purpose without exacting the time, cost, and effort of developing a measure from step one. However, would-be test developers who find a measure like their own should not rush to jettison their plans. The measure may have an inadequate norming sample or low reliability and validity (see Step 8). Additionally, recent years have seen a profusion of tests that have undergone *partial* analyses, or whose empirical properties merely *approach* satisfactory levels. In fact, a number of prominent tests fit this description. Moreover, many neuropsychological tests have serious methodological flaws (Reynolds, 1989). For these reasons, a test developer willing to commit time and effort, and able to obtain funding, can fill important lacunae. This article should provide enough background to identify weakly and partially supported tests, given that their manuals present the necessary information. In some cases, a test's author or publisher can furnish needed information.

STEP 2. DEFINING THE CONSTRUCT

A complete definition of the characteristic of interest allows a test developer to systematically and thoroughly sample that construct. A definition can include behaviors, skills, deficiencies, problems, and traits that suggest the presence of the target characteristic. A definition may also include brief descriptors such as *quiet, energetic,* and *aggressive* (Walsh & Betz, 1995). For personality tests, the DSM-IV (American Psychiatric Association, 1994) can provide an excellent basis for a useful definition. The literature review should also inform the construct definition.

STEP 3. TEST PLANNING AND LAYOUT

For any study, researchers try to draw a representative sample from their population of interest. A *representative sample* is one that has the same levels of any relevant characteristics as the population does. For example, a scientist planning to study self-confidence might want a sample whose levels of self-esteem, a related trait, match those of the population. If the sample diverges from the population in self-esteem, it may also diverge from it in self-confidence. As a result, the scientist could not make inferences about the population's self-confidence based on that of the sample.

Like any scientist, the test constructor seeks to develop an instrument that measures a representative sample drawn from a population. This sample and population, however, consist of behaviors rather than people. Anastasi (1984, 1988) calls the population a *behavior domain.* It includes the entire range of behaviors that a test purports to measure. For example, a researcher may seek to measure verbal ability, a sweeping and ambitious goal. This researcher would try to write items that elicit representative verbal behaviors from the entire domain of verbal ability. Another researcher may want to explore the ability to find a small picture concealed in a larger one. Again, the items should sample the full range of this ability. This

Table 2.1. Specifications

	CONTENt[a]				
PROCESS[a]	VERBAL	FIGURAL	SPATIAL	VISUAL	AUDITORY
Reception	6	6	6	6	6
Retention	6	6	6	6	6
Encoding	4	4	4	4	4
Transformation	4	4	4	4	4
Elaboration	4	4	4	4	4

Note: Adapted, in part, from *Introduction to measurement theory,* by M. J. Allen and W. M. Yen, 1979, Monterrey, CA: Brooks/Cole Publishing Company.

[a]Content areas are listed across the top, and processes, down the side.

smaller, better defined domain could result in a sounder measure. In comparison with a behavior domain, a behavior *sample* includes all the behaviors that the items themselves actually cover. The items in a hidden-pictures test would cover most of the behaviors in its well-defined domain. In contrast, a measure of verbal ability would have to omit some behaviors.

The behavior sample, then, should represent the behavior domain. To meet this goal, a personality test should cover all important aspects of the characteristic or characteristics of interest. An ability test should measure every important content and process included in the behavior domain. For example, suppose that a clinician develops a cognitive processing test. The items might present auditory, verbal, figural, spatial, and visual material as content areas. As process areas, the items might sample test-takers' reception, retention, encoding, transformation, and elaboration.

After selecting the areas to measure, the clinician should design a table of specifications such as the one shown in Table 2.1. In a table of specifications, each row or column sums, giving the total number of reception items, the total number of retention items, and so forth. The clinician in the example shown has included equal numbers of items in each content area, but differing numbers can also be used. The importance of an area should determine the number of items it contributes. Accordingly, effective test constructors should acquire a thorough knowledge of the characteristics they want to measure (Golden, Sawicki, & Franzen, 1990).

Some tables of specifications make use of Bloom's influential *taxonomy of educational objectives* (Bloom, 1956; Robertson, 1990). Bloom included six processes in the taxonomy: knowledge, comprehension, application, analysis, synthesis, and evaluation. Of these, Bloom saw knowledge as the most basic. Each subsequent process became progressively more advanced.

Krathwohl, Bloom, and Masia (1964) designed a taxonomy of affective objectives, such as willingness to follow rules and appreciation of literary classics. Psychometricians tend to overlook this work, perhaps in part because it includes attending as the first affective process, but draws only a weak distinction between this process and cognition. The affective character of attention deficit disorders, however, suggests that the inclusion of attending in an affective taxonomy has merit. Indeed, attention regarded as an affective process has a long history in psychology (e.g., Wundt, 1896/1907, (1906/1912). For their part, Krathwohl, Bloom, and Masia (1964) argue that people must attend to a rule, or a work of art, for instance, before they can value it, internalize it, and so on.

A second difficulty with this affective taxonomy lies in its far-reaching advanced objectives. Educators may not need to instill in their students an outright devotion to art or a dedication to following rules. Additionally, attending and responding, the second affective objective, may be difficult to measure with a classroom test, even if an educator were in a position to assess and treat attention difficulties. Still, clinicians and clinical test developers might find affective taxonomy useful. Additionally, test developers who want to measure attitudes can derive insight from Krathwohl, Bloom, and Masia's (1964) treatment of the last three objectives: valuing, organizing one's values into a system, and developing a unified *weltanschauung,* or worldview, that pervasively controls one's behavior.

Ironically, the wide-ranging applicability of Bloom's cognitive taxonomy also argues against

its affective counterpart. A cognitive taxonomy can apply as much to students' understanding of music or art as to their comprehension of history or science. Ultimately, many test users can measure cognitive and affective material using one kind of taxonomy.

STEP 4. DESIGNING THE TEST

Instructions and Demographic Information

For a test developer, the instructions given to test takers should carry as much weight as the items themselves. Kline (1986) presents basic rules for writing instructions. They appear here in an adapted form. First, make the instructions brief. Next, make them as simple, clear, and free from modifying clauses as possible. Present examples only if they clarify the instructions. Kline (1986) suggests interviewing test takers who have failed items, possibly during item tryout, to find out what they were trying to do. Every section of a test should begin with a set of intructions, even if they are basically the same for each section. Phrases such as *as before* and *as above* can make the repetition palatable.

Test constructors should ensure that a test is not "biased or offensive" on grounds of race, sex, height, native language, ethnicity, geographic region, or some other characteristic (Rudner, 1996, p. 3). Most test forms include a demographic section containing items such as *Sex: M, F* and *Ethnicity: Black, White, Hispanic, Asian, Other.* With ability measures, however, demographic items that remind test takers of their group membership may interfere with their performance by threatening their self-esteem. In a series of studies, Steele and colleagues (Spencer, Josephs, & Steele, 1993; Steele & Aronson, 1995) found that African Americans scored lower than whites on an ability test when asked to indicate their ethnicity. However, the two groups scored about the same with the ethnicity item removed. This item may have generated its effect by threatening African American test-takers' self-esteem in a phenomenon called *stereotype threat.* One solution to this difficulty might be to place items that may evoke stereotypes *after* the ability items. Researchers have not yet identified the possible effects of stereotype threat on personality test results.

The Manual and Directions for Test Users

In most cases, the test constructor should develop an extremely brief set of directions for administering and scoring, and place any elaborations in the manual. The directions should remain unambiguous and clear throughout. They should provide complete enough coverage that all examiners can give the test in the same way. Examiners should not have to rely on guesswork on even a seemingly trivial point. Indeed, every aspect of a test and its administration should be *standardized,* that is, the same for all test-takers. If coaching is permissible on the first few items of a subtest, the manual should include complete instructions for doing so. Additionally, the administration directions should closely match the conditions under which the members of the norming sample took the test (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1985; Anastasi, 1988; Rudner, 1996). Also, directions that duplicate commonly used procedures can ease the the examiners' mental strain. Training sessions for any colleagues giving a test can help to uncover and remedy areas of uncertainty. The test developer should try to anticipate and account for any difficulties that might occur later, during administrations of the final version.

Besides detailed instructions, a clear explanation of the appropriate use and interpretation of a test should also appear in the manual. In particular, the manual should specify inappropriate uses, such as inferring that an examinee has clinical depression from results on a test of normal depression. Additionally, the manual should accurately describe the test's statistical quality, as addressed in this chapter, and identify the populations for whom the test was normed. In general, a manual should be clear, complete, accurate, and understandable (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1985; Rudner, 1996).

Ability Tests

Item type. Important considerations in designing items for a test include item *type* or *format* and item difficulty. An ability-test developer has several item types to choose from. The *free-response* or open-ended item type places relatively few restrictions on possible responses. Essay and short-

answer items fall into this category. The familiar essay question calls upon test takers to respond in paragraph or sentence form. Examples might include "Describe the three stages of the Egyptian empire" and "Could the moon support life? Why or why not?" Over time, the rubric *essay question* has come to refer rather loosely to any free-response item. The short-answer item calls for brief responses, such as a word or two, or perhaps one or two sentences. Short-answer items might include "A _____ surrounds the axons of many neurons" and "Olive Schreiner's tale, _____, recounted a villager's lifelong search for truth."

The *fixed-response* or closed-ended item type, also called the objective item type, presents all permissible responses to the test taker, who chooses among them. Naturally, this arrangement limits the number and content of possible answers. These constraints allow the test user to score most responses as correct or incorrect without resorting to subjective judgment. This objectivity is a hallmark of the fixed-response format. Examples of this item type include multiple-choice, true-false, matching, ranking, and rearrangement items (Anastasi, 1988). Fixed-response items produce more agreement among scorers, or *interrater reliability,* than their older, open-ended counterpart. This is so because the test's author specifies the right answers unambiguously before administration. As Anastasi (1988) notes, fixed-response items have additional advantages over their free-response counterpart. They produce a better measure of the target characteristics; they take less time to complete and score; the smaller, briefer items permit wider coverage; and this improved coverage reduces error due to chance, making the test fairer to the individual.

Potential test authors will find the paper-and-pencil format most economical and easiest to work with, but other item types are possible. An observer might have difficulty classifying many items as multiple-choice, short-answer, and so on. A number of intelligence, memory, and neurological tests, for example, have innovative formats. The digits subtests found in major intelligence and memory batteries (Ramsay & Reynolds, 1995) require the test taker to repeat a series of digits in the order presented. Clinicians present the correct response only, a kind of ultra-closed-ended feature. However, test takers can repeat the digits in any order as an incorrect response, and they can and do respond with digits not presented. If items fall on a continuum from closed- to open-ended, as Pedhazur and Schmelkin (1991) suggest, then digits tasks lie in the midrange of that continuum.

Intelligence batteries (e.g., Kaufman & Kaufman, 1983a, 1983b, 1983c; see also Willson, Reynolds, Chatman & Kaufman, 1985) also include many puzzle-like activities, fixed-response subtests that call upon test-takers to rearrange shapes or blocks to match a model. Finally, intelligence, memory, and achievement tests often include subtests that make use of pictures. Many of these tasks have a fixed-response format. Elaborate batteries that incorporate many ability tests have few successful competitors. Smaller, separate tests along the same lines may find their niche more readily.

Item Difficulty

For beginners, the most common error in designing ability items lies in making them too difficult. Novice test-designers often find that most of their subjects have obtained abysmal scores. This result means that the items are not working properly. The test is uninformative. It fails to distinguish, or discriminate, between individuals. The opposite problem, making a test uninformative by making the items too easy, can also present itself. Test designers can gauge the difficulty of their items in advance by administering them to friends, colleagues, and others. Techniques for calculating item difficulties following the tryout phase appear in Step 6.

Items: Personality Tests

For personality tests, key concerns in item writing include item type and item *attractiveness*, a counterpart of item difficulty. Personality tests as ability tests, employ two types of items, free- and fixed-response. Most present-day personality tests use a fixed-response item type, but some of the most intriguing have free-response items.

Item type

Fixed-response items defy easy classification. Here, they are broken down into four groups:

dichotomous, rating-scale, midpoint, and *forced-choice.* Each of these item types takes many forms.

Dichotomous items force test takers to choose between two alternatives, such as *true* and *false, yes* and *no, like* and *dislike,* or *describes me* and *does not describe me.* Examples of dichotomous items include "I often get angry. T, F" and "Lawbreakers can change. Agree, Disagree." The Minnesota Multiphasic Personality Inventories (e.g., MMPI-2, Butcher, Dahlstrom, Graham, Tellegen, & Kaemmel, 1991), the Millon Clinical Multiaxial Inventories (e.g., MCMI-II; Millon, 1987), and parts of the Behavior Assessment System for Children (BASC; Reynolds & Kamphaus, 1992a, 1992b) use dichotomous items.

Rating-scale items might better be labeled continuous items; their responses fall on a rough continuum. A rating-scale item may call upon test takers to respond on a five-point scale from *agree* to *disagree,* with only the two extremes labeled. Another test might feature rating-scale items with every response point labeled. For example, the item "I enjoy making jokes about people" might include the response choices *rarely, sometimes, often,* and *very often.* More specific response options, like *once a week or more,* can boost validity. In still other rating-scale tests, each item has its own set of responses. Some measures even omit the item stem. Rating-scale tests include the State-Trait Anxiety Inventory (STAI; Spielberger, 1977) and the Beck Depression Inventory (BDI; Beck, 1978), which omits the stem. Items with midpoints have an odd number of response choices, with a central point marked *neutral, undecided, neither agree nor disagree,* or some other variant.

Midpoints tend to attract large numbers of responses, making a scale uninformative. The Strong Vocational Interest Blank (SVIB; Hansen & Campbell, 1985) employs a three-point scale with the midpoint labeled *indifferent.*

The forced-choice item type is really a test format. Many forced-choice tests present paired sentences. For each item, test takers select the sentence that best describes them. Other tests of this type present three or more alternatives. For each item, test takers select the most and least applicable alternatives. In forced-choice tests, the sentences appear repeatedly, in different pairs or groups. The Myers-Briggs Type Indicator (Myers, 1987) makes use of the forced-choice technique.

The most intriguing examples of free-response personality tests are *projective* tests. These tests present vague stimuli, such as inkblots or shadowy pictures, to the test taker. Since the stimuli themselves are nondescript, the content of a response must presumably come from the test taker. Thus, projective tests presuppose that people unknowingly disclose their conflicts, drives, or cognitions in their answers to nondirective items (Kaplan & Saccuzzo, 1989). The best-known projective test, the Rorschach test (Rorschach, 1921; Exner, 1974), consists of colored inkblots. Clients view each inkblot and tell the clinician what they see in it. Other projective tests have clients view a drawing and tell a story about it, or draw a person, scene, or object and describe what they have drawn. Besides projectives, free-response items can also take the form of incomplete sentences, often with only one or two words provided. The Rotter Incomplete Sentences Blank (RISB) (Rotter & Rafferty 1950; Rotter Incomplete Sentences Blank, 1977) is a widely known example. Its items include "My greatest fear _____" and "A mother _____."

Item Attractiveness

Excessive attractiveness causes difficulty for many personality-test designers. An item's *attractiveness* is its likelihood of eliciting a positive response, such as *yes* or *true.* People tend to agree with an item. The statements "I like my neighbors" and "I dislike my neighbors" would both draw many positive responses. A test developer should rephrase or discard items that most people are likely to agree with.

Items: All Tests

The author of any test may find the following suggestions helpful. First, have a colleague review the test for clarity, or set it aside and review it yourself later (Anastasi, 1988). Avoid using difficult words, unless the test measures abilities like vocabulary and word recognition. Similarly, eschew complex grammatical constructions. Replace emotionally loaded language with neutral terms. For example, the item "I suffer from debilitating anxiety" might be reworded, "My anxiety interferes with my day-to-day activities." This particular paraphrase has the added advantage of being specific, an important consideration (Kline, 1986). Similarly, rephrase items that imply bias

toward ethnic and other groups (Rudner, 1996). To the extent that a test measures unintended characteristics, such as reading ability or ethnic sensitivity, it fails to measure the intended ones. Finally, a measure with more than one item type complicates item analysis (Qualis, 1995).

Empirical Item Writing

In contrast with the comparative rigor of item analysis, many test constructors still treat item writing as an art. Yet research suggests that several factors relevant to item writing affect test performance (Willson, 1989; Willson, Kulikowich, Alexander, & Farrel, 1988). An enterprising test designer can take steps to account for such influences. Researchers have investigated item components (e.g., Alexander, Willson, White, & Fuqua, 1987; Butterfield, Nielson, Tangen, & Richardson, 1985; Willson, Goetz, Hall, & Applegate, 1986), motivation (Atkinson, 1980; Hill, 1980), prior knowledge (Alexander & Judy, 1988), short-term memory (Chase & Simon, 1973; Willson, Goetz, Hall, & Applegate, 1986; Willson & Olivarez, 1987), and self-monitoring (Flavell, 1985). Some researchers (e.g., Birenbaum & Tatsuoka, 1987; Roid & Haladyna, 1982) have generated items empirically. An approach involving several classes of incorrect responses (Willson, Kulikowich, Alexander, & Farrell, 1988) led to the successful classification of 6th-grade students (Goetz, Kulikowich, & Alexander, 1988). While these methods have been used chiefly with achievement tests, adaptations to ability and personality tests appear feasible.

Layout

The overall organization of a test can influence its effectiveness in measuring the intended characteristics. Naturally, the test form should be readable, clear, and free from distractions. The length and partitioning of an ability test can affect its difficulty level. As such, longer tests should be broken down into sections of perhaps 10 to 20 items, depending on the age groups tested. Ability tests should begin and end with a few easy items (Allen & Yen, 1979). Interviews with the initial test takers can generate feedback regarding a test's length and difficulty.

The number of items to include in a test should receive careful consideration. The initial version of the test should have 1 ½ to 2 times as many items as the final product, because weak items are later discarded. Yet test takers in the tryout sample should not be overburdened with items. To resolve this dilemma, test developer can distribute the items over two or three test forms for item tryout, then combine the forms for the final version of the test. Naturally, tests for young children and for students in the lower grades should have relatively few and easy items (Robertson, 1990).

A test developer should strive to minimize any adverse impact that a test may engender. Recall and retrieval research (Anderson, 1990, 1995) suggests that personality tests should begin with positive, reassuring items. Convention holds that disturbing or negative items should be rephrased or replaced, where possible. If needed, they should alternate with positive items and appear chiefly toward the middle of the test form. The test should conclude with its most upbeat items. In general, a test should not instill adverse experiences into test-takers' memories. Test-layout issues, however, have received little research attention. A negative item that elicits a response of *false* may evoke positive feelings. Conversely, a positive item may need to be fairly easy to endorse if it is to elicit a positive reaction.

STEP 5. ITEM TRYOUT

The Tryout Sample

In time, the diligent test-designer should have a completed instrument attractively set in readable type. If you have produced a well-designed test form or forms, you can now administer them to a representative sample of people, called a development or *tryout* sample (Walsh & Betz, 1995). A randomly selected sample of sufficient size eliminates chance error. Additionally, the tryout sample should match the final, target population on relevant characteristics (Rudner, 1996). For example, many test developers include people from all major geographic regions. You may need to take this approach if you are measuring characteristics that could vary from region to region. If you plan to explore the effects of ethnicity on your test, as is done in bias studies, you will need to oversample

minority groups to ensure that you have large enough samples to generalize to their respective populations (Robertson, 1990; Rudner, 1996). Small samples can yield results that differ from those that the population would generate. For example, tryout results might suggest that Native Americans have below-average spatial ability, when in fact they score high in this ability.

If you plan to evaluate your test using only your overall sample, having a large overall sample will suffice. However, if you plan separate analyses for each gender or ethnicity, you will need large samples of every such group that you include. Authors disagree somewhat on what number qualifies as large. If possible, select samples of 200 to 500 people (Henrysson, 1971; Robertson, 1990; Thorndike, 1982). Allen and Yen (1979) permit a tryout sample as small as 50. Kline calls a sample of 100 "the absolute minimum" (1993, p. 136). This standard may mean thoughtfully reducing the number of groups that you analyze separately.

Besides sample size, missing responses can also influence the generalizability of a test. If many subjects fail to complete a test, the ones who remain may differ from the target population. As mentioned above, random differences should cancel each other out with randomly selected samples. Systematic differences, however, can distort test-takers' results.

Ages and Grade Levels of Tryout Samples

Robertson (1990) describes additional concerns that a test constructor should attend to in the item tryout-phase. First, try out the test with people of the same grades or ages as the target population (Rudner, 1996). Include enough age or grade levels to make an informed decision to retain, discard, or alter an item. However, you may be able to try out an adult test with only a few age groups, or even one. This is so because adults show less marked developmental changes than children for many characteristics. Of course, you should familiarize yourself with research rather than simply assume that adults fall into only one group.

Training Tryout Examinees

Some test constructors have confederates conduct their item tryouts. These confederates, like all examiners, should have appropriate qualifications. If the test requires little activity, for example, reading brief instructions and passing out test forms, any capable person can administer it. At the opposite extreme, administering intelligence tests typically requires a master's degree, or its equivalent, and training in intelligence testing. Many personality measures, though easy to administer, require advanced knowledge and training to interpret the results. In general, the training needed varies with test type. Most well-designed tests have detailed manuals with a section devoted to examiner qualifications. A reading of this section can acquaint prospective test designers with the qualifications an examiner should have.

Informed consent

With few exceptions, anyone administering a test should obtain informed consent from the test takers. Informed consent means apprising test takers of (a) the types of tests to be administered, (b) the reasons for the testing, (c) the intended uses of the results, and (d) the possible consequences of that use. The disclosure should also specify (e) the parties who will receive information about the testing, and (f) the information that they will obtain. Children and people with mental retardation should receive a simple explanation that will permit them to provide informed consent (American Education Research Association, American Psychological Association, & National Council on Measurement in Education, 1985; American Psychological Association, 1982).

Other Considerations at Tryout

With time-limited tests, establish a separate time limit for each section. Robertson (1990) suggests setting each time limit so that 90 percent of the examinees can complete the entire section. Testing a few people before item tryout and noting how long they take is helpful. Robertson also points out that if a test is meant for a specific season or time of year, as with achievement tests

given in the Fall, item tryout should also occur at that time.

STEP 6. ITEM ANALYSIS

Discriminating Power

A test should convey as much information as possible about differences between test takers on the characteristic being measured. People with high levels of this characteristic should earn high scores on the test. People with low levels should earn low scores. For most tests, scores should also spread out widely across the range of possible scores, rather than clump together at one or two points. When a measure meets such standards, psychometricians say that it has high *discriminating power* (Kline, 1986), that is, it distinguishes well between people with high or low levels of the characteristic being measured. A good test discriminates among, not against, test takers.

Item Statistics: Classical True-Score Theory

To estimate a test's discriminating power, the test developer can calculate item statistics. The item statistics explained next derive from classical true-score theory. Statistics grounded in this theory afford the advantage of *weak assumptions.* Test responses can meet such assumptions, or conditions needed to ensure accurate results, fairly easily (Hambleton & Swaminathan, 1985). Classical item statistics include (a) indices of item difficulty for ability tests, or item attractiveness for personality measures, and (b) either item-discrimination indices or item-total point-biserial correlations. Point-biserial correlations are appropriate when correlating a continuous variable with a dichotomous variable, such as an item scored *correct* or *incorrect.*

Item Difficulties and Attractiveness Indices

An item's difficulty affects its discriminating power. If an item is too difficult, even people fairly high in the targeted characteristic will respond incorrectly. Thus, the item fails to distinguish between people high and low in the characteristic. The item does furnish information about the few people who responded correctly, but not about the majority who did not. An excessively easy item also fails to discriminate. Test scores should vary considerably from one individual to another (Rudner, 1996). This principle also applies to personality tests. An item that is too attractive, or too unattractive, fails to identify people who have the desired characteristic.

The item-difficulty index represents the proportion of test takers, p_i, who give the *keyed,* or designated, response (Kline, 1986; Robertson, 1990). For ability tests, this response is simply the correct alternative. For personality tests, the response itself suggests that the test taker has the trait of interest. When applied to personality measures, p_i might best be called the *item-attractiveness index.*

To calculate p_i, the number of test takers who give the keyed response by the total number of test takers in the sample. For example, suppose an examiner gives an ability test to a tryout sample of 100 people. Of these, 30 people respond correctly to a certain item. This item, then, has a difficulty of 30/100 or .3. However, 75 people respond correctly to another item. That item's difficulty is .75. In another example, 50 people take a self-esteem test. Of these, 23 people respond *true* to the item, "I take great pride in my achievements". Item attractiveness is 23/50, or .46.

The test constructor takes great pride in this figure, because an item with an attractiveness index near .5 provides more information about individual differences than one with a high or low value. In other words, the item can make more differentiations between test takers. If the item read only, "I take pride in my achievements," nearly everyone would endorse it. The resulting p_i might reach .80 or .90. The item would convey scant information (Allen & Yen, 1979; Anastasi, 1988).

Item difficulties of .5, however, do present a few problems. First, any item with two response options would possess this difficulty level if every test taker merely guessed randomly at the answer (Allen & Yen, 1979). According to Lord (1953, 1953b; Allen & Yen, 1979), objective items should have an average difficulty a little below the midpoint between 1.0 and their chance success levels. An item's *chance success level* (CSL) denotes the difficulty level that would result if all test-takers guessed at the answer, or otherwise responded randomly. An item's CSL is simply $1/n$, where n signifies the number of

response options. The optimal difficulty level lies just under (CSL + 1.0)/2. For example, an item with three response options would have a chance success level of .33. The best average difficulty for such items would be a little under (.33 + 1.0)/2 or .67. Thus, a difficulty level of roughly .65 would be optimal. For a two-response true-or-false item, the chance success-rate would, again, amount to $1/n = .5$. This figure would yield a best average difficulty just under (.5 + 1.0)/2 = .75, or about .72 (Allen & Yen, 1979).

In practice, however, item difficulties should vary. If all items had difficulties of exactly .50 and were perfectly intercorrelated, the same 50 percent of test takerss would answer every item correctly. These all-or-none results would permit examiners to make only the grossest distinctions among individuals. Generally, items should have a range of difficulties that centers on the best average difficulty. If this figure is about .50, difficulties should fall roughly between .30 and .70. The range should be larger for more highly intercorrelated items (Anastasi, 1988). Allen and Yen (1979) describe an exception, where a decision has been made to accept everyone at or above a predetermined score, called a *cutting score*. For example, an employer might decide to interview every applicant who scores above 75 on an in-house test. Ideally, the items included in this test should have difficulties of 50 percent, when completed by people whose total scores equal the cutting score.

The Item-Discrimination Index

Test developers can choose from more than 50 indices of discriminating power. In practice, however, most of these indices convey essentially the same information (Anastasi, 1988). Additionally, the indices presented below have desirable mathematical properties and enjoy widespread acceptance. First, the *item-discrimination index* denotes the difference between the number of high and low scorers who have given the keyed response. The formula shown below yields this index.

d_i = item-discrimination index

$$d_i = n_{hi}/n_h - n_{li}/n_l$$

n_{hi} = number of test takers in the high-scoring group who passed item i

n_h = number of test takers in the high-scoring group

n_{li} = number of test takers in the low-scoring group who passed item i

n_l = number of test takers in the low-scoring group

The notation is explained in terms of ability, but the formula also applies to personality tests. The high- and low-scoring groups mentioned are the test takers earning the upper and lower 27 percent of total scores (Anastasi, 1988; Kline, 1986). These groups yield a serviceable estimate of d_i, and the best estimate when total scores have a normal distribution (Allen & Yen, 1979; Kelley, 1939).

If the high- and low-scoring groups are equal in number ($n_h = n_l$), the formula becomes a little simpler:

$$d_i = (n_{hi} - n_{li})/n_r,$$

where $n_r = n_h = n_l$, and r stands for *range*. To illustrate this formula, assume that a tryout sample of 200 people take a spatial analogies test. The upper and lower ranges each include 200(27%) = 54 people. In the high-scoring group, 30 people give the keyed response to item i, compared with 9 people in the low-scoring group. In this example, the item-discrimination index equals

$$\begin{aligned} d_i &= (n_{hi} - n_{li})/n_r \\ &= (30 - 9)/54 \\ &= 21/54 \\ &= .39, \end{aligned}$$

a lackluster showing. However, if 39 people in the high-scoring group gave the keyed response, compared with only 6 people in the low-scoring group,

$$\begin{aligned} d_i &= (39 - 6)/54 \\ &= 34/54 \\ &= .63, \end{aligned}$$

an impressive performance for a single item.

The item-total point-biserial correlation. If they wish, test constructors can use this statistic in place of the item-discrimination index, described above. Both indices reflect the number of test takers who obtained a high score and also answered item i correctly. The item-total correlation makes use of the test's standard deviation. Any elementary statistics text explains how to calculate this value. Additionally, computer packages like SAS or SPSS can out-

put a standard deviation with very little effort, if the user has a basic familiarity with them.

The formula below yields the item-total correlation (Allen & Yen, 1979):

$$r_{iX} = (X_i - X)/s_x \sqrt{p_i}/(1 - p_i)$$

r_{iX} = correlation between scores on item and total scores

X_i = mean of total scores of all test takers passing item *i*

X = mean of all total scores

s_x = standard deviation of all total scores

p_i = item difficulty

Suppose that 300 students took a hyperactivity test. The students who passed the item being analyzed obtained an average total score of 65. The average total score of the entire group amounted to 60. The standard deviation was 10, and the item's difficulty reached .62. For this group,

$$\begin{aligned} r_{iX} &= (65 - 60)/10 \sqrt{.62/(1-.62)}. \\ &= 5/10 \sqrt{.62/.38} \\ &= .5 \sqrt{1.63} \\ &= .5(1.28) \\ &= .64, \end{aligned}$$

an excellent result for a single item.

The item-total correlation produces a somewhat inflated estimate of an item's discriminability, because the test-takers' total scores include their scores on the item being examined (Kline, 1986). An item that completely fails to measure the intended characteristic generates a positive item-total correlation nonetheless, because, like any other variable, the item correlates with itself. For tests of more than about 100 items, however, little distortion occurs with the point-biserial correlation. Kline (1986, p. 139) calls the point-biserial item-total correlation, presented here, the "best measure" of an item's correlation with the test's total score.

Item-discrimination statistics such as those described above should yield a positive value for any defensible item. A negative value means that the item behaves opposite the overall test. In the case of a large, negative figure, the test developer can score the item in reverse. Thus, if an item is scored *a = 1, b = 2, c = 3,* the rescoring would yield *a = 3, b = 2, c = 1.* This maneuver produces a large, positive figure, the desired result. The item may simply have been coded incorrectly. If an item has a low value, whether positive or negative, the problem could lie with its wording, its scoring, or both. The test's author may recode or rephrase such items, or discard them, saving the time, expense, and effort of a follow-up administration.

Item Response Theory (IRT)

The first hints of item response theory, also called *latent-trait analysis* and *item-characteristic curve theory,* emerged in the mid-1930s and late 1940s. By the 1970s, IRT had come to predominate in the work of measurement specialists. Works by Wright (e.g., Wright & Panchapakesan, 1969; Wright & Stone, 1979), Bock (1972; Bock & Lieberman, 1970), and especially Lord (1952; 1953a, 1953b; 1968; 1977a, 1977b; Lord & Novick, 1968) gave impetus to this alternative to classical true-score theory (Hambleton & Swaminathan, 1985). In time, Lord (1980) wrote the influential volume, *Applications of Item Response Theory to Practical Testing Problems.* Later, Hambleton and Swaminathan (1985) produced a comprehensive textbook on IRT. A subgroup of IRT models developed by Rasch (1960, 1980; Fischer & Molenaar, 1995) have become popular in their own right. Finally, Samejima (1983) developed a series of item-response models for small data sets.

IRT represents a family of models called *item response models* and the measurement theory that they all have in common. These models describe the mathematical relationship between an observed characteristic, measured by a test, and the underlying, latent characteristic that gives rise to it (Weiss, 1983). Item response models describe how the latent characteristic influences test-takers' performance. A successful model allows the test user to estimate an individual's standing on this characteristic (Allen & Yen, 1979; Hambleton & Swaminathan, 1985).

No one can observe or measure this hypothetical characteristic directly. Indeed, the word *latent* in this context means not directly measurable (Hambleton & Swaminathan, 1985). Despite the term

latent-*trait* analysis, the characteristic can take the form of (a) an ability, or (b) a trait, that is, a personality characteristic. In IRT, this ability or trait can take on an infinite range of values. In classical theory, it cannot. IRT also differs from classical models in a second way. IRT does not presuppose that the test takers' true score on the latent-trait or ability equals their expected observed score on the test. In fact, the true score is not even a linear function of the observed score. However, computer programs can generate estimates of the latent characteristic (Allen & Yen, 1979).

IRT users commonly assume that a single, dominant trait or ability explains test takers' performance on a given test (Hambleton & Swaminathan, 1985). The relationship between a sample's (or a test taker's) trait or ability level and the test taker's performance on an item is called the *item-characteristic function.* IRT users often plot this function using an *item-characteristic curve* (ICC) like the one shown in Figure 2.1. In an ICC, the Greek letter theta, θ, symbolizes the sample's trait or ability level, represented by their total scores on the test. Similarly, $p_i(\theta)$ symbolizes the item's difficulty level, represented by the proportion of test takers who passed the item. In Figure 2.1, 60 percent of test takers with a total score of 5 passed item *i.* Put differently, the test takers who scored 5 had a 60 percent chance of passing this item. The test takers who scored 6 had a 75 percent chance of passing the item (Allen & Yen, 1979).

A steep ICC indicates that the item discriminates well, a shallow slope, that it discriminates poorly. Also, an ICC provides b_i, an index of item difficulty that increases, rather than decreases, with difficulty. This index is simply the total test score that corresponds to a $p_i(\theta)$ of .50. Thus, in Figure 2.1, b_i is about 4.5. Transforming characteristic levels to give them a mean of zero and a standard deviation of 1 allows a test constructor to compare the b_i indices of two or more measures. The transformed b_i values usually range from +2 to −2 (Allen & Yen, 1979; Hambleton, Swaminathan, & Rogers, 1991). ICCs can alert a test constructor if an item is functioning differently for different groups of people (Allen & Yen, 1979). If the item has different ICCs for different groups, it is measuring different characteristics for the two groups. In such cases, the item is not working properly for at least one group. On the other hand, if the ICCs look the same, the item *might* be working appropriately for all groups. A comparison of ICCs can identify a bad item but not a good one. It is still useful, however, because it can signal the test developer to discard certain items.

Obtaining an ICC entails estimating item parameters and latent-trait values. This process is complex, and most test constructors simplify it by using a computer. Hambleton, Swaminathan, and Rogers (1991) list several sources of IRT computer programs, including BICAL, RASCAL, RIDA, LOGIST, BILOG, and MIRTE. Additionally, many universities offer courses that provide a basic foundation in data analysis. Inquirers should ensure that the course they are contemplating covers the kinds of analyses they plan to conduct.

STEP 7. BUILDING A SCALE

After obtaining item-analysis results, the test developer categorizes them according to their statistical properties. As mentioned above, items with moderate difficulty and high discriminability are best. The items are typically sorted into three categories: (a) items with acceptable statistics, (b) items with marginal statistics that could become serviceable with revision, and (c) statistically weak items that should be discarded (Most & Zeidner, 1995).

Many test developers use factor analysis for building a scale. As Kline (1993) points out, sterling-item statistics can simply mean that a test measures the wrong characteristic reliably and discriminably. Test constructors can sidestep this problem by factor-analyzing their test with another test, one already known as a good measure of the targeted characteristic. The next section addresses factor analysis.

Items judged acceptable become part of the scale or scales, if more than one scale has undergone analysis or emerged during factor analysis. At this point, a second consideration, item arrangement, comes into play. Step 3 introduced a number of issues pertaining to item arrangement. A few more receive attention here.

Intelligence tests typically consist of several subtests, each containing a set of similar items. For example, the widely used Wechsler (Wechsler, 1981, 1991a, 1991b) and Kaufman batteries (Kaufman & Kaufman, 1983a, 1983b, 1983c) have separate scales for arithmetic, vocabulary, mazes, immediate memory, and other items. For intelli-

Table 2.2. Facsimile of SAS Output Showing Cronbach's Alpha Results for a Five-Item Test

CRONBACH COEFFICIENT ALPHA
FOR RAW VARIABLES: .712769
FOR STANDARDIZED VARIABLES: .721678

	RAW VARIABLES		STD.[a] VARIABLES	
DELETED VARIABLE	CORRELATION WITH TOTAL	ALPHA	CORRELATION WITH TOTAL	ALPHA
ITEM1	.678923	.823774	.696623	.833642
ITEM2	.886435	.486724	.876037	.494759
ITEM3	.713067	.711307	.723402	.715638
ITEM4	.428661	.954231	.428661	.974375
ITEM5	.554268	.874024	.564876	.883756

Note: [a]Standardized.

gence and other ability tests, the first few items should be easy enough to encourage weaker test takers. To distinguish among even the ablest test takers, item difficulties within a subtest should gradually increase, so that only a few people pass the concluding items. The items in a subtest should vary enough to hold the test-takers' attention, unless the subtest measures attention.

Personality measures often consist of one form, with the items of all its subtests scattered indistinguishably throughout it (Most & Zeidner, 1995). Items arranged by subscale could reveal the nature of the subtests and introduce bias. The MMPI batteries (e.g., MMPI-2; Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1991; Hathaway, McKinley, Butcher, Dahlstrom, Graham, & Tellegen, 1989; see also Dahlstrom, Welsh, & Dahlstrom, 1972, 1975) present a single test form containing several hundred items that represent 13 basic scales. Dividing the form into sections would help avert fatigue effects. These sections would not reflect the traits measured, since examinees could then infer the nature of the subtests.

STEP 8. STANDARDIZING THE TEST

After assembling the items into scales, the test developer, or the publisher, sets the test in its final form, the form that will be administered and perhaps marketed. Next, qualified examiners administer the test to a second representative sample called a norming or *standardization sample*. This sample and administration should fulfill the criteria described for the tryout administration. For example, the sample should be large, and its characteristics, such as gender and ethnicity, should be specified (Reynolds, 1989). This administration should take place under the same conditions as the actual, day-to-day use of the final version. The test-takers' responses then become the basis for computing statistics that estimate the test's reliability and validity. If the test has sufficient reliability and validity, its author can calculate percentiles, standard scores, and other statistics needed by test users. If not, the next step is to return to item writing or item analysis.

Reliability and Validity Indices

The indices explained above describe test items. Most test developers obtain them during item analysis. By contrast, the indices presented next describe the test itself. A test developer obtains them during standardization and, if desired, during item analysis. A test demonstrates *reliability* to the extent that it displays consistency across times, conditions, scorers, items, or test forms. Reliability is essentially freedom from error. A test demonstrates *validity* to the extent that it measures what it purports to measure (Anastasi, 1988; Most & Zeidner, 1995; but see Stewart, Reynolds, & Lorys-Vernon, 1990 for an alternative definition). A valid test supports inferences that are "appropriate, meaningful, and useful" (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1985). See Pedhazur and

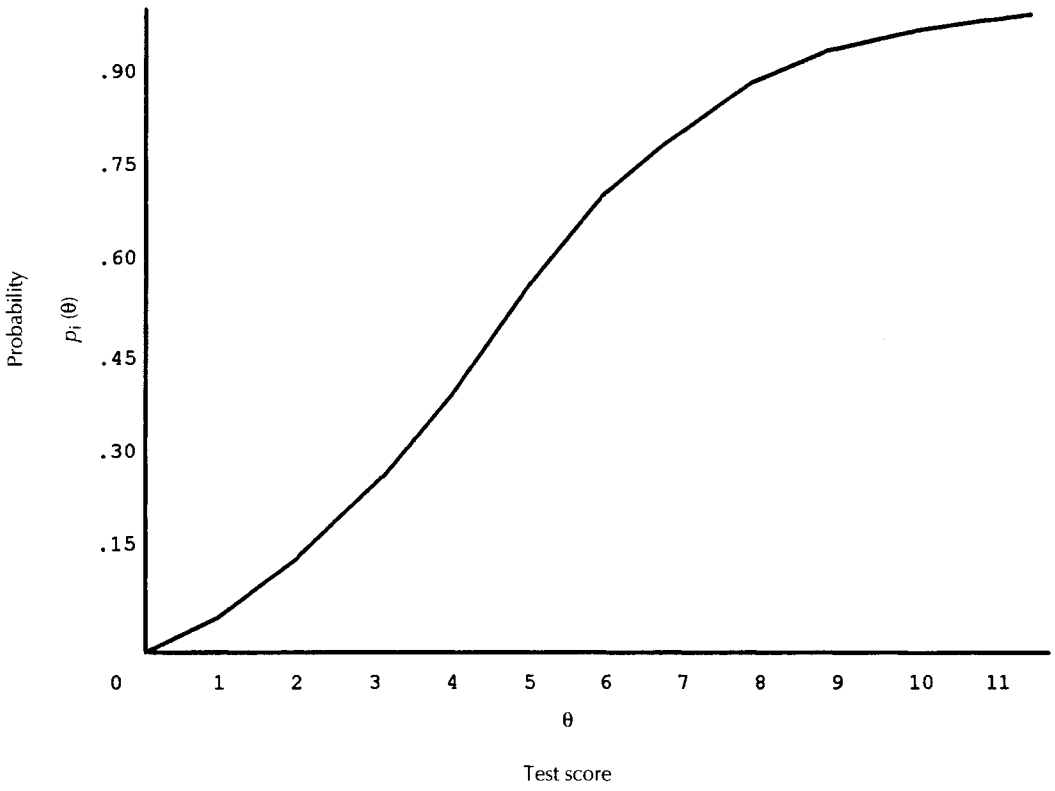


Figure 2.1. An Item-Characteristic Curve, p_i = probability that a test taker with trait value θ passes item *i*.

Schmelkin (1991) for an exploration of reliability and validity issues.

Reliability

Cronbach's Alpha. Sometimes called coefficient alpha, Cronbach's alpha (Cronbach, 1951) indexes the internal consistency reliability of a test, that is, the extent to which its items all measure the same characteristic. Cronbach's alpha derives from the following formula:

$$r_{xx} = [n/(n-1)]\cdot[SD^2 - \Sigma(SD_i^2)]/SD^2,$$

(Anastasi, 1988), which breaks down as follows,

$$a = n/(n-1)$$

$$b = SD^2 - \Sigma(SD_i^2)$$

$$c = SD^2$$

$$r_{xx} = a(b/c).$$

In the formulas above, n = the number of items, $\Sigma(SD_i^2)$ = the sum of the variances obtained from the items, and SD^2 = the variance of total scores.

Obtaining Cronbach's alpha by computer eliminates a great deal of time-consuming calculation. For test analysis, SAS works more effectively than other computer packages. The SAS statement,

PROC CORR ALPHA

generates output like that found in Table 2.2. A novice testp-developer should ignore the figures referring to standardized variables.

Table 2.2 displays Cronbach's alpha at the top. The first column contains the items on the test or subtest being analyzed. The second column shows each item's correlation with total test performance. Column three, the most important column, presents the coefficient alpha that would result if the test did not include the

Table 2.3. Facsimile of SAS Output Showing Factor Structure of an Eight-item Test

ROTATION METHOD: PROMAX

	FACTOR STRUCTURE (CORRELATIONS)			
	FACTOR 1	FACTOR 2	FACTOR 3	FACTOR 4
ITEM1	*.84**	.32	.21	.13
ITEM2	*.75**	.19	.23	.10
ITEM3	.49*	*.73**	.31	.23
ITEM4	.43	*.68**	.28	.31
ITEM5	.43	.11	*.71**	.18
ITEM6	.32	.12	*.67**	.05
ITEM7	.13	.07	.06	*.57**
ITEM8	.08	.08	.07	*.55**

item in that row. If this figure exceeds the current alpha appearing at the top, the removal of that item would increase the test's reliability. If the figure falls below the current alpha, the item's removal would decrease reliability. Thus, Cronbach's alpha helps guide the selection of items. Removing the items indicated can substantially increase reliability. Generally, psychometricians consider an alpha of at least .90 to be satisfactory. A test with an alpha below .93, however, can misclassify a small but nonnegligible number of people.

Ferguson's Delta. Ferguson's delta, or the *coefficient of test discrimination* (Ferguson, 1949), denotes the ratio between the number of discriminations made by the test and the greatest number that a test could generate, given the size of the sample and the number of items (Kline, 1986). According to Ferguson's theory (1949), the process of administering a test sets up relations between every item on the test and every individual taking the test. If a test taker, Pat, passes an ability item, the relation specifies that Pat's ability equals or exceeds (≥) the level needed to pass that item. If not, Pat's ability falls below (<) that level.

When the test is scored, every ≥ relation is expressed as a 1. Every < becomes a 0. These scores sum to a total score. Any test yields n + 1 possible total scores, including 0, where n denotes the number of items on the test. These total scores, in turn, lead to a second set of relations, this time between individuals. That is, each individual's total score equals (=) or differs from (> or <) every other individual's score. If Pat's total score exceeds Jon's, the test user can infer that Pat has more ability, or, presumably, a more pronounced trait, than Jon. Equal and lesser relations support similar inferences. Ferguson (1949) reasoned that a test should maximize the number of *difference relations* (> and <), given the number of items and the size of the sample, because people administer tests to detect relations of difference rather than equality.

The formula below (Ferguson, 1949; Kline, 1986) yields Ferguson's coefficient of discrimination, δ.

$$\delta = [(m + 1)(n^2 - \Sigma f_s^2)]/mn^2$$

In the formula, n = the number of test takers, m = the number of items, and f_s = the number of test takers obtaining each score. To calculate Ferguson's delta, a test developer should list each score, followed by the number of test takers who obtained it. A frequency distribution results:

SCORE	FREQUENCY
25	2
21	3
19	5
18	8
17	11

and so on. Next, square each frequency and add the squared frequencies. The result can take the name *sum of squared frequencies* (SSF) for the present. Then proceed as the following example illustrates.

Assume that 200 people complete a social-stress inventory. The inventory comprises 20 items. You, as the test's author, construct a frequency distribution, square the frequencies, and sum the results. You obtain an SSF of

3,500. Thus, you have $n = 200$, $m = 20$, and $\Sigma f_s^2 = 3{,}500$. You square n in advance for computational ease: $n^2 = 40{,}000$. Then, you calculate Ferguson's delta:

$$\delta = [(m + 1)(n^2 - \Sigma f_s^2)]/mn^2$$
$$= [(20 + 1)(40{,}000 - 3{,}500)]/(20 \times 40{,}000)$$
$$= (21 \times 36{,}500)/(20 \times 40{,}000)$$
$$= 766{,}500/800{,}000$$
$$= .96$$

Arriving at this figure, you realize that your test has substantial discriminating power.

Validity

Test developers find evidence that their instruments measure what they purport to measure in numerous ways. A complete exposition of the many ways to validate a test exceeds the scope of this chapter. Anastasi (1988) provides thorough coverage. For present purposes, validation can be divided into two general classes, convergent and divergent. A test shows *convergent* validity when it behaves similarly to other tests of the same or similar constructs. It shows *divergent* validity when it behaves differently from tests of dissimilar constructs. Anastasi (1986) writes that validation starts with construct definition (see Step 2 in this chapter). If a definition is accurate, complete, and comprehensive, a test developed from it has a good chance of measuring what its name implies. Factor analysis represents a key method for estimating a test's validity.

Factor Analysis. Many test constructors find that factor analysis is the most fascinating part of constructing a test. Conducted by hand, factor analysis takes a heavy toll in time and effort. Few psychologists attempted it before it became available by computer. Since then, however, it has exploded into widespread use. When given a SAS statement like the following,

```
PROC FACTOR N = 4 R = PROMAX SCREE
ROUND REORDER;
```

the computer generates a matrix similar to the one shown in Table 2.3. Each factor is a latent characteristic, one that no one can measure directly. In the table, Factor 1 correlates higher with items 1 and 2 than with items 3 through 8. In other words, factor 1 loads highest on items 1 and 2. Factor 2 loads highest on items 3 and 4, factor 3 on items 5 and 6, and so forth. Also, the various items load on the factors, just as the factors load on the items.

A test developer infers the nature of a factor by examining its loadings on the items. Factor 1 in the table illustrates this process. Suppose that high-loading items 1 and 2 call for verbal fluency. Mid-loading items 3, 4, and 5 entail a moderate amount of verbal fluency, together with other abilities. Low-loading items 7 and 8 require very little verbal fluency. The test constructor might infer that Factor 1 represents verbal fluency.

In practice, labeling factors in this manner involves subjective judgment. Test constructors should consider additional evidence. For example, they might factor-analyze a test designed to measure spatial ability with well-established measures of spatial ability and a dissimilar characteristic, like verbal ability. A test of spatial ability should correlate highly with the first measure and less so with the second, demonstrating convergent and divergent validity, respectively. However, test developers should also correlate their measures with nontest, real-world outcomes. For example, a measure of verbal ability should correlate with school achievement, especially in the verbal domain.

The challenge of developing a scientific test demands perseverance and rigor. A good test cannot stand alone; instead, it emerges from the scientific literature of the field of interest. It rests upon a theoretical foundation. It undergoes a timeworn process of development and validation to ensure that it accomplishes the purpose for which it was designed. Test users can and should compare it with existing measures to determine its relative merits. Though the task of constructing a test may seem daunting, an innovative, well-designed measure can win a long-term following and become an established part of research or clinical practice. The rewards, both tangible and intangible, can be great.

AUTHOR NOTE

Michael C. Ramsay, Department of Educational Psychology; Cecil R. Reynolds, Department of Educational Psychology.

The authors would like to thank Victor L. Willson for suggesting sources, providing direction, and reviewing the manuscript.

Correspondence regarding this chapter should be addressed to Michael C. Ramsay or Cecil R. Reynolds, Texas A&M University, College of Education, Department of Educational Psychology, College Station, Texas 77843. Electronic mail may be sent via Internet to either author at mike.ramsay@tamu.edu or crrh@bluebon.net, respectively.

REFERENCES

Alexander, P. A., & Judy, J. E. (1988). The interaction of domain-specific and strategic knowledge in academic performance. *Review of Educational Research, 58,* 375–404.

Alexander, P. A., Willson, V. L., White, C. S., & Fuqua, J. D. (1987). Analogical reasoning in young children. *Journal of Educational Psychology, 79,* 401–408.

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory.* Monterey, CA: Brooks/Cole Publishing Company.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

American Psychological Association (1986). Psychological Abstracts Information Services (PsychINFO) [Online].

American Psychological Association (1982). *Ethical principles in the conduct of research with human participants.* Washington, DC: Author.

American Psychological Association (1986). *Guidelines for computer-based tests and interpretations.* Washington, DC: Author.

Anastasi, A. (1984). The curious case of the indestructible strawperson. In S. N. Elliot & J. V. Mitchell, Jr. (Series Eds.) & B. S. Plake (Vol. Ed.), *Buros-Nebraska symposium on measurement & testing: Vol. 1. Social and technical issues in testing: Implications for test construction and usage.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psychology, 37,* 1–15.

Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan.

Anderson, J. R. (1990). *The adaptive character of thought.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Anderson, J. R. (1995). *Learning and memory: An integrated approach.* New York: Wiley.

Anstey, E. (1966). *Psychological tests.* London: Nelson.

Atkinson, J. W. (1980). Motivational effects in so-called tests of ability and educational achievement. In L. J. Fyans, Jr. (Ed.), *Achievement motivation: Recent trends in theory and research* (pp. 9–21). New York: Plenum.

Bauer, E. J. (1994). Diagnostic validity of an instrument for assessing behaviors of children with anatomically correct dolls. *Dissertation Abstracts International, 54,* 12B.

Beck, A. T. (1978). Beck Depression Inventory (BDI). New York: Psychological Corporation.

Binet, A., & Simon, T. (1980). The development of intelligenece in children. Nashville, TN: Williams Printing Company. (Original work published 1916)

Birenbaum, M., & Tatsuoka, K. K. (1987). Open-ended versus multiple choice response formats: It does make a difference for diagnostic purposes. *Applied Psychological Measurement, 11,* 385–395.

Bloom, B. M. (Ed.). (1956). *Taxonomy of educational objectives. Handbook 1: Cognitive domain.* New York: McKay.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37,* 29–51.

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for *n* dichotomously scored items. *Psychometrika, 35,* 179–197.

Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (1991). *Minnesota Multiphasic Personality Inventory—2: Manual for Administration and Scoring.* Minneapolis, MN: University of Minnesota Press.

Butterfield, E. C., Nielson, D., Tangen, K., & Richardson, M. B. (1985). Theoretically based psychometric measures of inductive reasoning. In S. E. Embretson (Ed.), *Test design: Developments in psychology and psychometrics* (pp. 77–147). New York: Academic Press.

Cattell, R. B. (1971). *Abilities: Their structure, growth, and action.* New York: Houghton Mifflin.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology, 1,* 55–81.

Committee on Professional Standards, & Committee on Psychological Tests and Assessment (1986). *Guidelines for Computer Based Tests and Interpretation.* Washington, DC: Authors.

Costa, P. T., Jr. & McCrae, R. R. (1992a). *NEO PI-R professional manual: Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI).* Odessa, FL: Psychological Assessment Resources.

Costa, P. T., Jr. & McCrae, R. R. (1992b). *Revised NEO Personality Inventory (NEO-PI-R): Item booklet—Form S.* Odessa, FL: Psychological Assessment Resources.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297–334.

Dahlstrom, W. G., Welsh, G. S., & Dahlstrom, L. E. (1972). *An MMPI handbook. Vol. 1: Clinical interpretation* (Rev. ed.). Minneapolis, MN: University of Minnesota Press.

Dahlstrom, W. G., Welsh, G. S., & Dahlstrom, L. E. (1975). *An MMPI handbook. Vol. 2: Research applications* (Rev. ed.). Minneapolis, MN: University of Minnesota Press.

Das, J. P., Kirby, J. R., & Jarman, R. F. (1979). *Simultaneous and successive processes.* New York: Academic Press.

DeLoache, J. S. (1995). The use of dolls in interviewing young children. In M. S. Zaragoza, J. R. Graham, G. C. N. Hall, R. Hirschman, & Y. S. Ben-Porath (Eds.), *Applied psychology: Individual, social, and community goals: Vol. 1. Memory and testimony in the child witness* (pp. 160–178). Thousand Oaks, CA: Sage Publications.

Embretson, S. E. (1985). Introduction to the problem of test design. In S. E. Embretson (Ed.), *Test design: Developments in psychology and psychometrics.* Orlando, FL: Academic Press.

Exner, J. E. (1974). *The Rorschach: A comprehensive system.* New York: Wiley.

Eyde, L. D., & Primoff, E. S. (1992). Responsible test use. In M. Zeidner & R. Most (Eds.), *Psychological testing: An inside view.* Palo Alto, CA: Consulting Psychologists Press.

Ferguson, G. A. (1949). On the theory of test development. *Psychometrika, 14,* 61–68.

Fischer, G. H., & Molenaar, I. W. (1995). *Rasch models: Foundations, recent developments, and applications.* New York: Springer-Verlag.

Flavell, J. H. (1985). *Cognitive development.* Englewood Cliffs, NJ: Prentice-Hall.

Gardner, H. (1983). *Frames of mind: The theory of multiple intelligences.* New York: Basic Books.

Goetz, E. T., Kulikowich, J. M., & Alexander, P. A. (1988, April). Investigating effects of domain specific and strategic knowledge on sixth graders and college students' performance on two analogy tasks. In R. Garner (Chair), *The interaction of domain specific and strategic knowledge in academic performance.* Symposium presented at the meeting of the American Educational Research Association, New Orleans, LA.

Golden, C. J., Sawicki, R. F., & Franzen, M. D. (1990). Test construction. In A. P. Goldstein & L. Krasner (Series Eds.) & G. Goldstein & M. Hersen (Vol. Eds.), *Pergamon General Psychology Series, No. 131: Handbook of psychological assessment* (2nd ed.). New York: Pergamon.

Guilford, J. P. (1967). *McGraw-Hill series in psychology: The nature of human intelligence.* New York: McGraw-Hill.

Guilford, J. P. (1977). *Way beyond the IQ.* Buffalo, NY: Creative Education Foundation.

Guilford, J. P., & Zimmerman, W. S. (1956). Fourteen dimensions of temperament. *Psychological monographs, 70*(10, Serial No. 417).

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston: KluwerÞNijhoff Publishing.

Hambleton, R. K., Swaminathan, H., Rogers, H. J. (1991). Concepts, models, and features. In: R. M. Jaeger (Series Ed.), *Measurement methods for the social sciences: Vol. 2. Fundamentals of item response theory.* Newbury Park, CA: Sage Publications.

Hansen, J.-I. C., & Campbell D. P. (1985). Strong-Campbell Interest Inventory of the Strong Vocational Interest Blank. Stanford, CA: Stanford University Press.

Hathaway, S. R., McKinley, J. C., Butcher, J. N., Dahlstrom, W. G., Graham, J. R., & Tellegen, A. (1989). *Minnesota Multiphasic Personality Inventory-2.* Minneapolis, MN: University of Minnesota Press.

Henrysson, S. (1971). Gathering, analyzing, and using data on test items. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 130-159). Washington, DC: American Council on Education.

Hill, K. T. (1980). Motivation, evaluation, and educational testing policy. In L. J. Fyans, Jr. (Ed.). *Achievement motivation: Recent trends*

in theory and research (pp. 34–95). New York: Plenum.

John, O. P. (1990). The "big five" factor taxonomy: Dimensions of personality in the natural language and in questionnaires. In L. A. Pervin (Ed.), *Handbook of personality theory and research.* New York: Guilford Press.

Kaplan, R. M., & Saccuzzo, D. P. (1989). *Psychological testing: Principles, applications, and issues.* Pacific Grove, CA: Brooks/Cole.

Kaufman, A. S., & Kaufman, N. L. (1983a). *Kaufman Assessment Battery for Children (K-ABC).* Circle Pines, MN: American Guidance Service.

Kaufman, A. S., & Kaufman, N. L. (1983b). *Kaufman Assessment Battery for Children (K-ABC): Administration and scoring manual.* Circle Pines, MN: American Guidance Service.

Kaufman, A. S., & Kaufman, N. L. (1983c). *Kaufman Assessment Battery for Children (K-ABC): Interpretive manual.* Circle Pines, MN: American Guidance Service.

Kelley, T. L. (1939). The selection of upper and lower groups for the validation of test items. *Journal of Educational Psychology, 30,* 17–24.

Kline, P. (1986). *A handbook of test construction: Introduction to psychometric design.* London: Methuen.

Kline, P. (1993). *The handbook of psychological testing.* London: Routledge.

Krathwohl, D. R., Bloom, B. S., & Masia, B. B. (1964). *Taxonomy of educational objectives. Handbook 2: Affective domain.* New York: McKay.

Lord, F. M. (1952). *A theory of test scores.* (Psychometric Monograph, No. 7). New York: Psychometric Society.

Lord, F. M. (1953a). An application of confidence intervals and of maximum likelihood to the estimation of an examinee's ability. *Psychometrika, 18,* 57–76.

Lord, F. M. (1953b). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement, 13,* 517–548.

Lord, F. M. (1968). An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement, 28,* 989–1020.

Lord, F. M. (1977a). A broad-range tailored test of verbal ability. *Applied Psychological Measurement, 1,* 95–100.

Lord, F. M. (1977b). Practical applications of item characteristic curve theory. *Journal of Educational Measurement, 14,* 117–138.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Luria, A. R. (1966). *Human brain and psychological processes.* New York: Harper & Row.

Luria, A. R. (1972). *The working brain.* New York: Penguin.

Luria, A. R. (1980). *Higher cortical functions in man.* (B. Haigh, Trans. 2nd ed., rev. and exp.). New York: Basic Books. (Original work published 1962)

Millon, T. (1987). *Millon Clinical Multiaxial Inventory-II: Manual for the MCMI-II* (2nd ed.). Minneapolis, MN: National Computer Systems.

Mischel, W. (1990). Personality dispositions revisited and revised: A view after three decades. In L. A. Pervin (Ed.), *Handbook of personality and social psychology.* New York: Guilford Press.

Most, R. B., & Zeidner, M. (1995). Constructing personality and intelligence instruments: Methods and Issues. In D. H. Saklofske & M. Zeidner (Eds.), *International Handbook of Personality and Intelligence.* New York: Plenum.

Murray, H. A. (1938). *Explorations in personality.* New York: Oxford University Press.

Myers, I. B. (1987). *Introduction to type.* Palo Alto, CA: Consulting Psychologists Press.

Norman, W. T. (1963). Toward an adequate taxonomy of personality attributes: Replicated factor structure. *Journal of Abnormal and Social Psychology, 66,* 574-583.

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, design, and analysis: An integrated approach.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Qualis, A. L. (1995). Estimating the reliability of a test containing multiple item formats. *Applied Measurement in Education, 8,* 111-120.

Ramsay, M. C., & Reynolds, C. R. (1995). Separate digits tests: A brief history, a literature review, and a reexamination of the factor structure of the Test of Memory and Learning (TOMAL). *Neuropsychology Review, 5,* 151-171.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests.* Chicago: University of Chicago.

Reynolds, C. R. (1986). Measurement and assessment of childhood exceptionality. In I. B. Weiner (Series Ed.) & R. T. Brown & C. R. Reynolds (Vol. Eds.), *Wiley series on personality processes. Psychological perspectives on childhood exceptionality: A handbook.* New York: Wiley-Interscience.

Reynolds, C. R. (1989). Measurement and statistical problems in neuropsychological assessment of children. In C. R. Reynolds & E. Fletcher-Janzen (Eds.), *Handbook of clinical child neuropsychology* (pp. 147-166). New York: Plenum.

Reynolds, C. R., & Bigler, E. D. (1995a). *Test of Memory and Learning (TOMAL).* Austin, TX: Pro-Ed.

Reynolds, C. R., & Bigler, E. D. (1995b). *Test of Memory and Learning (TOMAL): Examiner's manual.* Austin, TX: Pro-Ed.

Reynolds, C. R., & Kamphaus, R. W. (1992a). Behavior Assessment System for Children (BASC). Circle Pines, MN: American Guidance Services.

Reynolds, C. R., & Kamphaus, R. W. (1992b). *Behavior Assessment System for Children (BASC): Manual.* Circle Pines, MN: American Guidance Services.

Robertson, G. J. (1990). A practical model for test development. In C. R. Reynolds & R. W. Kamphaus (Eds.), *Handbook of psychological and educational assessment of children: Vol. 2. Intelligence and achievement* (pp. 62-85). New York: Guilford.

Robertson, G. J. (1992). Psychological tests: Development, publication, and distribution. In M. Zeidner & R. Most (Eds.), *Psychological testing: An inside view.* Palo Alto, CA: Consulting Psychologists Press.

Roid, G. H., & Haladyna, T. M. (1982). *A technology for test-item writing.* New York: Academic Press.

Rorschach, H. (1921). *Psychodiagnostik.* Bern: Bircher.

Rotter Incomplete Sentences Blank (RISB) 1977. College Response Sheet. San Antonio, TX: Psychological Corporation.

Rotter, J. B., & Rafferty, J. E. (1950). *Manual: The Rotter Incomplete Sentences Blank (RISB) College Form.* Cleveland, OH: Psychological Corporation.

Rudner, L. M. (1996). *Questions to ask when evaluating tests.* Lincoln, NE: Buros Institute of Mental Measurements.

Samejima, F. (1983). Some methods and approaches of Estimating the operating characteristics of discrete item responses. In H. Wainer & S. Messick (Eds.), *Principals of modern psychological measurement: A festschrift for Frederic M. Lord.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Skinner, L. J., Giles, M. K., Berry, K. K. (1994). Anatomically correct dolls and validation interviews: Standardization, norms, and training issues. *Journal of Offender Rehabilitation, 21,* 45-72.

Spearman, C. E. (1904). "General intelligence," objectively determined and measured. *American Journal of Psychology, 15,* 201-292.

Spearman, C. E. (1927). The abilities of man: Their nature and measurement. London: MacMillan and Company, Ltd.

Spearman, C. E. (1973). The nature of 'intelligence' and the principles of cognition. New York: Arno Press. (Original work published 1923)

Spearman, C. E., & Jones, L. W. (1950). Human ability: A continuation of "The Abilities of Man." London: MacMillan & Company, Ltd.

Spielberger, C. D. (1977). *State-Trait Anxiety Inventory (STAI).* Palo Alto, CA: Consulting Psychologists Press.

Spencer, S. J., Josephs, R. A., & Steele, C. M. (1993). Low self-esteem: The uphill struggle for self-integrity. In R. F. Baumeister (Ed.), *Self-esteem: The puzzle of low self-regard* (pp. 21-36). New York: Plenum.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality & Social Psychology, 69,* 797-811.

Sternberg, R. J. (1985). *Beyond IQ: A triarchic theory of human intelligence.* New York: Cambridge University Press.

Sternberg, R. J. (1990). *Metaphors of mind: Conceptions of the nature of intelligence.* Cambridge, England: Cambridge University Press.

Sternberg, R. J., & Wagner, R. K. (1986). *Practical intelligence: Nature and origins of compe-*

tence in the everyday world. Cambridge, England: Cambridge University Press.

Stewart, K. J., Reynolds, C. R., & Lorys-Vernon, A. (1990). Professional standards and practice in child assessment. In C. R. Reynolds & R. W. Kamphaus (Eds.), *Handbook of educational & psychological assessment.* New York: Guilford.

Thorndike, R. L. (1982). *Applied psychometrics.* New York: Wiley.

Thurstone, L. L. (1938). *Primary mental abilities.* Chicago: Chicago University Press.

Thurstone, L. L., & Thurstone, T. G. (1941). *Factorial studies of intelligence.* (Psychometric Monograph, No. 2). Chicago: University of Chicago Press.

Walsh, W. B., & Betz, N. E. (1995). *Tests and Assessment* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

Wechsler, D. (1981). *Wechsler Adult Intelligence Scale—Revised (WAIS-R) manual.* New York: Psychological Corporation.

Wechsler, D. (1991a). *Wechsler Intelligence Scale for Children—Third Edition (WISC-III).* San Antonio, TX: Psychological Corporation.

Wechsler, D. (1991b). *Wechsler Intelligence Scale for Children—Third Edition (WISC-III) manual.* San Antonio, TX: Psychological Corporation.

Weiss, D. J. (Ed.). (1983). *New horizons in testing: Latent trait test theory and computerized adaptive testing.* New York: Academic Press.

Willson, V. L. (1989). Cognitive and developmental effects on item performance in intelligence and achievement tests for young children. *Journal of Educational Measurement, 26,* 103-119.

Willson, V. L., Goetz, E. T., Hall, R. J., & Applegate, E. B., III. (1986, April). *Effects of varying the number of elements and transformations of matrix analogies on children ages 5-12.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Willson, V. L., Kulikowich, J. M., Alexander, P. A., & Farrell, D. (1988, April). *A cognitive theory for test design: Implications for assessing domain-specific and strategic knowledge.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Willson, V. L., & Olivarez, A. O., Jr. (1987). *Information processing of pictorial arithmetic problems.* Unpublished manuscript, Texas A&M University, College Station, TX.

Willson, V. L., Reynolds, C. R., Chatman, S. P., & Kaufman, A. S. (1985). Confirmatory analysis of simultaneous, sequential, and achievement factors on the K-ABC at 11 age levels ranging from 2 ½ to 12 ½ years. *Journal of School Psychology, 23,* 261-269.

Wright, B. D., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement, 29,* 23-48.

Wright, B. D., & Stone, M. H. (1979). *Best test design.* Chicago: MESA.

Wundt, W. M. (1907). *Outlines of psychology* (3rd rev. ed.; C. H. Judd, Trans.). Leipzig, Germany: W. Engelmann. (Original work published 1896)

Wundt, W. M. (1912). *An introduction to psychology* (R. Pintner, Trans.). New York: Arno Press. (Original work published 1906)

Zeidner, M. (1995). Personality trait correlates of intelligence. In D. H. Saklofske & M. Zeidner (Eds.), *International handbook of personality and intelligence.* New York: Plenum.

CHAPTER 3

SCALING TECHNIQUES

Mark D. Reckase

INTRODUCTION

When assessment instruments are administered, the goal is to gather information about the characteristics of the individual being assessed. The characteristics of interest may be directly observable, such as crying or thumb sucking in children, and the assessment instrument may be a form that is used to record the observations. A more complex situation exists when the characteristics of interest are not directly observable or when the required observations are much too extensive to be practical. In those cases, the information obtained from the assessment instrument is used to infer the characteristics of the person. Examples of such characteristics include intelligence, aptitude for foreign language, artistic interests, repression, and anxiety. The vast majority of psychological assessment instruments provide information about characteristics of the latter type.

An interesting research question is whether characteristics such as intelligence exist as physical entities or whether they are statistical constructions. Goldstein (1994) suggests that many of the constructs that are the focus of assessment instruments do not have any stable physiological meaning, but rather are arbitrary functions of items selected for the assessment instrument. This chapter does not explicitly address the issue of whether constructs such as intelligence exist in a physiological sense. Rather, it is assumed that if individuals can be consistently grouped on the basis of responses to items, there is a "real" underlying cause of the responses, but that the cause may not be easily determined. The assessment scale provides a means for summarizing the responses, but it does not necessarily define a physiological truth.

In addition to gathering information to describe the characteristics of a person, there is usually interest in determining the relative amount of each characteristic exhibited by a person. Is the person expressing considerable anxiety, or not very much? Will the individual learn a foreign language quickly or slowly? This interest in the relative amount of a characteristic implies that it is desirable to quantify observations in some way. If done in a reasonable way, the resulting numerical values not only give an indication of the relative amount of each characteristic, but also allow comparisons to be made between persons and give a convenient procedure for summarizing observations. In addition, the numerical values lend themselves to further analyses that may help reveal relationships that exist among different characteristics. That is, the numerical values are used to infer relationships among the underlying causative variables (hypothetical constructs) that explain or describe a person's behavior.

The discovery of relationships between quantitative measures of hypothetical constructs is a necessary first step in the development of an area of science. It is difficult to envision an area of science that has advanced without quantitative information. For example, proportions of phenotypes were needed to develop basic laws of genetics, and atomic weights were needed to develop the molecular theory in chemistry. So it is also with psychology. The advances in the quantification of psychological variables have

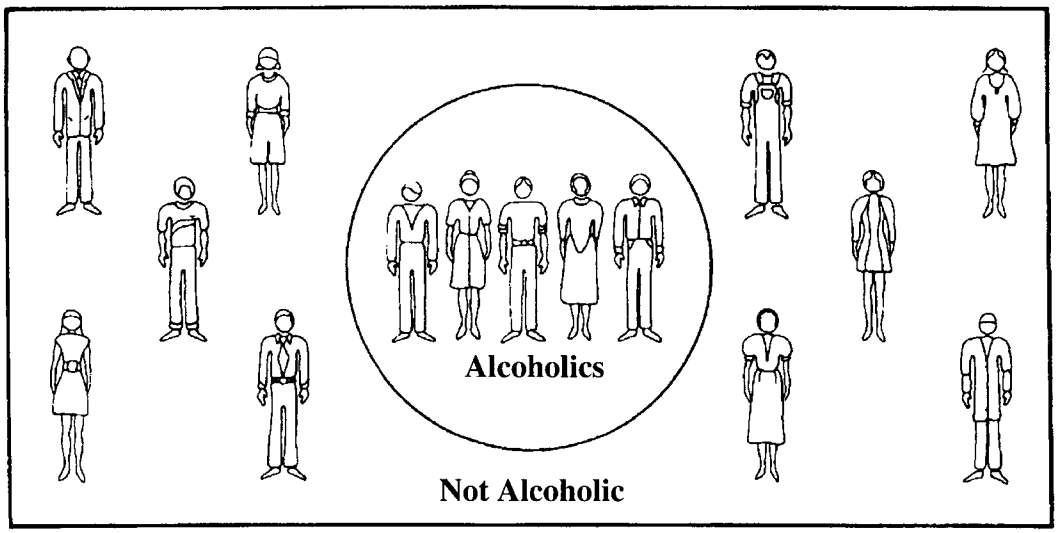


Figure 3.1. Alcoholics/Not Alcoholics

facilitated both the theory and practice of psychology (for example, see Cattell, Eber, & Tatsuoka, 1970; Guilford & Hoepfner, 1971; Holland, 1966).

The process that is used to assign numbers to collections of observations is the topic of this chapter. This process is called *scaling*. If scaling is successful, the numerical value that is obtained from an assessment instrument can be used to infer accurately the characteristics of a person and the relationships among the characteristics. In a very basic psychological sense, scaling can be defined as the assignment of meaning to a set of numbers derived from an assessment instrument.

The purpose of this chapter is to present some fundamental concepts about the features and uses of numerical scales developed to describe psychological constructs. The chapter is organized around two major topics: (a) the theory behind scale formation, and (b) the relationship of that theory to the scales produced by several psychological scaling procedures. This chapter will not present a catalog of scaling procedures, although some specific procedures will be described. Rather, it offers a basic philosophy of scale development that can be used to construct new scales for the assessment of psychological traits. Although practical methodologies will be presented, the results of these methodologies will always be related to the basic philosophy of scale formation rather than be presented as "cookbook" procedures to be followed.

SCALING THEORY

The basic concept in the theory of scale formation is that of a property (see Rozeboom, 1966, for a more abstract development of the concepts presented here). A property can be thought of in at least two different ways. It is commonly used to denote a characteristic, trait, or quality of an entity. Human beings are mammals; being a mammal is a property of human beings. This usage of the term *property* is sufficient for conversational use, but it is not precise enough for use in scaling theory.

For the purposes of this chapter, a *property* will be defined by a set of entities, the entities of interest usually being people. Any set of entities defines a property, but some sets are more interesting in a psychological sense than others. For example, the set of all persons who are alcoholics defines the property "alcoholic." If a person belongs to that set, he or she is an alcoholic. Theoretically, we can determine whether a person is an alcoholic by checking whether he or she is a member of the set (see Figure 3.1). A less interesting property, X, might be made up of some random selection of entities. Then, each entity in that set has the property that it is a member of X. Although the sets "alcoholic" and "member of X" are equally good

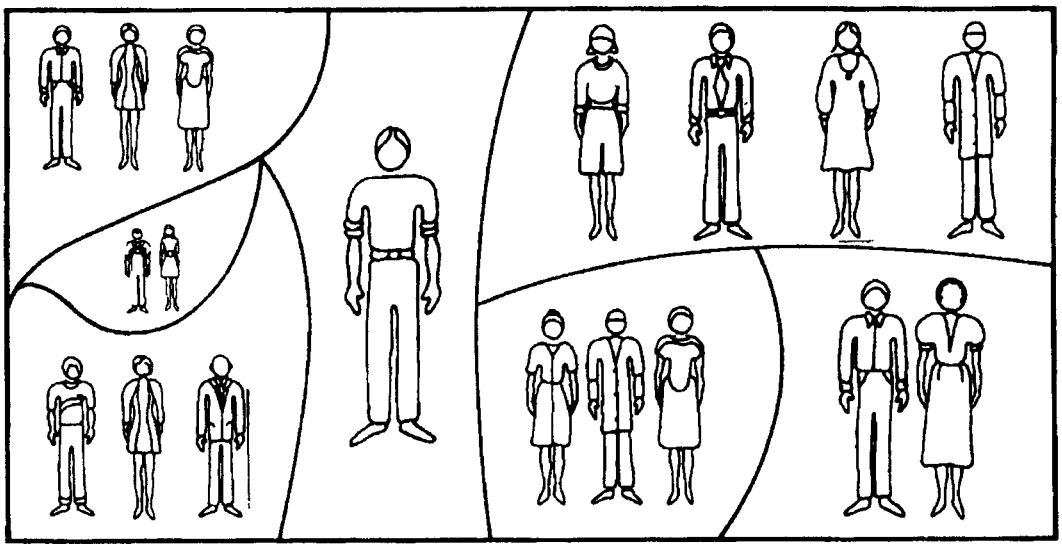


Figure 3.2. The Natural Variable *Height*

definitions of properties, the random set does not have psychological meaning. The process of defining and assigning meaning to a scale will requires that properties be defined in a meaningful way.

The definition of a property used here is very similar to that of a "concept" in the psychological study of concept formation. Just as people are presented with exemplars and nonexemplars of a concept until they develop a personal, empirically useful definition of the concept, members of a property set can be thought of as exemplars and those not in the set as nonexemplars. These two groups define the property. The "concept" that a person forms in studying the characteristics of the two groups is an abstract generalization that summarizes the property for the person, but the abstraction is not itself the property. Only the two sets, the exemplar group and the nonexemplar group, contain all the nuances of the property.

When used in a psychological context, the sets of people that define properties are often more restricted in their definition than the "alcoholics" example given above. Most psychological characteristics exist at a number of levels. Therefore, psychological properties are usually defined by sets of people having the same level of the characteristic of interest rather than membership in a global, single class. For example, a set of people who all have the same amount of test anxiety defines the property of that level of test anxiety. Another set of people defines another, different level of test anxiety. A different set of people is hypothesized to exist for each different level of test anxiety, and each of those defines a property. Thus, when a person is said to have a high anxiety level, in theory that means the person belongs to the set of people who are as highly anxious as the person in question. All of the people in that set have the property of having a high level of anxiety.

Defining a Property

The actual process required to define a property is that of determining equivalence. All persons who have a property are equivalent on the characteristic of interest and are different in at least the level of the characteristic from those persons who do not have the property. If a procedure can be developed to determine whether two individuals are equivalent on the characteristic of interest to some practical level of precision, then the first step toward scale formation has been taken.

An example of the formation of the sets of people that form properties can easily be given if height is used as the characteristic of interest. Imagine a room full of people of varying heights. It would not be a difficult task to sort the people into groups of individuals who have approximately the same height. Each of the groups would define a

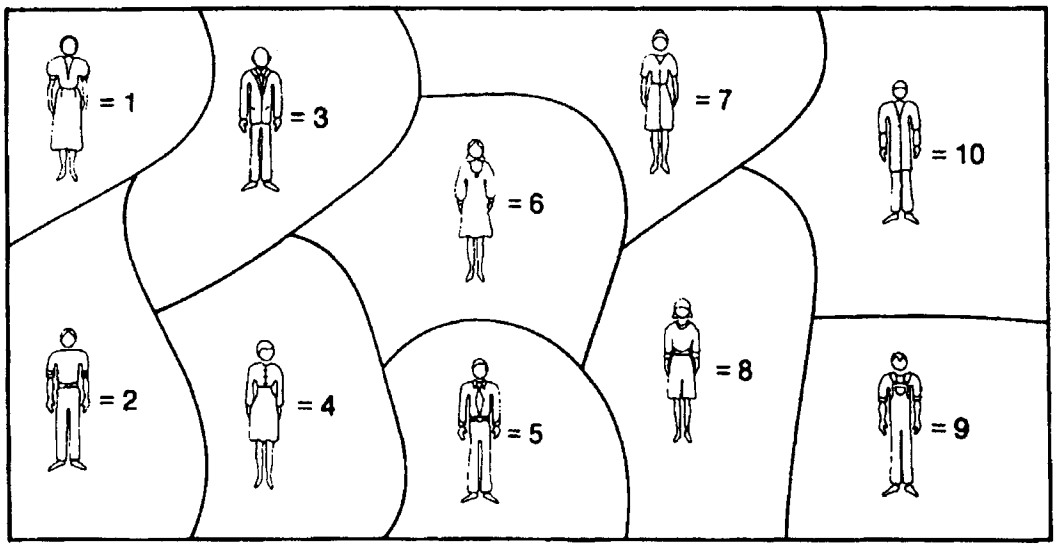


Figure 3.3. A Scaled Variable

property of being an individual of a particular height.

Since the members of a property set are equivalent on the characteristic of interest, these sets are also called *equivalence classes* (Krantz, Luce, Suppes, & Tversky, 1971). Equivalence classes can be shown to divide the full set of entities in such a way that all entities are in an equivalence class and no entity is in more than one. The forming of collections of equivalence classes that include all individuals of interest, called *partitioning* of the population of entities, is important for the definition of a *variable.*

Definition of a Natural Variable

A further step toward the formation of a scale can be demonstrated using the height example given above. If the room full of people contained the total population of people of interest as far as the characteristic "height" is concerned, the sets of people of equal height contain all of the possible properties related to the concept *height.* This situation is depicted in Figure 3.2. Each person belongs to only one set, and every person belongs to a set, even if that person is standing alone because no one else is of the same height. Sets containing one person are perfectly legitimate. Thus, each person has a height property and no person has more than one.

The collection of sets that define the properties for different levels of height together define a concept called a *natural variable.* A natural variable is a collection of properties in which every entity is included in a property and no entity is in more than one property. The variable is called "natural" because it is defined using the actual objects of interest and it does not depend on abstract symbols such as numbers.

All the variables that are commonly dealt with in psychology are assumed to be natural variables. When the variable "mechanical aptitude" is used, it is assumed that at any moment in time numerous groups could be formed, each of which contain persons who are equivalent in their level of mechanical aptitude. All persons are assumed to have some level of mechanical aptitude, and no person is assumed to have more than one level of mechanical aptitude at a given time. This set of conditions holds for any psychological trait for which a scale can be formed.

Of course, the procedure described above for forming a natural variable is impractical in reality. The example was given only to illustrate the concept of a variable that is commonly used in psychology. A variable is merely a collection of sets of individuals such that individuals in a given set are equal on the trait of interest. In order for the con-

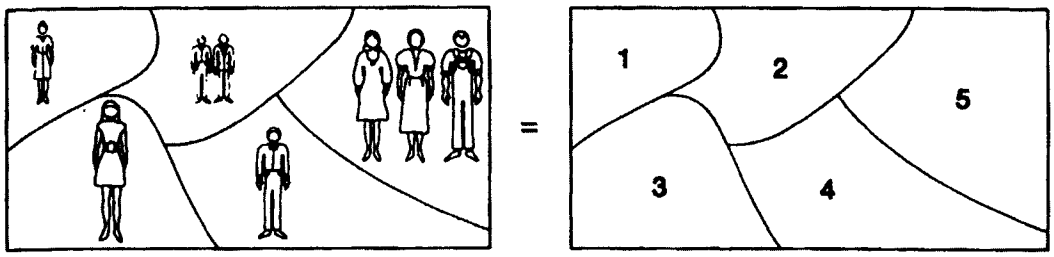


Figure 3.4. An Example of a Nominal Scaling

cept of *variable* to be of use, some means must be determined to identify the particular set to which a person belongs without going through the physical sorting process. The procedure that is commonly used is assigning a unique number to each property set and then developing a set of rules for determining the number assigned to each person so that their property set can be uniquely determined.

Definition of a Scaled Variable

Up to this point, rather cumbersome language has been used to describe a property and a natural variable. These concepts can be simplified considerably if abstract symbols are used to represent actual individuals. Suppose, in the height example given above, that each person in the room had been randomly assigned a number between 1 and 10. The individuals could then be grouped according to the number that had been assigned to them to form a collection of sets. If each person is given one number, and no person receives more than one number, this collection of sets forms a variable that can be called "the number assigned to each person." However, this variable is not a natural variable because it was not defined using naturally occurring features. This variable may or may not have a connection to an underlying trait. It is strictly an abstract variable. This type of variable will be labeled a *scaled variable* (see Figure 3.3).

An infinite number of scaled variables are possible. Any numbers can be assigned to a set of individuals, and if the conditions of a property and a variable are met by the assignment (i.e., each person gets one number and no person gets more than one), then the result is a scaled variable. If, however, the properties in the scaled variable are related to the properties in a natural variable, a very powerful result is obtained.

Definition of a Scaling

If each person having the same height property is assigned the same number, then the grouping of sets that defines the natural variable "height" is exactly the same as the grouping of sets that defines the scaled variable. If this relationship between the variables occurs, the result is called a *scaling* or a *nominal scaling* of the variable height. The relationship between a natural variable and a scaled variable for a nominal scaling is shown in Figure 3.4.

The scaling of a natural variable yields a very powerful result. The individuals no longer have to be physically present for one to know whether they are equal in height. The numbers assigned to them need only be compared. If two persons have been assigned the same number, they are equal in height. If two persons have been assigned different numbers, they are different in height.

It should be clear that the critical part of scaling a variable is the procedure for assigning the numbers. If the numbers are assigned in such a way that persons with the same number have the same property on the trait of interest, a perfect scaling results, and the information present in the natural variable is present in the scaled variable. For most natural variables of interest to psychologists, every member of a property is not assigned the same number because of errors in the assignment process. In most cases, the scaling would be considered successful if most persons in a property were assigned the same number. To the extent that the correct assignment is made, the numerical assignment is

said to be valid. The greater the frequency of persons assigned the wrong numbers, the more invalid is the numerical assignment. The numerical values assigned to the properties are called *measurements* when the assignment is reasonably valid.

Representation versus Cause and Effect

The approach taken here is based on an assumption that certain unobservable psychological traits exist and the goal is to "represent" these unobservable traits with a numerical scaled variable. This is the philosophical framework provided by Krantz, Luce, Suppes, and Tversky (1971) in the first and later two volumes of the monumental work, *Foundations of Measurement.* An alternative approach is suggested by Bollen and Lennox (1991) who construct scales from convenient indicator variables. For example, they define "exposure to discrimination" as a function of indicator variables for race, sex, age, and disabilities. They label the measures that are produced in this way as *causal indicators* because the indicators determine the latent variable rather than represent the variable. This chapter takes the representational measurement approach, but it is important to realize that alternative measurement philosophies exist.

Scale Types

When the results of the scaling of a psychological variable are used, more information is usually desired than an indication of whether a person does or does not have a property (i.e., belong to a property set). Information about the magnitude of the level of the trait of interest is also desired. For this information to be represented, it must first be possible to order the properties of the natural variable. For the height example given above, the procedure for doing the ordering is quite obvious. Persons with the various height properties can be compared and ranked according to height. If the numbers assigned to the property sets have the same order as the properties in the natural variable, the scaling that results will contain information about the ordering and is called an *ordinal scale.* Still more information can be included in the scaling of a natural variable if an assumption can be made about the properties in the natural variable. If it can be assumed that when the ordered properties of a natural variable differ by an equal amount on the characteristic of interest, the numerical values in the corresponding scaled variable also differ by an equal amount, then the resulting scaling is called an *interval scale.* That is, if numbers are assigned in such a way that when the distances between sets of numerical values are equal, the psychological differences in the elements of the corresponding properties of the natural variable are also equal, an interval scaling is the result.

The measurement of temperature using the Celsius scale is a common example of measurement at the interval-scale level. When Anders Celsius developed this scale he assigned 0 to the freezing point of water and 100 to the boiling point of water and divided the temperature range between into 100 units. This numerical-assignment rule defined the scaled variable now labeled Celsius temperature. The equal numerical units on the Celsius scale correspond to the increase in the temperature of one cubic centimeter of water brought about by the application of one calorie of heat. The physical sets of objects of equal temperature define the properties in the natural variable. Thus, for this temperature scale, equal differences in the natural variable correspond to equal numerical differences on the Celsius scale. Therefore, the Celsius temperature scale is an interval scale.

The classification of scales as ordinal or interval takes on importance because psychometric theorists (e.g., Stevens, 1959) have suggested that many common statistical procedures (e.g., the mean, standard deviation, etc.) require interval-scale measurements for proper application. These procedures use the difference between scores to compute the descriptive statistics. Because the distances between scores from ordinal scales do not provide information about the differences in properties on the natural variable, Stevens indicated that the interpretation of the statistics computed on these scales did not apply to the natural variable,

The opposing point of view is that most psychological scales give a reasonably close approximation to interval scales and, therefore, that interval-based statistics can be applied and interpreted relative to the natural variables. Labovitz (1970) performed a study that supported this point of view. He demonstrated that only if the match of the scale intervals for the scaled and natural variables varied by great amounts were the interpretations of the statistics adversely affected. Adams, Fagat, and Robinson (1965) also argued that it is the interpretations of the

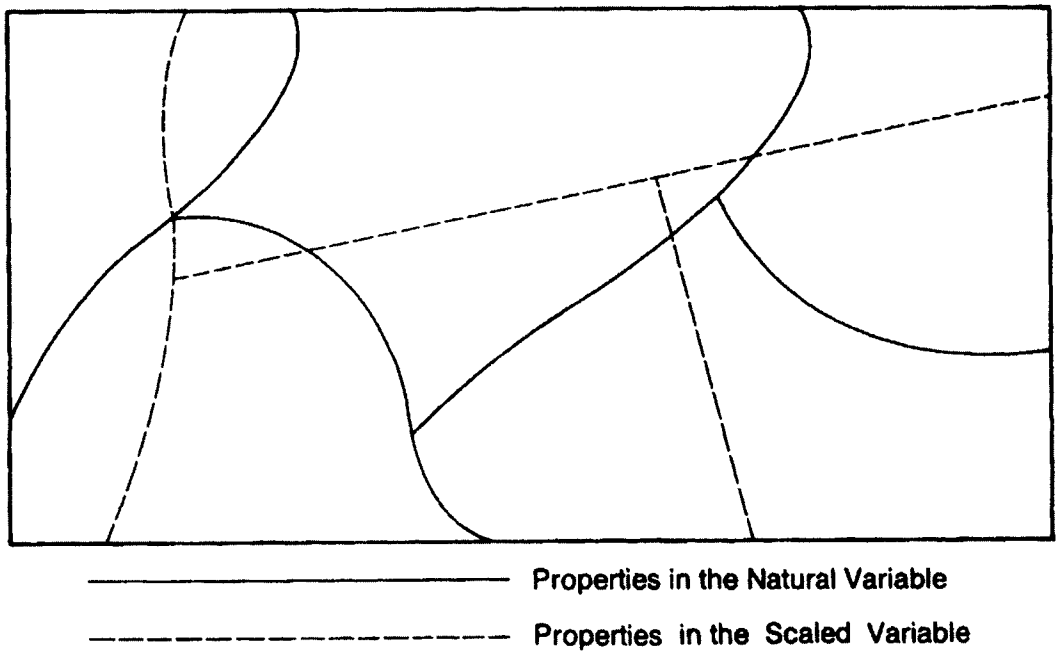


Figure 3.5. Example of an Invalid Scaling

natural variable that are important. If the scaled variable and the statistics applied to it yield useful information about the natural variable (e.g., the scaled variable is found to correlate with other variables of interest), the level of measurement is not a concern. Few psychologists are very dogmatic about the relationship between scale type and the use of particular statistical procedures. However, scale types should still be considered when interpreting the results of an analysis.

One other type of scale has been included in the topology of scales developed by Stevens (1959). For this type of scale, one of the property sets is defined by the group of individuals who quantitatively have none of the variable of interest. This property defines the true zero point of the scale. In addition to the existence of the property defining the true zero point, the natural variable must also meet all of the requirements for an interval scale. That is, equal differences in the numbers assigned to the properties must correspond to equal psychological differences in the properties. Of course, the entities in the true zero property must be assigned the number zero.

If all of these conditions are met, the resulting scaling is called a *ratio scale.* Ratio scales are relatively rare when psychological traits are scaled because of the difficulty in defining the zero point. While objects approaching zero height are relatively easy to find (e.g., very thin paper), persons approaching zero intelligence are hard to imagine. Even if an object such as a rock is defined as having zero intelligence, the equal steps of intelligence required to get the interval properties of the scale are difficult to determine. For example, is the difference in intelligence between a dog and a rock the same as the difference between a person and a dog? At some point in the future we may be able to develop psychological scales with ratio-scale properties, but currently the best that can be expected are interval scales.

Definition of Validity

Up to this point, various scale types have been defined, and the relationship between scaled variables and natural variables has been considered. However, as was mentioned earlier, the match between the scaled variable and the natural variable is seldom exact. In some cases, a scaled vari-

Property Characteristics

No unusual reactions	Violent pounding of heart	Violent pounding Sinking feeling	Violent pounding Sinking feeling Trembling all over

Figure 3.6. Characteristics of Individuals in Four Property Sets of the Natural Variable, *Fear*

able can be formed that has properties that do not conform at all to the properties of the natural variable in question. In these cases, there is more of a problem than occasionally misclassifying a person—the sets clearly do not match. Such a case is illustrated in Figure 3.5.

When the sets do not match, the scaling does not result in a *valid* measure of the natural variable. The scaled variable does not give useful information about an entity's membership in the properties of the natural variable. Obviously, the most important task in forming a scale is insuring that a valid scale is the result. The next section deals with the techniques that are available for forming scales and checking their validity.

Scaled variables can be reproducible in that assignment of numbers at different times or with different techniques yields the same set of scaled variable properties, but that are not valid because the properties do not match those in the natural variable. These scalings are *reliable* because the assignment is consistent, but they are not valid. Fairly consistent assignment of numbers to individuals is a minimal condition for validity but such consistency does not guarantee validity. The sets from the scaled and natural variables must match for the scaling to be valid.

TECHNIQUES FOR SCALE FORMATION

Guttman's Scalogram Approach

According to Guttman's scalogram approach to the formation of a scale (1950), the properties in a natural variable can be ordered in such a way that individuals in a higher-level property include all of the characteristics of those in lower-level properties plus at least one more. That is, if the properties in a natural variable are labeled in increasing order starting with a_1 to a_{n-1}, then those individuals in property a_n have all the characteristics of the persons in properties a_1 to a_{n-1} plus at least one more. The task involved in scale formation is to find a series of behaviors such that all of those persons who exhibit a particular set of behaviors belong to the same property, and those in the next higher property exhibit at least one additional behavior.

The classic example of a Guttman scale is the measure of fear developed for use with soldiers in World War II (Stouffer, 1950). For that scale, those who did not experience "violent pounding of the heart" formed the lowest property set, while those who did formed the next higher property in the natural variable. If a sinking feeling in the stomach as well as a violent pounding of the heart were reported, the person belonged in the next higher property. If, in addition to the other two characteristics, trembling all over was reported, the person belonged in the next higher property level on the natural variable "fear" (see Figure 3.6). In all, 10 fear properties were defined through sets of physiological characteristics.

The scaled variable corresponding to the natural variable was formed by simply counting up the number of characteristics that were present. If no characteristics were present, the person was assigned a numerical score of 0. If only violent pounding of the heart were present, a score of 1 was assigned. If both violent pounding of the heart and a sinking feeling in the stomach were reported, a score of 2 was assigned. Because of the cumulative nature of this type of natural variable, the case where a person has a sinking feeling in his or her stomach but does not have a pounding heart occurs infrequently. Therefore, the meaning of a score of one is unambiguous.

The scaled variable is formed by grouping together into a property all the individuals who

Properties in Natural Variable	Characteristics: Pounding	Sinking	Trembling	Sick	Weak
a_5	1	1	1	1	1
a_4	0	1	1	1	1
a_3	0	0	1	1	1
a_2	0	0	0	1	1
a_1	0	0	0	0	1
a_0	0	0	0	0	0

Figure 3.7. Representation of a Perfect Guttman Scale

have been assigned the same numerical score. If all of the individuals with the same numerical score have the same level of the trait (i.e., belong to the same property of the natural variable), a scaling results. Usually this scaling is of at least the ordinal level because of the cumulative nature of the Guttman procedure. If the added characteristics that distinguish the different levels of the properties indicate an equal amount of change in the trait level from a psychological point of view, an interval scale is formed.

The relationship between the properties in the natural variable and the presence of characteristics is usually shown by a two-way table. Across the top of the table are listed the characteristics of the individuals that are used to classify them into the properties. Down the side of the table are listed the properties in the natural variable. In the body of the table, a "1" is placed at the intersection of a property and a characteristic if all persons in the property have the characteristic. An example of such a table is presented in Figure 3.7. If all persons in a property do not have the characteristic, a 0 is placed in the table. If a Guttman scale is present, the 1s in the table form a triangular pattern when the properties are arranged by order of magnitude of the trait being measured and the characteristics are arranged according to their level of severity.

In reality, we do not know the composition of the properties of the natural variable and must substitute the properties of the scaled variable for the rows of the table. In this case, the perfect triangular form may not be present. To the extent that the relationship between the scaled variable and the characteristics cannot be put into the triangular form, a scaling has not taken place. There are two possible reasons why a proper scaling might not be accomplished. First, the trait for which the measure is being developed may not easily be put into the hierarchical form required by the Guttman procedure. For example, holding liberal political beliefs does not mean that a person also holds all the beliefs of a person of conservative bent, even though the properties in the natural variable defined by political beliefs can generally be ordered along a continuum. The second reason a scaling may not be possible is that the properties of the scaled variable do not match the properties of the natural variable because of errors in the assignment of the numerical values. A person may not report a characteristic when it is really present, an observer may miss an important activity, or a record may be inaccurately kept.

In order to judge whether the scaled variable matches the natural variable sufficiently closely to form a scaling, Guttman (1950) suggested that a statistic called the coefficient of reproducibility be computed. This coefficient is simply the proportion of ones and zeros in the person-by-characteristic table that are in the appropriate places to produce the triangle form when the rows and columns have been ordered according to the total

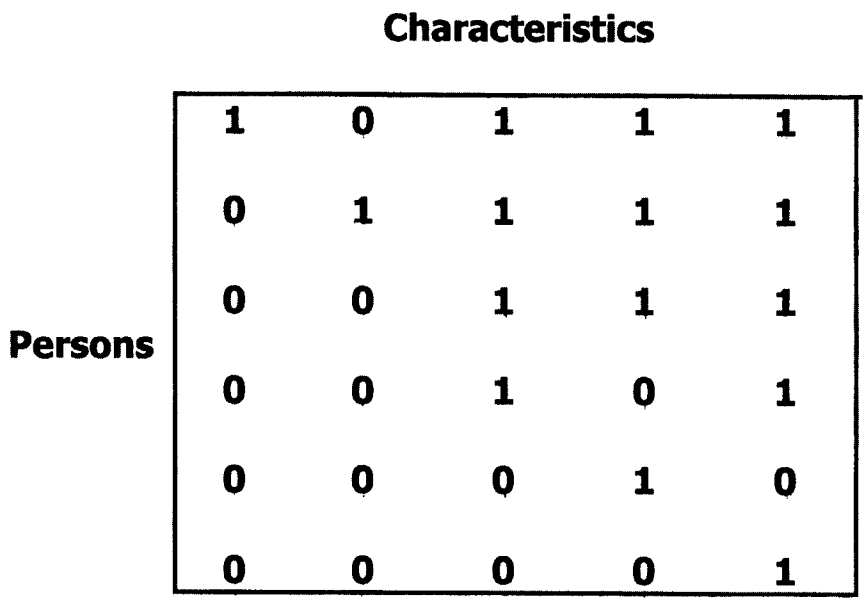


Figure 3.8. An Imperfect Guttman Scale

number of ones in them. If a one or zero is not in the appropriate place to produce the triangular form, the perfect Guttman scale will not be possible. The number of inappropriately placed zeros and ones is given by the number of ones below the diagonal and the number of zeros above the diagonal. In Figure 3.8, the number of inappropriately placed zeros and ones is 3 out of a total of 30 entries. The number of appropriate values is then 30 − 3 = 27. The coefficient of reproducibility is 27/30 = .90. Guttman felt that the coefficient of reproducibility should be at least .90 for the scaling to be considered reasonable.

Since Guttman's early work, procedures for determining the quality of a Guttman scale have become much more elaborate (see McIver & Carmines, 1981; see also and White & Saltz, 1957, for examples). However, these procedures are all conceptually related to the coefficient of reproducibility. They all check to determine whether the properties of the scaled variable have the necessary cumulative relationship with the observed characteristics.

An analysis of the assumptions of the Guttman Scalogram procedure can be used to determine whether this approach should be used to form a scale. The first step in this process is to evaluate the properties of the natural variable in question to determine whether they have the required cumulative relationship. If they do not, the Guttman procedure should not be used. One of the other methods given later in this chapter may be an appropriate alterative. If the necessary cumulative relationship does exist among the properties, the next step is to determine the characteristics that distinguish the properties of the natural variable. For example, a particular type of self-destructive behavior may distinguish one type of psychological disorder from another. This behavior can then be used as one of the items to assign the score to form the scaled variable. Usually a number of different behaviors are tentatively selected and only those that can be used to form the triangular pattern of responses shown above are used to form the scale. It is usually difficult to find more than five or six behaviors that have the required cumulative relationship.

Once the behaviors have been selected and data have been collected on the presence or absence of the behavior for a new group of individuals, a variant of the reproducibility coefficient is computed to determine whether a reasonable Guttman scale has been obtained. If the value of this coefficient is sufficiently high, the scaling is accepted.

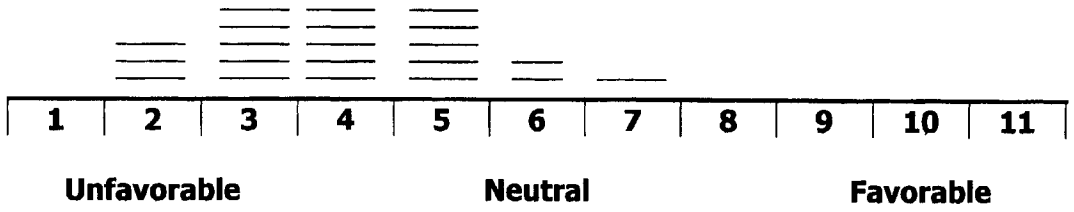


Figure 3.9. Judgments for a Statement with a Scale Value of approximately 4

Thurstone's Method of Equal-Appearing Intervals

Guttman's method of scale formation is fairly limited in its application because of the requirement of cumulative properties in the natural variable. Many natural variables do not have the cumulative property. Yet, the properties in the natural variable are distinguishable. In order to identify the persons belonging to each property, Thurstone (1927) developed a model of the interaction between a person and statements describing possible attitudes toward an object. Thurstone's model indicates that a person who is a member of a particular property set will endorse some attitude statements and not others. Persons in a different property set will endorse a different, although possibly overlapping, set of alternatives. Those persons who endorse similar sets of statements are hypothesized to belong to the same property set.

By merely sorting persons into categories on the basis of the responses to a set of attitude statements, a variable can be defined, but this variable does not contain any information about the level of an attitude toward an object. All that is obtained is groups of individuals, each of which is composed of persons with similar attitudes. In order to add the information about the relative level of attitude into the scaling, Thurstone suggested that the attitude statements themselves first be scaled.

The scaling of the attitude statements is performed in a very straightforward manner. A set of 11 properties is hypothesized for the natural variable of interest. These properties range from sets of statements that are very unfavorable to the object of interest to those that are very favorable. The sixth property is assumed to contain those statements that are neutral. The 11 properties can be arranged in order from very unfavorable through neutral to very favorable. This set of properties is the scaled variable for the attitude statements.

To determine which statements belong in each of the property sets, a number of judges (Thurstone used 300) are asked to sort the statements (usually over 100) into the appropriate sets (see Figure 3.9). The judges were instructed to perform this sorting on the basis of the favorableness or unfavorableness of the statements, not on the statements' level of agreement with the judges' position. If the statements differed solely on their degree of favorableness, and if the judges were totally consistent in their judgments, it would be expected that a statement would be put into the same property set by each judge. In reality, variations in the classifications are found which Thurstone called discriminal dispersion. In other words, the placement of the statements into the property sets is not perfectly reliable.

Because there usually is variation in the placement of the statements, a procedure is needed for forming a scaled variable using the statements. The procedure suggested by Thurstone first assigns the numbers 1 to 11 to the properties. When a statement is sorted into one of the properties by a judge, the corresponding number is assigned to the statement. After all the judges classify all the statements, the median and quartile deviations are computed using the numbers assigned to each statement.

If the quartile deviation for a statement is large, the statement is ambiguous in some sense, as indicated by the fact that the judges could not agree on the property set into which the statement should be placed. For a statement with a low quartile deviation, the median value is used as the scale value for the statement. The numbers that are assigned in this way are used to form the scaled variable for the statements. Two statements that have been assigned the same number are assumed to fall into the same property set. The statements and their associated scale values are used to produce the

instrument that is used to assign numerical values to individuals and thereby form the scaled variable for people.

Recall that individuals who endorse roughly the same sets of statements are assumed to come from the same property on the natural variable. If the mean scale value for these statements were computed, each person in the same property set would obtain the same mean scale value. Thus, Thurstone decided to form the scaled variable on people by assigning each person the mean scale value of the statements that they endorsed. In order to have a sufficient range of statements for all the persons who are being measured, Thurstone suggested producing the measuring instrument by selecting two statements from each of the 11 property sets. This results in an attitude-measuring device consisting of 22 statements. To use it, a person is asked simply to check the statements with which they agree. Their score is the average scale value for the statements endorsed.

Of course, there is some error in the procedure because persons can obtain approximately the same score although agreeing with different sets of statements. To the extent that this occurs, the scaled variable does not match the natural variable and the results of the scaling are invalid.

The level of scaling of the scores obtained from the Thurstone equal-appearing interval procedure depends on the quality of the judgments made concerning the attitude statements. Clearly a person who endorses favorable statements has a more positive attitude toward the topic in question than one who endorses less favorable statements. Therefore, the procedure results in at least an ordinal scale. Whether an interval scale is achieved depends on whether the 11 properties of the natural variable used to classify the attitude statements are equally spaced. Thurstone and Chave (1929) contended that the judges would subjectively make adjacent properties equally distant when they classified the items. To the extent that this conjecture is true, the scaling procedure results in an interval scale.

At this point, an example of the application of the Thurstone equal-appearing interval technique may prove useful in clarifying the steps in this procedure. Suppose it were desirable to develop a measuring instrument for determining attitudes toward nuclear power. The first step in the process would be to write more than 100 statements that vary in their degree of favorableness toward nuclear power. These should be statements of opinion, not fact. For example, the statement "Nuclear power will vastly improve the quality of life" is a favorable statement. "The use of nuclear power will destroy this country" is a negative statement. After these statements have been produced, several hundred individuals should be asked to rate the statements, based on whether the statements represent positive or negative attitudes toward nuclear power, using the 11-point scale. Next, the median and quartile deviations of each statement are computed. Those statements with large quartile deviations are dropped and, from the statements remaining, two statements are selected from each of the 11 categories. For this purpose, the median for the statement is used as a scale value. The resulting 22 statements form the measuring device for attitudes toward nuclear power.

To use the measuring instrument that has been developed, individuals are asked to check the statements with which they agree. Each person's attitude score is the average of the scale values for the statements they have checked.

Item Response Theory

Within the last 10 years, a new approach to the formation of scales of measurement has become popular. This approach, called item response theory or IRT (Lord, 1980), has been applied mostly to aptitude and achievement measurement, but it can also be used for other types of psychological assessment problems (see Kunce, 1979, for example). As with the Guttman and the Thurstone procedures, this scaling procedure assumes that the properties in the natural variable can be arranged along a continuum based on the magnitude of the trait possessed by the persons in each property set. If a test item is administered to the persons in one of these properties, this model assumes that all the persons will have the same probability of responding correctly or endorsing the item, but that they may not all give the same response to the item because of errors in measurement.

For example, suppose the item "Define democracy" is given to all the persons in a particular property set. Because of errors in the persons' responses, ambiguities in the question, problems in deciding whether the answer is correct or incorrect, and so on, some of the persons miss the item and others answer it correctly. However, the IRT model assumes that all persons in that property set have the same probability of answering the item correctly. Persons in a different property set will

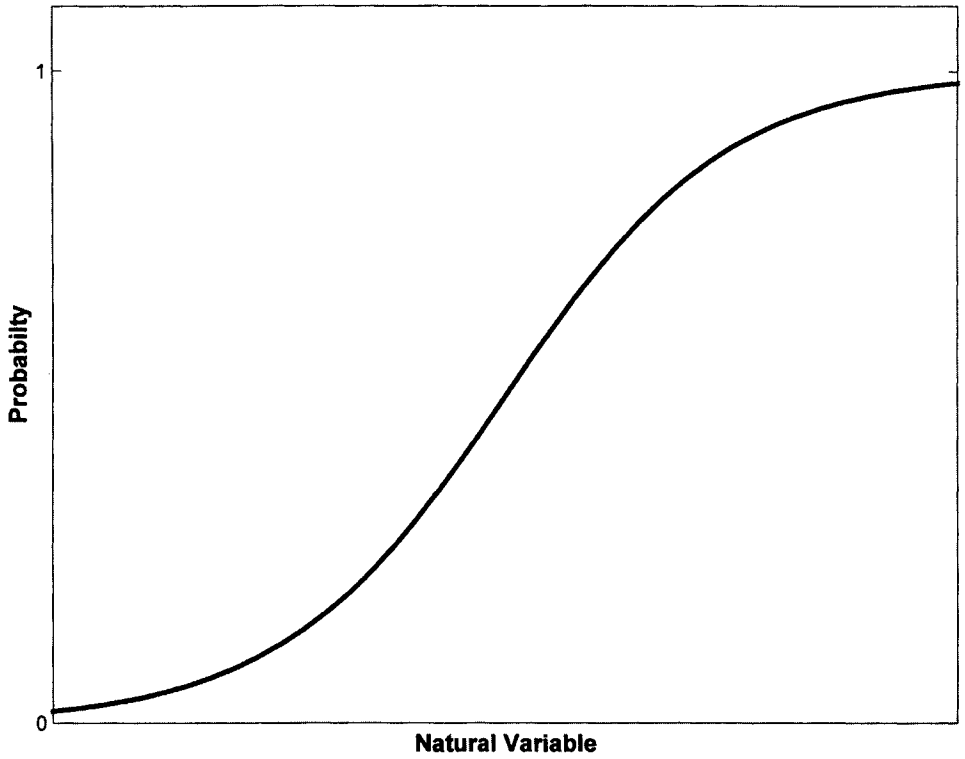


Figure 3.10. Probability of Correct Response for Properties of a Natural Variable

have a different probability of a correct response. If the probability of a person's correct response to an item is known, then the person can be classified into the appropriate property of the natural variable.

One of the basic assumptions usually made for IRT models is that if the properties are ordered according to the probability of correct response to an item, they are also ordered according to increasing trait level on the natural variable. That is, the probability of a correct response is assumed to have a monotonically increasing relationship to the trait of interest. Thus, if persons can be placed into the properties on the basis of the probability of correct response, at least an ordinal scale results.

If the natural variable has properties that are evenly spaced, the relationship between the ability properties and the probability of a correct response for persons in a property is assumed to have a particular form. The mathematical forms commonly used for this purpose are the one-parameter logistic model (Rasch, 1960), the two-parameter logistic model (Birnbaum, 1968), the three-parameter logistic model (Lord, 1980), and the normal ogive model (Lord, 1952). However, other forms, including ones that are nonmonotonic, are also being considered. The usual practice is to assume one of these forms for all the items in the measuring instrument to be produced. Figure 3.10 presents an example of the relationship typically found between the properties of a natural variable and the probability of a correct response.

The relationship shown here assumes that the probability of a correct response increases with increased magnitude of the trait. This type of model is most appropriate for items that have a single positive or correct response that is more likely for persons belonging to property sets defining high magnitudes of the trait. Other models are more appropriate for rating scale type items (Masters, 1988; Samejima, 1969). These models assume that the probability of a particular response to the rating-scale item first increases and then decreases as the level of the trait increases. The relationships for each rating-scale category are shown in Figure 3.11.

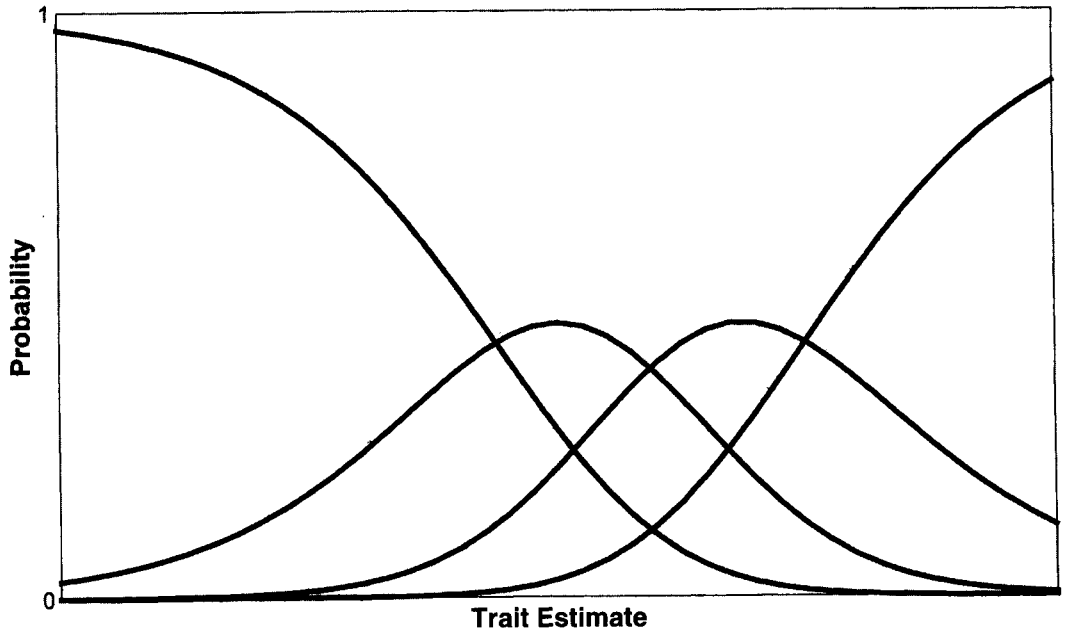


Figure 3.11. Probability of Various Responses for Properties in the Natural Variable

		Property A	Property B
Item	1	.1	.7
	2	.6	.9

Figure 3.12. Probability of a Correct Response to Two Items

As with the other two procedures described earlier, the purpose of the IRT analysis procedure is to determine the property in the natural variable to which a person belongs. If the probability of a person's responses to a single item could be observed, the determination of the appropriate property could be accomplished with one item. Of course, when a person is administered an item, a discrete score is

obtained (usually a 0 or a 1), and no probability is observed. Therefore, a person cannot be classified into a property using one item if IRT methods are used. Instead, an instrument composed of many items is administered, and a person is classified into the property that has the highest probability of giving the observed responses to the set of items.

For example, suppose two items are administered to the persons in two different property sets. Suppose further that the probability of a correct response for the two items for persons in the two properties is given in Figure 3.12.

If a person answered the first item incorrectly and the second correctly, that set of responses would have a probability of $(1 - .1) \times .6 = .54$ for those in property A but a probability of $(1 - .7) \times .9 = .27$ for those in property B. Because the probability of the responses was higher for property A, the examinee would be estimated to belong in property A. This principle of classification is called maximum likelihood trait estimation.

In practice, the properties of the scaled variables are indexed by numerical values, and the probability of a correct response to each item is determined by a mathematical formula. For example, the formula for the two-parameter logistic latent trait model is given by

$$P(x_{ij} = 1) = \frac{e^{a_i(\theta_j - b_i)}}{1 + e^{a_i(\theta_j - b_i)}}$$

where $P(x_{ij} = 1)$ is the probability of a correct response for person j on item i, e is the constant, 2.718..., θ_j is the ability of person j, and a_i and b_i are the item parameters that control the shape of the mathematical function. The estimate of θ_j indicates the property on the scaled variable to which the person belongs.

The values of a_i and b_i for an item are determined in much the same way as the scale values in Thurstone's equal-appearing interval procedure. A set of test items is administered to a large number of individuals and values of a_i and b_i are computed from the responses. The values of b_i are related to the proportion responding correctly to the item, and the values of a_i are related to the correlation between the item score and the values of the scaled variable (see Lord & Novick, 1968, for a discussion of this relationship). These values are determined from a scaling of each item along two dimensions, while the ability estimates are a scaling of the people responding to the items. The process of determining the values of a_i and b_i for a set of items is called *item calibration.*

The use of item response theory for the process of scaling is conceptually more complicated than use of the Guttman or Thurstone procedures because of the complex mathematics involved. In practice, however, the procedures are simpler because computer programs are available to perform the necessary analyses. Suppose we want to measure a personality characteristic by administering a series of items with instructions to the examinee to check those that apply to him or her. If this scale is to be developed using item response theory, the items would first be administered to a large number of individuals who vary on the trait of interest. The resulting data are analyzed using one of the available calibration programs to determine the item parameters. The calibration program is selected depending on which of the item response theory models is assumed. If the items are assumed to vary only in their rate of endorsement, the one-parameter logistic model may be appropriate, and a program such as BIGSTEPS (Wright & Linacre, 1992) can be used to obtain the item parameters. If the items are assumed to vary also in their discriminating power, the two-parameter logistic model is appropriate, and the BILOG program (Mislevy & Bock, 1990) can be used for calibration. If there is a non-zero base rate for positive responses to the items, the three-parameter logistic model is appropriate, and the ASCAL (Vale & Gialluca, 1985) programs can be used for item calibration. Finally, if a rating-scale item form is used, a program like MULTILOG (Thissen, 1986) may be appropriate. New programs are constantly being produced for these methods, and the literature should be checked for the most current versions for a particular model before the item calibration is performed.

After the items are calibrated, the items for the measuring instrument are selected. A procedure similar to Thurstone's can be used if the population to be measured ranges widely on the trait of interest. The items may also be selected on both level of difficulty and discriminating power if high precision at a point on the score scale is required. Many computer programs also give measures of fit between the models and the data. These fit measures may also be used to select items to insure that the model being used is appropriate for the response data.

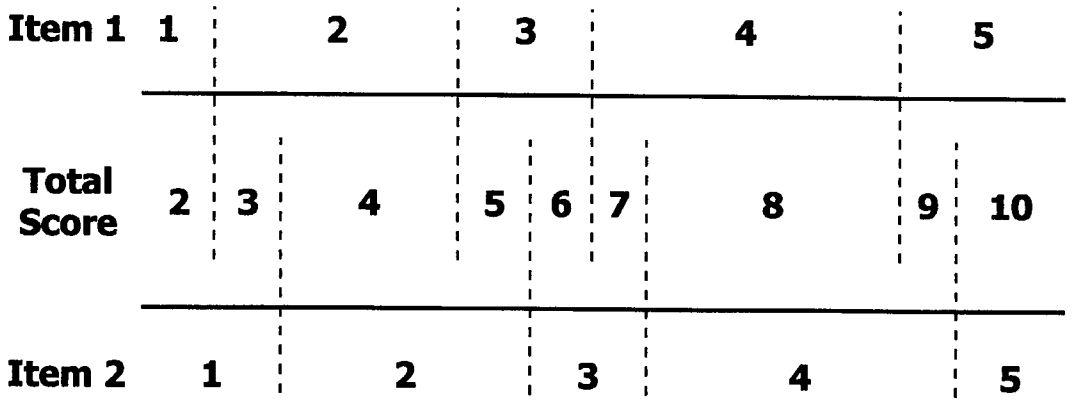


Figure 3.13. Score Categories for a Two-Item Likert Scale

Once the items for the instrument are selected, an estimation program or a conversion table can be used to obtain an estimate of the level of the trait for each person. In general, as with the Guttman and Thurstone procedures, those persons with similar patterns of responses will receive similar trait estimates. These trait estimates form the scaled variable for the trait in question. As with any of the other procedures, this scaled variable must be checked to determine whether it matches the natural variable and therefore yields a reliable and valid measure.

Likert Scaling Technique

Another commonly used procedure for forming attitude measuring instruments was developed by Likert (1932). This procedure also begins by assuming a natural variable with properties that can be ordered according to the magnitude of the trait possessed by the persons in each property set. The form of the item used by the Likert procedure is a statement concerning the concept in question, followed by five answer-choices ranging from *strongly agree* to *strongly disagree.* It is assumed that the five answer-choices divide the natural variable into five classes that are ordered with respect to the attitude toward the concept.

If only one item is used in the measuring instrument, the five categories are numbered from 1 (strongly disagree) to 5 (strongly agree) and each person is assigned the score corresponding to the response selected. If the statement being rated has a negative connotation, the scoring is reversed. The score assignment forms the scaled variable for this procedure.

In reality, more than one item is usually used with the Likert procedure. Each of these items is assumed to divide the natural variable in a similar, but not exactly the same, way. Thus, for two items the natural variable may be divided as shown in Figure 3.13. In this figure, the boundaries between the sets of properties are not exactly aligned. Therefore, it is possible for one person to respond with *strongly agree* responses to two items, while another person may respond with *strongly agree* and *agree.* The latter person has a slightly lower trait level than the former. To indicate this fact on the scaled variable, the scores on the two items are simply summed. The first person receives a score of 10 on the scaled variable while the second receives a 9.

As more items are added to the instrument, the score for each person is obtained by simply summing the numbers assigned to each response category. Because the division of the natural variables into five categories is seldom exactly the same, each additional item brings about a greater subdivision of the natural variable. If 20 items were used in an instrument, the natural variable could be divided into as many as $(5 - 1)20 + 1 = 81$ categories. Each of these would be assigned a score which is the sum of the item scores. The persons with the same score would constitute properties in the scaled variable. To the extent that the properties in the scaled variable match those of the natural variable, a valid scaling is the result.

Although for the 20-item example given above each score is assumed to result from only one pattern of responses (one region on the natural variable), in reality there are many ways to obtain the same score. A total of $5^{20} = 9.5 \times 10^{13}$ patterns of responses are possible. To the extent that categories other than the 81 consistent categories mentioned above are present, the underlying model does not hold. These additional response patterns are usually attributed to errors of measurement and result in a mismatch between the scaled score and natural variable reducing the reliability of the results of the scaling. The Likert procedure tends to be robust to the violations, however, and items that result in many inappropriate responses are usually removed at a pretesting phase in instrument construction. This is done by correlating the score for each item with the total score on the instrument and dropping those that have a low correlation.

The level of scaling obtained from the Likert procedure is rather difficult to determine. The scale is clearly at least ordinal. Those persons from the higher level properties in the natural variable are expected to get higher scores than those persons from lower properties. Whether an interval scale is obtained depends on a strong assumption. In order to achieve an interval scale, the properties on the scaled variable have to correspond to differences in the trait on the natural variable. Because it seems unlikely that the categories formed by the misalignment of the five response categories will all be equal, the interval scale assumption seems unlikely. However, as the number of items on the instrument is increased, each property of the scaled variable contains a smaller proportion of the population, and the differences in category size may become unimportant. Practical applications of the Likert procedure seem to show that the level of scaling for this method is not an important issue. That is, treating the scores as if they were on an interval scale does not seem to cause serious harm.

An example of the construction of an attitude scale using the Likert procedure should clarify all the issues discussed. As with the Thurstone procedure, the first step in producing a Likert-scaled attitude instrument is to write more statements about the concept of interest than are expected to be used. In this case, about twice as many statements as are to be used should be enough. These should be statements of opinion, not fact, and both positive and negative statements should be included in approximately equal numbers. The five response categories (strongly agree, agree, neither agree nor disagree, disagree, strongly disagree) are then appended to each statement. For positive statements the categories are scored 5, 4, 3, 2, and 1, and for negative statements they are scored 1, 2, 3, 4, and 5.

If a measure of body image were desired, one item might be the following:

I have a well-proportioned body.

(a) strongly disagree
(b) disagree
(c) neither agree nor disagree
(d) agree
(e) strongly agree.

A negatively phrased item might be

I am noticeably overweight.

(a) strongly disagree
(b) disagree
(c) neither agree nor disagree
(d) agree
(e) strongly agree.

For the first item, (a) would be scored as 1, (b) as 2, (c) as 3, (d) as 4, and (e) as 5. For the second item, the scoring would be reversed: (a) 5, (b) 4, (c) 3, (d) 2, (e) 1.

The attitude items are next tried on a sample of approximately 100 individuals who represent the population of interest. For each statement, the correlation is computed between the score on the statement and the sum of the scores on all the statements. If the correlation is negative, the phrasing for the statement has probably been misclassified as to whether it is positive or negative. If it has not been misclassified, the statement should be deleted from the instrument as being ambiguous. Statements with low correlations (less than .3) are also dropped from consideration because the correlation indicates that these statements do not form a scaled variable that is consistent with the other items. From the items that meet the above criteria, 10 to 20 are selected with about equal numbers that are positively and negatively phrased. Both positively and negatively phrased items are needed to reduce response bias. The items selected constitute the measuring device.

REQUIREMENTS FOR SCALE FORMATION

In the formation of measurement devices, there is a common starting point for all techniques. In all cases, a natural variable is hypothesized to exist. Without this initial step, the concept of instrument validity is meaningless because the focus of the instrument is unknown. Once the natural variable has been defined, the scale-construction task becomes one of devising a method for determining which persons belong in each of the property sets of the natural variable. Conceptually this could be done by developing a detailed description of the persons in each property set and then observing each individual until he or she could be accurately classified into a property. This is essentially the procedure that is used for some infant intelligence-scales.

The more common procedure is to develop a series of items and use these items to obtain a highly structured sample of behavior (i.e., item responses). Those persons who exhibit similar behavior are assigned the same numerical score and are assumed to belong to the same property of the natural variable. For the Guttman procedure, behaviors that are cumulative in nature are used, and the numerical assignment rule is based on a count of the number of behaviors present. For the Thurstone procedure, the behavior is the endorsement of an attitude statement, and the numerical-assignment rule is based on the average scale-value for the items endorsed. The IRT approach is very similar to Thurstone's procedure in that the items are first scaled, and the results are then used to obtain an estimate of the trait level for a person. For Likert's procedure, the behavior used in the scaling is the rating of attitude statements, and the numerical assignment is based on the sum of the ratings.

Note that for all of the procedures some sort of prescreening of items is required. The Thurstone and IRT procedures require a scaling of items and some measure of fit to the underlying model. The Guttman and Likert procedures also use a measure of fit: the reproducibility coefficient in the former case and the item total score correlation in the latter. The existence of these procedures for evaluating the quality of the items in the measuring instruments reflects the fact that merely assigning numbers to persons does not result in the meaningful measurement of a trait. The numbers must be assigned in a way that is consistent with the natural variable. The procedures described for each of the methods provide a check on the consistency. Even when the scales produced by these methods are shown to be internally consistent, this fact does not insure that measurements obtained from the instruments are valid. The measures must still be shown to interact with other variables in ways suggested by an understanding of the natural variable. If this is not the case, a good measure has been developed of some unknown quality.

REFERENCES

Adams, E. W., Fagat, R., & Robinson, R. E. (1965). A theory of appropriate statistics. *Psychometrika,* 30, 99–127.

Birnbaum, A. (1968). Some latent trait models and their use of inferring an examinee's ability. In F. M. Lord & M. R. Novick, (Eds.), *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin,* 110(2), 305–314.

Cattell, R. B., Eber, H. W., & Tatsuoka, M. M. (1970). *Handbook for the sixteen personality factor questionnaire.* Champaign, IL: Institute for Personality and Ability Testing.

Goldstein, H. (1994). Recontextualizing mental measurements. *Educational Measurement: Issues and Practice,* 13(1), 16–19, 43.

Guilford, J. P., & Hoepfner, R. (1971). *The analysis of intelligence.* New York: McGraw-Hill.

Guttman, L. L. (1950). The basis for scalogram analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. W. Lazarsfeld, S. A. Star, & J. A. Clausen, *Studies in social psychology—World War II* (Vol. 4). Princeton, NJ: Princeton University Press.

Holland, J. L. (1966). *The psychology of vocational choices.* Waltham, MA: Blaisdell.

Krantz, D. H., Luce, R. D., Suppes, P. & Tversky, A. (1971). *Foundations of measurement: Vol. I. Additive and polynomial representations.* New York: Academic Press.

Kunce, C. S. (1979). *The Rasch one-parameter logistic model applied to a computerized, tailored administration of Mini-Mult scales.* Unpublished doctoral dissertation, University of Missouri Columbia.

Labovitz, S. (1970). The assignment of numbers to rank order categories. *American Sociological Review, 35,* 515–524.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 140,* 44–53.

Lord, F. M. (1952). A theory of test scores. *Psychometrika, Monograph,* 1.

Lord, F. M. (1980). *Applications of item-response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Masters, G. N. (1988). Measurement models for ordered-response categories. In Langeheine, R., & Rost, J. (Eds.), *Latent trait and latent class models.* New York: Plenum Press.

Mclver, J. P., & Carmines, E. G. (1981). *Unidimensional scaling.* Beverly Hills, CA: Sage Publications.

Mislevy, R. J., & Bock, R. D. (1990). BILOG 3: Item Analysis and Test Scoring with Binary Logistic Models [Computer Software and Manual]. Chicago: Scientific Software, Inc.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research.

Rozeboom, W. W. (1966). Scaling theory and the nature of measurement. *Synthese, 16,* 170–233.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika* (Monograph Supplement No. 17).

Stevens, S. S. (1959). Measurement. In C. W. Churchsman (Ed.), *Measurement: Definitions and theories.* New York: Wiley.

Stouffer, S. A. (1950). An overview of the contributions to scaling and scale theory. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. W. Lazarsfeld, S. A. Star, & J. A. Clausen, *Studies in social psychology—World War II* (Vol. 4). Princeton, NJ: Princeton University Press.

Thissen, D. (1986). *MULTILOG version 5 user's guide.* Mooresville, IN: Scientific Software.

Thurstone, L. L. (1927). Psychophysical analysis. *American Journal of Psychology, 38,* 368–389.

Thurstone, L. L., & Chave, E. J. (1929). *The measurement of attitude.* Chicago: University of Chicago Press.

Vale, C. D., & Gialluca, K. A. (1985, November). *ASCAL: A microcomputer program for estimating logistic IRT item parameters* (Research Report ONR-85-4). St. Paul, MN: Assessment Systems Corporation.

White, B. W., & Saltz, E. (1957). Measurement of reproducibility. *Psychological Bulletin, 54(2),* 81–99.

Wright, B. D., & Linacre, J. M. (1993). *A user's guide to BIGSTEPS: A Rasch model computer program.* Chicago: MESA Press.

This Page Intentionally Left Blank

PART III

ASSESSMENT OF INTELLIGENCE

This Page Intentionally Left Blank

CHAPTER 4

ASSESSMENT OF CHILD INTELLIGENCE

Kristee A. Beres
Alan S. Kaufman
Mitchel D. Perlman

INTRODUCTION

Over the centuries, many definitions of intelligence have been postulated attempting to explain this elusive construct. This chapter provides a context in which to understand children's intelligence by including a chronology of historical landmarks, by expounding on popular intelligence measures, and by looking at future trends in intelligence testing. In the first part, a brief history of mental measurement is provided. The second part is partitioned into two subsections and describes tests currently available to measure children's intelligence (preschool through adolescence). The first subsection provides detail on five major intelligence tests for children: Wechsler Intelligence Scale for Children-Third Edition (WISC-III); Kaufman Assessment Battery for Children (K-ABC); Stanford-Binet: Fourth Edition (SB-IV); Kaufman Adolescent and Adult Intelligence Test (KAIT); and Woodcock-Johnson Psycho-Educational Battery-Revised: Tests of Cognitive Ability (WJ-R). The second subsection describes other popular tests of mental measurement. The final part of this chapter provides a sample case report that combines the WISC-III, WJ-R, and other tests in the assessment of a male adolescent with academic and emotional difficulties.

A CHRONOLOGY

Classifying and categorizing individuals is by no means a novel concept. From the beginning, so we are told, Adam was classified as being "man" and Eve was classified as being "woman." The fascinating but evasive concept of "intelligence," while called many things, is a concept that has been with us throughout time and has been used both positively and negatively to set mankind apart from beasts and to differentiate within the broad category of mankind itself. Measuring intelligence is a complex and historically sensitive issue that has often been misused.

Measuring intelligence emerged out of both theoretical interest and societal need. Theoretical interest relates to the desire to understand individual differences. Societal need involves the utilization of this understanding to solve practical problems. The evolution of the measurement of intelligence, therefore, did not emerge in a vacuum, but rather, by the interplay in the development of several paradigms: psychology, sociology, psychometrics, and law. The brief historical review that follows highlights landmark events in those areas and relates them to the birth and development of intelligence.

It is difficult to pinpoint when intelligence testing began, but certainly it was conceived prior to

the 1800s. Aristotle, for example, attempted to understand how people behaved by dividing mental functions into cognitive (cybernetic) and dynamic (orectic) categories. Cybernetic functions are thought processes; orectic functions are emotional and moral processes (Das, Kirby, & Jarman, 1979). Since the time of Aristotle, whenever personality is conceptualized, attempts are still made to keep separate these two functions, but the difficulty in doing so has been well recognized (Das et al., 1979; Perlman, 1986; Shapiro, 1965). Others, such as Firtzherbert in 1510, Huarte in 1575, Swinbourne in 1610, and Thomasius in 1692 gave credence to testing one's cognition. They proposed various definitions of cognition and gathered information on human mentality (Sattler, 1988).

The 1800s ushered in several important advances in intelligence testing. In fact, interest in cognition and in the measurement of cognition in the 19th century was part of the scientific movement that brought psychology into being as a separate and respected discipline. Esquirol (1828) focused on distinguishing between mental retardation and emotional disturbance. It was he who coined terms such as *imbecile* and *idiot* to describe diverse levels of mental deficiency. He pointed out that idiots never developed their intellectual capacities, whereas mentally deranged persons lost abilities they once possessed. After studying different methods of measuring intelligence, he concluded that language usage was the most dependable criterion, a philosophy prevailing in most intelligence measures today.

Seguin's (1907) philosophy was quite different. He stressed sensory discrimination and motor control as aspects of intelligence. The Seguin Form Board, which requires rapid placement of variously shaped blocks into their correct holes, is an application of his theory. Many of the procedures he developed were adopted or modified by later developers of performance and nonverbal tasks. It was during this time that intelligence testing and education had their first formal courtship, for Seguin convinced the authorities of the desirability of educating the "idiots" and "imbeciles." He is credited for beginning the first school for the feebleminded and for being the author of the first standard book dealing with educating and treating them (Pintner, 1949). Also, his methods provided the inspiration for Maria Montessori's approach to education.

Galton's approach was similar to Seguin's in that he also stressed discrimination and motor control. In accordance with his commitment to the notion that intelligence is displayed through the uses of the senses—sensory discrimination and sensory motor coordination—he believed that those with the highest IQ should also have the best discriminatory abilities. Therefore, he developed tasks such as weight discrimination, reaction time, strength of squeeze, and visual discrimination. Galton is credited both with establishing the first comprehensive individual intelligence test and with influencing two basic notions of intelligence: the idea that intelligence is a unitary construct (which eventually led others to postulate the notion of general intelligence, or the "g" factor), and that individual differences in intelligence are largely genetically determined (possibly influenced by the theory of his cousin, Charles Darwin) (Cronbach, 1970; Das, Kirby, & Jarman, 1979; Pintner, 1949). Perhaps Galton's greatest contributions to the field of intelligence testing, were two crucial psychometric concepts that he originated: regression to the mean and correlation. His concepts allowed for studying intelligence over time, as well as for studying relationships between intelligence scores of parents, children, etc. (a concept for which, on the basis of Galton's work, Pearson developed the product-moment correlation and other related formulas).

James McKeen Cattell (1888) (as cited in Pintner, 1949), an assistant in Galton's laboratory, brought Galton's concepts to the United States. He shared his mentor's philosophy that intelligence is best measured by sensory tasks, but expanded his use of "mental tests" (a term coined first by Cattell in the literature) to include standardized administration procedures. He pleaded for standardized procedures, and urged the necessity for the establishment of norms. Cattell's valuable contribution to psychology was that he took the assessment of mental ability out of the field of abstract philosophy and showed that mental ability could be studied experimentally and practically. Under Cattell's direction, the Pearson correlation technique seems to have been used for the first time for comparison of test with test, and tests with college grades.

By the late 1800s diverse notions of intelligence were conjectured, standardized procedures and norms were urged, and interest in classification had been implemented. Societal need provided the final impetus which led to the development of the individually administered Binet-Simon Scale in

1905. With the specific appointment by the French minister of public instruction to study the education of retarded children, the notion to separate mentally retarded and normal children in the Paris public schools arose. Binet, assisted by Theophile Simon and Victor Henri, rejected Galton's notions of what made up intelligence and proposed that tasks utilizing higher mental processes (memory, comprehension, imagination, etc.) would be more effective measures. Binet did, however, retain Galton's idea of general intelligence ("g"), which is reflected in his battery. This 1905 scale might be considered the first *practical* intelligence test.

The Binet has gone through a number of modifications and revisions throughout the years including the eventual introduction of the term "intelligence quotient" (IQ) in Terman's 1916 version, the Stanford-Binet. This ratio IQ was computed by dividing mental age by chronological age, and multiplying by 100. While these single IQ scores have become a popular means of classifying individuals, it is a clear departure from Binet's notion of intelligence as "a shifting complex of inter-related functions" (Tuddenham, 1962, p. 490). In fact, some doubt whether Binet would have accepted the concept of a single IQ score even with Terman's elaborate standardization (Wolf, 1969). In 1986, Thorndike, Hagen, and Sattler developed a completely modified version of the Stanford-Binet, the SB-IV. The test incorporates Wechsler's subtest format, and departs so much from the previous test that one wonders whether it merits the same name.

An important note is that this first major intelligence test battery, the Binet, arose to classify individuals. Classification has been fundamental to the history of mental assessment. It is no wonder that this philosophy continues today with such fervor, despite earnest attempts to move beyond single IQ scores in a desire to individualize profile analysis.

Like Binet, David Wechsler included the concept of global intelligence in his Wechsler-Bellevue Scale (published in 1939). Instead of having one global score, his battery included three separate IQ scores, a Verbal IQ, a Performance IQ, and a Full Scale IQ. The Full Scale IQ, for Wechsler, is an index of general mental ability ("g").

While the formats from the original Stanford-Binet and the Wechsler Scales differ considerably, the subtests themselves do not. Wechsler's tasks weren't novel concepts at all, but rather, were borrowed from other tests of cognitive abilities. In many ways, Wechsler combined the philosophies of Esquirol (1828) and Seguin (1866/1907), and the psychometrics of Cattell and Terman. As equal components of intelligence, the Verbal Scale roughly capitalizes upon a person's language abilities, (which expresses Esquirol's philosophy), while the Performance Scale roughly capitalizes upon a person's nonverbal and motoric abilities (as in Seguin's view). Wechsler's main ideas for the verbal tasks were the Stanford-Binet and the Army Group Examination Alpha. Ideas (and often specific items) for the performance tasks came primarily from the Army Group Examination Beta and the Army Individual Performance Scale Examination. The Army Alpha and Army Beta tests, published by Yerkes in 1917, were group administered intelligence tests developed to assess United States military recruits. The Wechsler Scales have gone through a number of revisions, but the basic test has remained structurally intact.

Although the Stanford-Binet and Wechsler scales were powerful tools to measure cognitive ability, theories of intelligence continued to be introduced and refined. In 1936, Piaget published *Origins of Intelligence.* He conceived of intelligence as a form of biological adaptation of the individual to the environment. Just as living organisms adapt to their environments biologically, individuals adapt to their environment through cognitive growth. Cognitive stages, therefore, emerge as a function of psychological structures reorganizing and/or developing out of organismic-environmental interactions (Piaget, 1950). Piaget's model of intelligence is developmental and hierarchical in that he believes individuals pass through four predetermined, invariant stages of cognition, each more complex than the preceding one: *sensorimotor* (birth–2 years), *preoperational thought* (2 years–7 years), *concrete operations* (7 years–11 years), *formal operations* (11 years–adult).

With the advancement of psychometrics, factor analytic theories of intelligence emerged espousing either a general-factor theory ("g") or a multiple-factor theory. Each method can be reduced to the other by either accepting the unrotated first factor, or by rotating the factors by various methods. Within this domain, J. P. Guilford (1959) developed a complex multifactor theory. His three-dimensional Structure of Intellect model (Guilford, 1967; Guilford & Hoepfner, 1971) posited five different operations, four types of content, and six products resulting in 120 possible factors ($5 \times 4 \times 6$). An even larger number of factors is possible

based on Guilford's (1988) modification of his model, in which he subdivided both on operation (Memory) and a content area (Figural) into two parts.

Also following the factor-analytic model were Raymond B. Cattell and John Horn (Cattell, 1963; Horn & Cattell, 1966, 1967), who postulated a structural model that separates fluid intelligence from crystallized intelligence. Fluid intelligence traditionally involves relatively culture-fair novel tasks and taps problem-solving skills and the ability to learn. Crystallized intelligence refers to acquired skill: knowledge and judgments that have been systematically taught or learned via acculturation. The latter type of intelligence is highly influenced by formal and informal education and often reflects cultural assimilation. Tasks measuring fluid ability often involve more concentration and problem solving than crystallized tasks, which tend to tap retrieval and application of general knowledge.

Another theoretical approach conceptualizing intelligence is an information-processing model focusing on the strategies individuals use to complete tasks successfully. Within this approach is the neuropsychological processing model which originated with the neurophysiological observations of Alexsander Luria (1966a, 1966b, 1973, 1980) and Roger Sperry (1968), the psychoeducational research of J. P. Das (1973; Das et al., 1975, 1979; Naglieri & Das, 1988, 1990), and the psychometric research of A. S. and N. L. Kaufman (1983c). This model possesses several strengths relative to previous models in that it (a) provides a unified framework for interpreting a wide range of important individual difference variables; (b) rests on a well-researched theoretical base in clinical neuropsychology and psychobiology; (c) presents a processing, rather than a product-oriented, explanation for behavior; and (d) lends itself readily to clear remedial strategies based on relatively uncomplicated assessment procedures (Das et al., 1979; Kaufman & Kaufman, 1983c; McCallum & Merritt, 1983; Naglieri & Das, 1988, 1990; Perlman, 1986).

This neuropsychological processing model describes two very distinct types of processes that individuals use to organize and process information received in order to solve problems successfully: successive or sequential, analytic-linear processing versus holistic/simultaneous processing, (Levy & Trevarthen, 1976; Luria, 1966a). These processes have been identified by numerous researchers in diverse areas of neuropsychology and cognitive psychology (Perlman, 1986). From Sperry's cerebral-specialization perspective, these processes represent the problem-solving strategies of the left hemisphere (analytic/sequential) and the right hemisphere (gestalt/holistic). From Luria's theoretical approach, successive and simultaneous processes reflect the "coding" processes that characterize "Block 2" functions.

Regardless of theoretical model, successive processing refers to the processing of information in a sequential, serial order. The essential nature of this mode of processing is that the system is not totally surveyable at any point in time. Simultaneous processing refers to the synthesis of separate elements into groups. The essential nature of this mode of processing is that any portion of the result is at once surveyable without dependence on its position in the whole. The model assumes that the two modes of processing information are available to the individual. The selection of either or both modes of processing depends on two conditions: (a) the individual's habitual mode of processing information as determined by social-cultural and genetic factors, and (b) the demands of the task (Das et al., 1975).

Many different theories and models of cognition underlie intelligence tests. However, it appears that recently factor-analytic and neuropsychological models have had a strong impact on test development in the field of intelligence testing. For example, the WJ-R utilizes a factor-analytic model while the K-ABC is based on a neuropsychological model. Both of these models can be translated into a unique way of viewing and assessing intelligence, and, when properly utilized, they have the ability to provide the examiner with a wealth of information about an individual's cognitive functioning.

Many laws and judicial decisions have addressed the need for the development of nonbiased IQ tests for those having various learning deficiencies and those in minority groups. These laws and opinions underscore some of the controversy surrounding the appropriate use of intelligence tests and place ethical, if not legal, responsibility on clinicians for determining the adequacy and appropriateness of intelligence tests for children. The American Psychological Association clearly addresses this issue in their Ethical Principles of Psychologists and, under Principle 2-Competence, requires clinicians to recognize differences among people (age, sex, socioeco-

nomic, and ethnic backgrounds) and to understand test research regarding the validity and the limitations of their assessment tools (American Psychological Association, 1990).

CURRENT MEASURES

Intelligence tests are administered for a variety of reasons including *identification* (of mental retardation, learning disabilities, other cognitive disorders, giftedness), *placement* (gifted and other specialized programs), and as a cognitive adjunct to a *clinical evaluation.* The Wechsler Scales, Kaufman Scales, Stanford-Binet, and Woodcock-Johnson battery, are probably the most commonly used and most widely accepted individual intelligence measures. Administration of one of these more traditional measures is recommended for the assessment of intelligence when a child has the necessary physical capacities to respond to test questions, when the child meets age requirements of the test, and when there are no time restraints. When verbal responses cannot be elicited from a child, when sensory or motor impairments or both place limits on a child's performance, or when time is at a premium, other measures become necessary. A review of these tests follows with a summary of other general cognitive measures and tests designed for special populations (infants and preschoolers, people with mental retardation, hearing and language impairment, visual impairment, orthopedic impairment, cultural minorities). The list of measures reviewed is by no means exhaustive, but represents the ones that are commonly used in the field today.

Five Major Intelligence Scales

Wechsler Intelligence Scale for Children-Third Edition (WISC-III)

Theory. Wechsler (1974) puts forth the definition that "intelligence is the overall capacity of an individual to understand and cope with the world around him" (p. 5). His tests, however, were not predicated on this definition. Tasks developed were not designed from well-researched concepts exemplifying his definition. In fact, as previously noted, virtually all of his tasks were adapted from other existing tests.

Like the Binet, Wechsler's definition of intelligence also ascribes to the conception of intelligence as an overall global entity. He believed that intelligence cannot be tested directly, but can only be inferred from how an individual thinks, talks, moves, and reacts to different stimuli. Therefore, Wechsler did not give credence to one task above another, but believed that this global entity called intelligence could be ferreted out by probing a person with as many different kinds of mental tasks as one can conjure up. Wechsler did not believe in a cognitive hierarchy for his tasks, and he did not believe that each task was equally effective. He felt that each task was necessary for the fuller appraisal of intelligence.

STANDARDIZATION AND PROPERTIES OF THE SCALE

The WISC-III was standardized on 2,200 children ranging in age from 6 through 16 years. The children were divided into eleven age groups, one group for each year from 6 through 16 years of age. The median age for each age group was the sixth month (e.g., 7 years, 6 months). The standardization procedures followed the 1980 U.S. Census data and the manual provides information by age, gender, race or ethnicity, geographic region, and parent education. "Overall, the standardization of the WISC-III is immaculate....a better–standardized intelligence test does not exist (Kaufman, 1994, p.351).

The WISC-III yields three IQ scores, a Verbal Scale IQ, a Performance Scale IQ, and a Full Scale IQ. All three are standard scores (mean of 100 and standard deviation of 15) obtained by comparing an individual's score with those earned by the representative sample of age peers. The WISC-III also yields four factor indexes, Verbal Comprehension, Perceptual Organization, Freedom from Distractibility and Processing Speed. The first two factors are in the cognitive domain, whereas the distractibility dimension is in the behavioral or affective domain. "The fourth factor seems to bridge the two domains; "processing" implies cognition, but "speed" has behavioral as well as cognitive components" (Kaufman, 1994, pp. 104, 105). The Verbal Comprehension Index measures abilities related to verbal conceptualization, knowledge, reasoning, and the ability to express ideas in words. The Freedom from Distractibility Index measures number ability and sequential process-

Table 4.1. Summary of Seven Steps for Interpreting WISC-III Profiles

Step 1. Interpret the Full Scale IQ

Convert it to an ability level and percentile rank and band it with error, preferable a 90% confidence interval (about ± 5 points).

Step 2. Determine if the Verbal-Performance (V-P) IQ Discrepancy Is Statistically Significant

Overall values for V-P discrepancies are *11 points* at the .05 level and *15 points* and the .01 level. FOr most testing purposes, the .05 level is adequate.

Step 3. Determine if the V-P IQ Discrepancy Is Interpretable–Or if the VC and PO Factor Indexes Should be Interpreted Instead

Ask four questions about the Verbal and Performance Scales

Verbal Scale

1. Is there a significant difference ($p < .05$) between the child's standard scores in VC versus FD?

Size Needed for Significant (VC-FD) =13+ points

2. Is there abnormal scatter (highest minus lowest scaled score) among the five Verbal subtests used to compute V-IQ?

Size Needed for Abnormal Verbal Scatter =7+ points

Performance Scale

3. Is there a significant difference ($p < .05$) between the child's standard scores on PO Versus PS?

Size Needed for Significant (PO-PS) =15+ points

4. Is there abnormal scatter (highest minus lowest scaled score) among the five Performance subtests used to compute P-IQ?

Size Needed for Abnormal Performance Scatter =9+ points

If all answers are no, the V-P IQ discrepancy is interpretable. If the answer to one or more questions is yes, the V-P IQ discrepancy may not be interpretable. Examine the VC-PO discrepancy. Overall values for VC-PO discrepancies are *12 points* at the .05 level and *16 points* at the .01 level.

Determine if the VC and PO indexes are unitary dimensions:

1. Is there abnormal scatter among the four VC subtests?

Size Needed for Abnormal VC Scatter =7+ points

2. Is there abnormal scatter among the four PO subtests?

Size Needed for Abnormal PO Scatter =8+ points

If the answer to either question is yes, then you probably shouldn't interpret the VC-PO index discrepancy–unless the discrepancy is to big to ignore (see Step 4). If both answers are no, interpret the VC-PO differences as meaningful.

Step 4. Determine if the V-P IQ Discrepancy (Or VC-PO Discrepancy) is Abnormally Large

Differences of at least *19 points* are unusually large for both the V-P and VC-PO discrepancies. Enter the tale with the IQ or indexes, whichever was identified by the questions and answers in Step 3.

If neither set of scores was found to be interpretable in Step 3, they may be interpreted anyway if the magnitude of the discrepancy is unusually large (19 + points).

Step 5. Interpret the Meaning of the Global Verbal and Nonverbal Dimensions and the Meaning of the Small Factors

Study the information and procedures presented in Chapter 4 (verbal/nonverbal) and Chapter 5 (FD and PS factors). Chapter 5 provides the following rules regarding when the FD and PS factors have too much scatter to permit meaningful interpretation of their respective indexes:

1. Do not interpret the FD index if the Arithemetic and DIgit Span scaled scores differ by *4 or more points*.
2. Do not interpret the PO index if the Symbol Search and Coding scaled scores differ by *4 or more points*.

Step 6. Interpret Significant Strengths and Weaknesses in the WISC-III Subtest Profile

If the V-P IQ discrepancy is less than 19 points, use the child's mean of all WISC-III subtests administered as the child's midpoint.

If the V-P IQ discrepancy is 19 or more points, use the child's mean of all Verbal subtests as the midpoint for determining strengths and weaknesses on Verbal subtests, and use the Performance mean for determining significant deviations on Performance subtests.

Using either the specific values in Table 3.3 of *intelligent Testing with the WISC-III* (Kaufman, 1994), rounded to the nearest whole number, or the following summary information of determining significant deviations:

± 3 points: Information, Similarities, Arithmetic, Vocabulary

± 4 points: Comprehension, Digit Span, Picture Completion, Picture Arrangement, Block Design, Object Assembly, Symbol Search

± 5 points: Coding

(*continued*)

Table 4.1. (Continued)

Step 7. Generate Hypotheses about the Fluctuations in the WISC-III Subtest Profile

Consult Chapter 6 in *Intelligent Testing with the WISC-III,* (Kaufman, 1994) as it deals with the systematic reorganization of subtest profiles to generates hypotheses about strengths and weaknesses.

Note: VC = Verbal Comprehension; PO = Perceptual Organization; FD = Freedom from Distractibility; PS = Processing Speed.
Source: From *Intelligent Testing with the WISC-III* (Table 3.4), by A. S. Kaufman, 1994, New York: John Wiley & Sons. Reprinted with permission.

ing, which require good nondistractible attention spans for success. The Perceptual Organization Index measures nonverbal thinking and visual-motor coordination. More specifically, it assesses an individual's ability to integrate visual stimuli, reason nonverbally, and apply visual-spatial and visual-motor skills to solve the kinds of problems that are not school-taught. Finally, the Processing Speed Index measures response speed in solving an assortment of nonverbal problems (speed of thinking as well as motor speed) (Kaufman, 1994).

Within the WISC-III there are 10 mandatory and 3 supplementary subtests, all of which span the age range of 6 to 16 years. The Verbal Scale's five mandatory subtests include: Information, Similarities, Arithmetic, Vocabulary, and Comprehension. The supplementary subtest on the Verbal Scale is Digit Span. Digit Span is not calculated into the Verbal IQ unless it has been substituted for another Verbal subtest because one of those subtests has been spoiled (Kamphaus, 1993; Wechsler, 1991).

The five mandatory Performance Scale's subtests include Picture Completion, Picture Arrangement, Block Design, Object Assembly, and Coding. The two supplementary subtests on the Performance Scale are Mazes and Symbol Search. The Mazes subtest may be substituted for any Performance Scale subtest; however, Symbol Search may only be substituted for the Coding subtest (Kamphaus, 1993; Wechsler, 1991).

"Symbol Search is an excellent task that should have been included among the five regular Performance subtests instead of Coding. Mazes is an awful task that should have been dropped completely from the WISC-III" (Kaufman, 1994, p.58). He goes further to say that "there's no rational reason for the publisher to have rigidly clung to Coding as a regular part of the WISC-III when the new Symbol Search task is clearly a better choice for psychometric reasons" (Kaufman, 1994, p. 59). Therefore, for all general purposes, Kaufman (1994) strongly recommends that Symbol Search be routinely substituted for coding as part of the regular battery, and to use Symbol Search to compute the Performance IQ and Full Scale IQ. The manual does not tell one to do this, but neither does it prohibit it.

Reliability of each subtest except Coding and Symbol Search was estimated by the split-half method. Stability coefficients were used as reliability estimates for the Coding and Symbol Search subtests because of their speeded nature. Across the age groups, the average reliability coefficients are: Information (.84), Similarities (.81), Arithmetic (.78), Vocabulary (.87), Comprehension (.77), Digit Span (.85), Picture Completion (.77), Coding (.79), Picture Arrangement (.76), Block Design (.87), Object Assembly (.69), Symbol Search (.76), and Mazes (.70). The average reliability, across the age groups, for the IQs and Indexes are: .95 for the Verbal IQ, .91 for the Performance IQ, .96 for the Full Scale IQ, .94 for the Verbal Comprehension Index, .90 for the Perceptual Organization Index, .87 for the Freedom from Distractibility Index, and .85 for the Processing Speed Index (*WISC-III Interpretive Manual,* 1991).

Analyzing the WISC-III Data. To obtain the most information from the WISC-III, one should be more than familiar with each of the subtests individually as well as with the potential information that those subtests can provide when integrated or combined. The WISC-III is maximally useful when tasks are grouped and regrouped to uncover a child's strong and weak areas of functioning, so long as these hypothesized assets and deficits are verified by multiple sources of information.

As indicated previously, the WISC-III provides examiners with a set of four Factor Indexes in addition to the set of three IQs. The front page of the WISC-III record form lists the seven standard scores in a box on the top right. The record form is quite uniform and laid out nicely; however, it is difficult to know just what to do with all of those scores. Kaufman (1994) has developed The Seven Steps which offer a unique and systematic method of WISC-III interpretation that allows the clinician

to organize and integrate the test results in a stepwise, easy-to-use approach. The Seven Steps provide an empirical framework for profile attack while organizing the profile information into hierarchies. Table 4.1 provides an overview of the seven interpretive steps for WISC-III profiles.

Critique. Professionals in the field of intelligence testing have described the third edition of the Wechsler Intelligence Scale for Children in a number of different ways. "The WISC-III reports continuity, the status quo, and only the smallest step in the evolution of the assessment of intelligence. Despite more than 50 years of advancement in theories of intelligence, the Wechsler philosophy of intelligence (not actually a formal theory), written in 1939, remains the guiding principle of the WISC-III" (Shaw, Swerdlik, & Laurent, 1993, p. 151). One of the principal goals for developing the WISC-III stated in the manual was merely to update the norms, which is "hardly a revision at all" (Sternberg, 1993). If one has chosen to use the WISC-III because he or she is looking for a test of new constructs in intelligence, or merely a new test, one should look elsewhere (Sternberg, 1993). In contrast to these fairly negative evaluations, Kaufman (1994) reports that the WISC-III is a substantial revision of the WISC-R and that the changes that have been made are considerable and well done. "The normative sample is exemplary, and the entire psychometric approach to test development, validation, and interpretation reflects sophisticated, state-of-the-art knowledge and competence" (Kaufman, 1994). For Kaufman, the WISC-III is not without its flaws but his overall review of the test is quite positive. Although the WISC-III has clearly had mixed reviews, it is one of the most frequently used tests in the field of children's intelligence testing.

Kaufman Assessment Battery for Children (K-ABC)

The K-ABC is a battery of tests measuring intelligence and achievement of normal and exceptional children ages 2½ through 12½ years. It yields four scales: the Sequential Processing Scale, the Simultaneous Processing Scale, the Mental Processing Composite (Sequential and Simultaneous) Scale, and the Achievement Scale. The K-ABC is becoming a frequently used test in intelligence and achievement assessment that is used by both clinical and school psychologists (Kamphaus, Beres, Kaufman, & Kaufman, 1995). In a nationwide survey of school psychologists conducted in 1987 by Obringer (1988), respondents were asked to rank the following instruments in order of their usage: Wechsler's scales, the K-ABC, and both the old and new Stanford-Binets. The Wechsler scales earned a mean rank of 2.69, followed closely by the K-ABC with a mean of 2.55, the L-M version of the Binet 1.98 and the Stanford-Binet: Fourth Edition 1.26. Bracken (1985) also found similar results of the K-ABC's increasing popularity. Bracken surveyed school psychologists and found that for ages 5 to 11 years the WISC-R was endorsed by 82 percent, the K-ABC by 57 percent, and the Binet IV by 39 percent of the practitioners. These results suggest that clinicians working with children should have some familiarity with the K-ABC (Kamphaus et al., 1995).

The K-ABC has been the subject of great controversy from the outset, as evident in the strongly pro and con articles written for a special issue of the *Journal of Special Education* devoted to the K-ABC (Miller & Reynolds, 1984). Many of the controversies, especially those regarding the validity of the K-ABC theory, will likely endure unresolved for some time (Kamphaus et al., 1995). Fortunately, the apparent controversy linked to the K-ABC has resulted in numerous research studies and papers that provide more insight into the K-ABC and its strengths and weaknesses.

Theory. The K-ABC intelligence scales are based on a theoretical framework of sequential and simultaneous information-processing, which relates to *how* children solve problems rather than *what* type of problems they must solve (e.g., verbal or nonverbal), which is in stark contrast to Wechsler's theoretical framework of the assessment of "g", a conception of intelligence as an overall global entity. As a result, Wechsler used the Verbal and Performance scales as a means to an end. That end is the assessment of general intelligence. In comparison, the Kaufmans emphasize the individual importance of the Sequential and Simultaneous Scales in interpretation, rather than the overall Mental Processing Composite (MPC) score (Kamphaus et al., 1995).

The sequential and simultaneous framework for the K-ABC stems from an updated version of a

Table 4.2. Representation of the Standardization Sample by Educational Placement (N = 2,000)

EDUCATIONAL PLACEMENT	K-ABC STANDARDIZATION SAMPLE		
	N	%	%
Regular Classroom	1,862	93.1	91.1
Speech Impaired	28	1.4	2.0
Learning Disabled	23	1.2	2.3
Mentally Retarded	37	1.8	1.7
Emotionally Disturbed	5	0.2	0.3
Other [b]	15	0.8	0.7
Gifted and Talented	30	1.5	1.9[c]
Total K-ABC Sample	2,000	100.0	100.0

[a]Data from U.S. Department of Education, National Center for Education Statistics, 1980. Table 2.7, *The Condition of Education*, Washington, DC, U.S. Government Printing Office.

[b]Includes other health impaired, orthopedically handicapped, and hard of hearing.

[c]Data from U.S. Office for Civil Rights, 1980, *State, Regional, and National Summaries of Data from the 1978 Child Rights Survey of Elementary and Secondary Schools*, p.5, Alexandria, VA, Killalea Associates.

variety of theories (Kamphaus, 1993). The foundation lies in a wealth of research in clinical and experimental neuropsychology and cognitive psychology. The sequential and simultaneous theory was primarily developed from two lines of theory: the information-processing approach of Luria (e.g., Luria, 1966a), and the cerebral-specialization work of Sperry (1968, 1974), Bogen (1975), Kinsbourne (1978), and Wada, Clarke, and Hamm (1975).

In reference to the K-ABC, simultaneous processing refers to the mental ability to integrate information all at once to solve a problem correctly. Simultaneous processing frequently involves spatial, analogic, or organizational abilities (Kaufman & Kaufman, 1983c; Kamphaus & Reynolds, 1987). There is often a visual aspect to the problem and visual imagery used to solve it. A prototypical example of a simultaneous subtest is the Triangles subtest on the K-ABC, which is similar to Wechsler's Block Design. To solve both of these subtests, children must be able to see the whole picture in their mind and then integrate the individual pieces to create the whole.

In comparison, sequential processing emphasizes the ability to place or arrange stimuli in sequential or serial order. The stimuli are all linearly or temporally related to one another, creating a form of serial interdependence within the stimulus (Kaufman & Kaufman, 1983c). The K-ABC subtests assess the child's sequential processing abilities in a variety of modes. For example, Hand Movements involves visual input and a motor response, Number Recall involves auditory input with a verbal response, and Word Order involves auditory input and visual response. These different modes of input and output allow the examiner to assess the child's sequential abilities in a variety of ways. The sequential subtests also provide information on the child's short-term memory and attentional abilities.

According to Kamphaus et al. (1995), one of the controversial aspects of the K-ABC was the fact that it took the equivalent of Wechsler's Verbal Scale and redefined it as "achievement". The Kaufmans' analogs of tests such as Information (Faces & Places), Vocabulary (Riddles and Expressive Vocabulary), and Arithmetic (Arithmetic) are included on the K-ABC as achievement tests. The Kaufmans viewed the above tests as diverse tasks that are united by the demands they place on children to extract and assimilate information from their cultural and school environment. The K-ABC is predicated on the distinction between problem solving and knowledge of facts. The former set of skills is interpreted as intelligence; the latter is defined as achievement. This definition presents a break from other intelligence tests, where a person's acquired factual information and applied skills frequently influence greatly the obtained IQ (Kaufman & Kaufman, 1983c).

Standardization and Properties of the Scale. Stratification of the K-ABC standardization sample was excellent and closely matched the 1980 U. S. Census data in age, gender, geographic region, community size, socioeconomic status, race and ethnic group, and parental occupation and education. Additionally, unlike most other intelligence measures for children, stratification variables also

included educational placement of the child (see Table 4.2).

Reliability and validity data are impressive. A test-retest reliability study was conducted with 246 children after a 2- to 4-week interval (mean interval = 17 days). The coefficients for the Mental Processing Composite were .83 for ages 2 years, 6 months through 4 years, 11 months; .88 for ages 5 years, 0 months through 8 years, 11 months; and .93 for ages 9 years, 0 months to 12 years, 5 months. Test-retest reliabilities for the Achievement scale composite for the same age groups were .95, .95, and .97 respectively (Kamphaus et al., 1995). The test-retest reliability research reveals that there is a clear developmental trend, with coefficients for the preschool ages being smaller than those for the school-age range. This trend is consistent with the known variability over time that characterizes preschool children's standardized test performance in general (Kamphaus & Reynolds, 1987).

Split-half reliability coefficients for the K-ABC global scales range from 0.86 to 0.93 (mean = 0.90) for preschool children, and from 0.89 to 0.97 (mean = 0.93) for children aged 5 to 12 ½ (Kamphaus et al., 1995).

There has been a considerable amount of research done on the validity of the K-ABC. The *K-ABC Interpretive Manual* (Kaufman & Kaufman, 1983c) includes the results of 43 such studies. Construct validity was established by looking at five separate topics: developmental changes, internal consistency, factor analysis (principal factor, principal components, and confirmatory), convergent and discriminant analysis, and correlations with other tests. Factor analysis of the Mental Processing Scales offered clear empirical support for the existence of two, and only two, factors at each age level, and for the placement of each preschool and school-age subtest on its respective scale. Analyses of the combined processing and achievement subtests also offered good construct validation of the K-ABC's three-scale structure (Kaufman & Kamphaus, 1984).

Although the K-ABC and the WISC-III differ from one another in a number of ways, there is strong evidence that the two measures correlate substantially (Kamphaus & Reynolds, 1987). In a study of 182 children enrolled in regular classrooms, the Mental Processing Composite (MPC) correlated 0.70 with WISC-R Full-Scale IQ (FSIQ), thus, sharing a 49 percent overlap in variance (Kamphaus et al., 1995; Kaufman & Kaufman, 1983c). There have also been numerous correlational studies conducted with handicapped and exceptional populations that may be found in the *Interpretative Manual.*

Critique. Although the K-ABC has been the subject of past controversy, it appears that it has held its own and is used often by professionals. The K-ABC is well designed with easels and manuals that are easy to use. The information in the manuals is presented in a straightforward, clear fashion, making use and interpretation of the K-ABC relatively easy (Merz, 1985). There has been a considerable amount of research done on the validity of the K-ABC and the authors have done a thorough job of presenting much of that information in the manual. The reporting of the reliability and validity data in the manual is complete and understandable. However, there is not enough information presented on the content validity of the test. The various tasks on the subtests on the K-ABC are based on clinical, neuropsychological or other research-based validity; however, a much clearer explication of the rationale behind some of the novel subtests would have been quite helpful (Merz, 1985).

The K-ABC measures intelligence from a strong theoretical and research basis, evident in the quality of investigation in the amount of research data presented in the manual (Merz, 1985). The K-ABC was designed to measure the intelligence and achievement of children 2 ½ to 12 ½ years old and the research done to date suggests that in fact the test does just that. The Nonverbal Scale significantly contributes to the effort to address the diverse needs of minority groups and language-handicapped children. Overall, it appears that the authors of the K-ABC have met the goals listed in the interpretative manual and that this battery is a valuable assessment tool (Merz, 1985).

In a number of studies, Keith and his colleagues (Keith, 1985; Keith & Dunbar, 1984) have called the K-ABC processing model into question by applying Wechsler-like content labels to the K-ABC scales. Keith (1985) used labels such as "nonverbal/reasoning" (Simultaneous), "achievement/verbal reasoning" (Achievement), and "verbal memory" (Sequential) for the K-ABC factors, making the scales similar to the tradition of psychological assessment. "The issue of what to call the K-ABC factors remains debated but unresolved" (Kamphaus, 1993).

Table 4.3. Representation of the Stanford-Binet, Fourth Edition

	SAMPLE PERCENT	U.S. POPULATION PERCENT
By Parental Occupation		
Managerial/Professional	45.9	21.8
Technical Sales	26.2	29.7
Service Occupations	9.7	13.1
Farming/Forestry	3.2	2.9
Precision Production	6.7	13.0
Operators, Fabricators, Other	8.3	19.5
Total	100.0	100.0
By Parental Education		
College Graduate or Beyond	43.7	19.0
1 to 3 Years of College	18.2	15.3
High School Graduate	27.5	36.5
Less Than High School Graduate	10.6	29.2
Total	100.0	100.0

Stanford-Binet: Fourth Edition (SB-IV)

Theory. Like its predecessor, the Fourth Edition (SB-IV) is based on the principal of a general ability factor, "g," rather than on a connection of separate functions. The Fourth Edition has maintained, yet to a much lesser degree, its adaptive testing-format. No examinee takes all the items on the scale, nor do all examinees of the same chronological age respond to the same tasks. Like its predecessor, the scale provides a continuous appraisal of cognitive development from ages two through adult.

One of the criticisms of the previous version is that it tended to underestimate the intelligence of examinees whose strongest abilities did not lie in verbal skills (or overestimate the intelligences of those whose verbal skills excelled). Therefore, consideration when developing the SB-IV was to give equal credence to several areas of cognitive functioning. The authors set out to appraise verbal reasoning, quantitative reasoning, abstract/visual reasoning, and short-term memory (in addition to a composite score representing "g").

This model is based on a three-level hierarchical model of the structure of cognitive abilities. A general reasoning factor is at the top level ("g"). The next level consists of three broad factors: crystallized abilities, fluid analytic abilities, and short-term memory. The third level consists of more specific factors: verbal reasoning, quantitative reasoning, and abstract/visual reasoning.

The selection of these four areas of cognitive abilities came from the authors' research and clinical experience of the kinds of cognitive abilities that correlate with school progress. This foundational emphasis on academic cognition continues the philosophy of the original Binet, which did not extend to measuring adult intelligence as did later versions, including the SB-IV. One wonders whether the same emphasis should be used when measuring adult intelligence. While subtests change (with considerable overlap) for various age groups and while selection reportedly has been subjected to rigorous research, there is considerable dispute whether children and adults utilize the same intellectual processes. After all, any task can be developed and normed for a variety of ages, but does that mean that each age group is calling upon the same processes to accomplish this task?

The SB-IV contains previous tasks combining old with new items and some completely new tasks. In general, test items were accepted if (a) they proved to be acceptable measurements of the construct; (b) they could be reliably administered and scored; (c) they were relatively free of ethnic or gender bias; and (d) they functioned adequately over a wide range of age groups (again, not making philosophical distinctions between intelligence of children and adults).

Standardization and Properties of the Scale. Standardization procedures followed 1980 U.S. Census data. There appears to be an accurate sample representation from geographic region, size of community, ethnic group, and gender. The standardization falls short, however, in terms of age, parental occupation, and parental education. The age representation extends from 2 years of age, 0 months to 23 years, 11 months. The concentration of the sample is on children 4 to 9 years old (41%). Not only were adults 24 years and older not represented, but also representation beyond age 17 years, 11 months was negligible (4%).

In order to assess characteristics of socioeconomic status (SES), information regarding parental occupation and parental education was obtained. A review of Table 4.3 demonstrates that children whose parents came from managerial or professional occupations and/or who were college graduates and beyond were grossly overrepresented in the sample. In other words, the norms are based on a large percentage of individuals from upper-socioeconomic classes. In order to adjust for this discrepancy, an after-the-fact weighting procedure was applied, which makes the norming sample suspect. Unquestionably, SES has been shown time and again to be the single most important stratification variable regarding its relationship to IQ (Kaufman, 1990a, Chapter 6; Kaufman & Doppelt, 1976).

According to McCallum (1990), there is a considerable amount of evidence for the general construct validity of the SB-IV. For example, the difficulty level of items within the various subtests is developmentally determined. In other words, age and cognitive maturity are highly correlated with success on items. Therefore, older children are more likely to succeed on the items than younger children. Additionally, the SB-IV measures intelligence in ways that are similar to older, established tests of intelligence. Correlation coefficients between the SB-IV global scores from the Wechsler scales, the Stanford-Binet (Form L-M), and the K-ABC range from .50 to .85 (McCallum, 1990).

Research also shows that the individual subtests of the SB-IV had impressive high to substantial loadings on "g" (.51-.79). Unfortunately, the four factors were given weak support by the confirmatory procedure. Additionally, exploratory factor analysis gave even less justification for the four Binet Scales; only one or two factors were identified by Reynolds, Kamphaus, and Rosenthal (1988) for 16 of the 17 age groups studied. Clearly, the factor analytic structure does not conform to the theoretical framework used to construct the test. Therefore, once again one is left with the composite score as the only clearly valid representation of a child's cognitive abilities.

Correlational studies, using non-exceptional children, between the SB-IV and the Stanford-Binet (Form L-M), WISC-R, Wechsler Adult Intelligence Scale-Revised (WAIS-R), Wechsler Preschool and Primary Scale of Intelligence (WPPSI), K-ABC have ranged from .80 to .91 (comparing full-scale composites). Correlational studies using exceptional children (gifted, learning impaired, mentally retarded) produced generally lower correlations, probably because of restricted variability in the test scores. These data and data from similar validity investigations are presented more extensively in the *Technical Manual* for the SB-IV (Thorndike, Hagen, & Sattler, 1986). Hodapp (1993) conducted a correlational study between the SB-IV and the PPVT-R with a group of 42 children ranging in age from 3 to 6 years. Correlations of .54, .60, and .50 were computed for Standard Age Scores on the SB-IV Composite, Vocabulary, and Absurdities with the PPVT-R standard score equivalent. The seven other SB-IV subtests showed correlations ranging from .25 to .38.

There appears to be a considerable amount of diversity in the conclusions drawn from the research on the validity and usefulness of the SB-IV. However, in general, the evaluations of the SB-IV tend to be rather negative, suggesting that its use in the field may be limited. The irresponsibly gathered normative data, and other difficulties with the SB-IV have led at least one reviewer to recommend that the battery be laid to rest (Reynolds, 1987); "To the SB-IV, *Requiescat in pace*" (p. 141).

Critique. The SB-IV was developed in an attempt to increase the popularity of the test as well as address some of the negative reviews that had plagued the previous edition. The test authors attempted to make the Fourth Edition significantly different from the previous L-M Edition; however, it does not appear that they succeeded in doing so. Canter (1990) describes the "rebirth" of the Stanford-Binet as giving way to "confusion and even dismay as the primary consumers of intelligence tests learned that the new edition offered a more complicated route to the same destination." Another reviewer describes the SB-IV as "in most

respects, a completely new version of a very old test" (Spruill, 1987). It appears that the Fourth Edition of the Stanford-Binet has been a disappointment for most professionals in the field of intelligence testing.

One of the major problems with the Fourth Edition is the fact that it went into publication too soon. As a result, the test was published without accompanying technical data to allow the user to evaluate the appropriateness and technical adequacy of the instrument. This made it difficult for the examiner to know if the test was appropriate for his or her client, not to mention that it is a violation of Standard 5.1 in the *Standards for Educational and Psychological Testing* (Spruill, 1987). Furthermore, there were errors in the norms tables in the first printing of the administration manual.

A common criticism of the SB-IV and previous versions is that there had been inadequate standardization. For example, in the Fourth Edition the standardization sample contained a larger percentage of high-socioeconomic-status subjects than in the population at large, as demonstrated previously. It is not clear whether or not the weighting procedure that was used to correct for sample bias was adequate (Spruill, 1987). Also, the test was designed to be used with individuals from age 2 to "adult"; however, there are no normative data for "adults" over the age of 23. This may also be misleading to an examiner.

Although there appears to be a number of flaws with the SB-IV, the test is still used and it is not without its strengths. The administration of some of the subtests allow the examiner a little flexibility, and young children seem to find the items challenging and fun. Despite its shortcomings, SB-IV continues to be a very good assessment of cognitive skills related to academic progress (Spruill, 1987). It also includes several excellent, well-constructed tasks that offer valuable supplementary information when they are administered as Weschler supplements (Kaufman, 1990a, 1994).

Kaufman Adolescent and Adult Intelligence Test (KAIT)

The KAIT (Kaufman & Kaufman, 1993), is an individually administered intelligence test for individuals between the ages of 11 and more than 85 years. It provides Fluid, Crystallized, and Composite IQs, each a standard score with a mean of 100 and a standard deviation of 15.

Theory. The Horn-Cattell theory forms the foundation of the KAIT and defines the constructs presumed to be measured by the separate IQs; however, other theories guided the test development process, specifically the construction of the subtests. Tasks were developed from the models of Piaget's formal operations (Inhelder & Piaget, 1958; Piaget, 1972) and Luria's (1973, 1980) planning ability in an attempt to include high-level decision making on more developmentally advanced tasks. Luria's notion of planning ability involves decision making, evaluation of hypotheses, and flexibility, and "represents the highest levels of development of the mammalian brain" (Golden, 1981, p.285).

Piaget's formal operations depicts a hypothetical-deductive abstract reasoning system that has as its featured capabilities the generation and evaluation of hypotheses and the testing of propositions. The prefrontal areas of the brain associated with planning ability mature at about ages 11 to 12 years (Golden, 1981), the same ages that characterize the onset of formal operational thought (Piaget, 1972). The convergence of the Luria and Piaget theories regarding the ability to deal with abstractions is striking; this convergence provided the rationale for having age 11 as the lower bound of the KAIT, and for attempting to measure decision making and abstract thinking with virtually every task on the KAIT (Kaufman & Kaufman, 1993).

Within the KAIT framework (Kaufman & Kaufman, 1993), Crystallized intelligence "measures the acquisition of facts on problem solving ability using stimuli that are dependent on formal schooling, cultural experiences, and verbal conceptual development" (p.7). Fluid intelligence "measures a person's adaptability and flexibility when faced with new problems, using both verbal and nonverbal stimuli" (Kaufman & Kaufman, 1993, p. 7). It is important to note that this Crystallized-Fluid construct split is not the same as Wechsler's (1974, 1981, 1991) verbal-nonverbal split. More specifically, the KAIT Fluid subtests stress reasoning rather than visual-spatial ability, include verbal comprehension or expression as key aspects of some tasks, and minimize the role played by visual-motor speed for correct responding.

The Core Battery of the KAIT is composed of three Crystallized and three Fluid subtests, and these six subtests are used to compute the IQs. The Expanded Battery also includes two supplementary subtests and two measures of delayed recall that evaluate the individual's ability to retain infor-

Table 4.4. Correlations of the Three KAIT IQ with Standard Scores and IQs yielded by Other Major Intelligence Tests

INTELLIGENCE TEST	AGE RANGE	CRYSTALLIZED	FLUID	COMPOSITE
WAIS-R Verbal IQ (*N*=343)	16–83	0.78	0.62	0.76
WAIS-R Performance IQ (*N*=343)	16–83	0.72	0.72	0.77
WAIS-R Full Scale IQ (*N*=343)	16–83	0.86	0.73	0.85
WISC-R Verbal IQ (*N*=118)	11–16	0.79	0.74	0.83
WISC-R Performance IQ (*N*=118)	11–16	0.67	0.67	0.72
WISC-R Full Scale IQ (*N*=118)	11–16	0.78	0.75	0.82
K-ABC Mental Processing Composite (*N*=124)	11–12	0.57	0.62	0.66
K-ABC Achievement (*N*=124)	11–12	0.81	0.64	0.82
SB-V Test Composite (*N*=79)	11–42	0.81	0.84	0.87

Note: Data in this table are based on data reported in the KAIT Manual (Kaufman & Kaufman, 1993, Tables 8.15–8.19 and 8.22–8.23. Data for the WAIS-R are averages of values reported separately for four age groups between (a) 16 and 19 years and (b) 50 and 83 years.

mation that was learned previously in the evaluation during two of the Core subtests.

The Core Battery of the KAIT consists of subtests one through six, and subtests one through ten comprise the Expanded Battery. Each subtest except the supplementary Mental Status task yields age-based scaled scores with a mean of ten and a standard deviation of three. Sample and teaching items are included for most subtests to ensure that examinees understand what is expected of them for each subtest.

The delayed-recall subtests are administered, without prior warning, about 25 and 45 minutes after the administration of the original, related subtests. The two-delayed recall subtests provide good measure of an ability that Horn (1985, 1989) calls TSR (Long-Term Storage and Retrieval). TSR "involves the storage of information and the fluency of retrieving it later through association" (Woodcock, 1990, p. 234).

The Mental Status subtest is comprised of ten simple questions that assess attention and orientation to the world. Most normal adolescents and adults pass at least nine of the ten items, but the task has special use with retarded and neurologically impaired populations. The Mental Status subtest may be used as a screener to determine if the KAIT can be validly administered to an individual.

Standardization and Properties of the Scale. The KAIT normative sample, composed of 2,000 adolescents and adults between the ages of 11 and 94 years, was stratified on the variables of gender, racial/ethnic group, geographic region, and socioeconomic status (Kaufman & Kaufman, 1993).

Mean split-half reliability coefficients for the total normative sample were .95 for Crystallized IQ, .95 for Fluid IQ, and .97 for Composite IQ (Kaufman & Kaufman, 1993). Mean test-retest reliability coefficients, based on 153 identified normal individuals in three age groups (11–19 years of age, 20–54 years of age, 55–85+ years of age), retested after a one-month interval, were .94 for Crystallized IQ, .87 for Fluid IQ, and .94 for Composite IQ (Kaufman & Kaufman, 1993). Mean subtest split-half reliabilities of the four Crystallized subtests ranged from .89 to .92 (median = .90). Mean values for the four Fluid subtests ranged from .79 to .93 (median = .88) (Kaufman & Kaufman, 1993). Median test-retest reliabilities for the eight subtests, based on the 153 people indicated previously, ranged form .72 to .95 (median = .78). Rebus Delayed Recall had an average split-half reliability of .91 and Auditory Delayed Recall had an average value of .71; their respective stability coefficients were .80 and .63 (Kaufman & Kaufman, 1993).

Factor analysis, both exploratory and confirmatory, gave strong construct validity support for the Fluid and Crystallized Scales, and for the placement of each subtest on its designated scale. Crystallized IQs correlated .72 with Fluid IQs for the total standardization sample of 2,000 (Kaufman & Kaufman, 1993). Table 4.4 provides the correlations of the three KAIT IQs with standard scores and IQs yielded by other major intelligence tests. The data found in this table are taken from the KAIT Technical Manual (1993). The values shown in Table 4.4 support the construct and criterion-related validity of the three KAIT IQs.

The KAIT benefits from an integration of theories that unite developmental (Piaget), neuropsychological (Luria), and experimental-cognitive (Horn-Cattell) models of intellectual functioning. The theories work well together and do not compete with one another. Together, the theories give the KAIT a solid theoretical foundation that facilitate test interpretation across the broad 11 to 94 year age-range on which the battery was normed.

The KAIT and WISC-R were administered to 118 individuals ages 11 to 16 years, and the KAIT and WAIS-R were administered to 338 individuals ages 16 to 83 years; these data were factor analyzed in two separate joint analyses. A number of analyses were conducted to determine what factors each of the tests have that are unique and what factors they share. "The most crucial finding from these analyses is that the Wechsler Performance subtests and the KAIT Fluid subtests seem to measure markedly different constructs" (Kaufman & Kaufman, 1993, p. 93). According to Horn, there are important differences between Performance IQ and fluid intelligence, noting that Performance IQ "involves visualization to a very considerable extent" (Horn & Hofer, 1992, p. 72). The following conclusions from the joint factor analyses of KAIT and Wechsler subtests were drawn:

1. Three factors define the joint matrices of the KAIT and the Wechsler scales: Crystallized/Verbal, Fluid, and Perceptual Organization.
2. The constructs underlying the KAIT Fluid and the Wechsler Performance Scales are distinctly different. The Fluid and Perceptual Organization factors correlate about as highly with each other as they do with the Crystallized/Verbal factor.
3. The constructs underlying the KAIT Crystallized the Wechsler Verbal scales seem virtually identical; all component subtests load substantially on the Crystallized/Verbal factor.
4. The KAIT Crystallized and Fluid subtests load consistently on the factors underlying their respective scales. The Wechsler subtests, however, sometimes do not load highly on the factor underlying the scale to which they belong (Kaufman & Kaufman, 1993).

Critique. The KAIT represents a reconceptualization of the measurement of intelligence that is more consistent with current theories of intellectual development (Brown, 1994). The fluid-crystallized dichotomy, the theory underlying the KAIT, is based on the original Horn-Cattell theory of intelligence, thus offering a firm and well-researched theoretical framework (Flanagan, Alfonso, & Flanagan, 1994). The fluid-crystallized dichotomy enhances the richness of the clinical interpretations that can be drawn from this instrument (Brown, 1994). The test materials are well constructed and attractive, and the manual is well organized and helpful (Dumont & Hagberg, 1994; Flanagan et al., 1994). Furthermore, the test materials are easy to use and stimulating to examinees (Flanagan et al., 1994).

"The KAIT has been standardized by state-of-the-art measurement techniques" (Brown, 1994). The psychometric properties of the KAIT regarding standardization and reliability are excellent and the construct validity evidence that is reported in the manual provides a good foundation for its theoretical underpinnings (Flanagan et al., 1994).

The theoretical assumption that formal operations are reached by early adolescence limits the application of the KAIT with certain adolescent and adult populations (Brown, 1994). If an individual has not achieved formal operations, many of the subtests will be too difficult for them and perhaps frustrating and overwhelming. Examiners should be aware of this when working with such individuals in order to maintain rapport. The KAIT can be a great assessment tool when working with high-functioning, intelligent individuals; however, it can be difficult to use with borderline individuals and some elderly clients. Elderly clients' scores on some of the subtests may be negatively impacted by poor reading, poor hearing, and poor memory (Dumont & Hagberg, 1994).

Flanagan and colleagues (1994) report that the inclusion of only three subtests per scale may limit or interfere with the calculation of IQs if a subtest

is spoiled. The usefulness of the Expanded Battery and Mental Status subtest of clinical populations is questionable given the reliability and validity data presented in the manual, suggesting that interpretations be made with caution (Flanagan et al., 1994).

Although there clearly are some limitations in the use of the KAIT with some populations, overall, the test appears to be well thought out and validated (Dumont & Hagberg, 1994). The KAIT represents an advancement in the field of intellectual assessment with its ability to measure fluid and crystallized intelligence from a theoretical perspective and, at the same time, maintain a solid psychometric quality (Flanagan et al., 1994).

Woodcock-Johnson Psycho-Educational Battery—Revised: Tests of Cognitive Ability (WJ-R)

The WJ-R is one of the most comprehensive test batteries available for the clinical assessment of children and adolescents (Kamphaus, 1993). The WJ-R is a battery of tests for individuals from 2 to 90+ years of age, and is composed of two sections, Cognitive and Achievement. The focus of this discussion is the Cognitive portion of the WJ-R battery.

Theory. The WJ-R Cognitive battery is based on Horn's (1985, 1989) expansion of the Fluid/Crystallized model of intelligence (Kamphaus, 1993; Kaufman, 1990). The standard and supplemental subtests of the WJ-R are aligned with eight of the cognitive abilities isolated by Horn (1985, 1989) (Kamphaus, 1993; Kaufman 1990). These abilities include: Long-Term Retrieval, Short-Term Memory, Processing Speed, Auditory Processing, Visual Processing, Comprehension-Knowledge and Fluid Reasoning. An eighth ability, Quantitative Ability, is measured by several Achievement subtests on the WJ-R.

The four subtests that measure Long-Term Retrieval (Memory for Names, Visual-Auditory Learning, Delayed Recall/Memory for Names, Delayed Recall/Visual-Auditory Learning), require the subject to retrieve information stored minutes or a couple of days earlier. In contrast, the subtests that measure Short-Term Memory (Memory for Sentences, Memory for Words, Numbers Reversed) require the subject to store information and retrieve it immediately or within a few seconds. The two Processing Speed subtests (Visual Matching, Cross Out) assess the subject's ability to work quickly, particularly under pressure, to maintain focused attention.

Within the Auditory-Processing domain, three subtests (Incomplete Words, Sound Blending, Sound Patterns) assess the subject's ability to fluently perceive patterns among auditory stimuli. The three Visual-Processing subtests (Visual Closure, Picture Recognition, Spatial Relations) assess the subject's ability to fluently manipulate stimuli that are within the visual domain.

Picture Vocabulary, Oral Vocabulary, Listening Comprehension, and Verbal Analogies are the four subtests that are linked to the Comprehension-Knowledge factor, also known as crystallized intelligence within Horn's theoretical model. These subtests require the subject to demonstrate the breadth and depth of his or her knowledge of a culture. Analysis-Synthesis, Concept Formation, Spatial Relations, and Verbal Analogies (which also loads on the Comprehension-Knowledge factor) assess the subject's Fluid Reasoning. Finally, from the Achievement portion of the WJ-R, both the Calculation and Applied Problems subtests assess the individual's Quantitative Ability.

The cognitive battery consists of 21 subtests, 7 of which comprise the standard battery; the remaining 14 are part of the supplemental battery. There are two composite scores, Broad Cognitive Ability and Early Development (for preschoolers), which are both comparable to an overall IQ score. The individual subtest scores, as well as the composite scores, have a mean of 100 and a standard deviation of 15.

Computer software is available for scoring the WJ-R and is essential if one is to obtain all of the information that the WJ-R is capable of providing. The WJ-R provides the examiner with percentile ranks, grade-based scores, age-based scores and the Relative Mastery Index (RMI). The RMI is unique and similar to a ratio with the second part of the ratio set at a value of 90. The denominator of the ratio means that children in the norm sample can perform the intellectual task with 90 percent accuracy. The numerator of the ratio refers to that child or adolescent's proficiency on that subtest (Kamphaus, 1993). For example, if a child obtains an RMI of 60/90, it would mean that the child's proficiency on the subtest is at a 60 percent level whereas the typical child of his or her age (or grade) mastered the material at a 90 percent level of accuracy.

The entire battery is quite lengthy and therefore can be timely to administer. The Standard Battery takes approximately 40 minutes to administer; however, all the clinician will obtain from it is, essentially, a measure of "g". In order to obtain all of the information that the WJ-R is capable of providing, a clinician should administer most of the subtests in both the Cognitive and Achievement batteries. Administration of a thorough cognitive and achievement assessment using the WJ-R would take approximately 3½ to 5 hours depending on the subject's age, abilities, and speed. However, individual subtests may be administered to test specific hypothesis without administering the entire battery.

Standardization and Properties of the Scale. The WJ-R was normed on a representative sample of 6,359 individuals selected to provide a cross-section of the U.S. population from 2 to 90+ years of age (Woodcock & Mather, 1989). The sample included 705 preschool children, 3,245 students in grades K through 12, 916 college or university students, and 1,493 individuals aged 14 to 90+ years who were not enrolled in school. Stratification variables included gender, geographic region, community size, and race. However, Kaufman (1990a) reports that although representation on important background variables was adequate, it was not excellent and therefore necessitated the use of a weighting procedure.

The internal consistency estimates for the standard battery are relatively high. The median coefficients are above .80 for five of the seven subtests. The Broad Cognitive Ability composite score based on seven standard battery subtests yields a median internal consistency coefficient of .94, and the Broad Cognitive Ability Early Development scale yields a coefficient of .96 at ages 2 and 4 years (Kamphaus, 1993).

The *Woodcock-Johnson Psycho-Educational Battery-Revised: Examiner's Manual* reports that "Items included in the various tests were selected using item validity studies as well as expert opinion" (Woodcock & Mather, 1989, p.7). Kamphaus (1993) states that the manual should have included more information on the results of the experts' judgements or some information on the methods and results of the studies that were used to assess validity.

It is clear that the WJ-R Cognitive battery is quite comprehensive, providing the clinician with a wealth of information. The standardization sample is large, the factor loadings reveal generally strong factor-analytic support for the construct validity for the battery for adolescents and adults, and the reliability coefficients are excellent (Kaufman, 1990a).

Critique. The WJ-R Cognitive battery was developed based on Horn's expansion of the Cattell-Horn Fluid-Crystallized model of intelligence. This theoretical rationale allows for further empirical analysis of both the WJ-R and the theory (Webster, 1994). The standardization of the battery appears to be sound and the various age groups are adequately represented. According to Webster (1994), the Cognitive battery is quite thorough, and when administered in its entirety, can provide the examiner with a wealth of information about an individual's intellectual functioning and abilities. The test materials and manuals are easy to use and well designed. The administration is fairly simple; however, scoring the test, especially when the Achievement battery is administered as well, can be quite a lengthy process. The scoring can be done by hand but is done more efficiently with the computer-scoring program. The computer-scoring program is easy to use and provides the examiner with the individual's raw scores, standard scores, percentile ranks, and age and grade equivalents for each subtest (Webster, 1994).

Kaufman (1986) reviewed the 1977 version of the Woodcock-Johnson (WJ) battery and concluded that it "is a mixture of extremes, possessing some outstanding qualities, yet hampered by glaring liabilities." He went further to add that the WJ represents a monumental and creative effort by its authors and he encourages examiners to take the time to master the test. Cummings (1985) agreed that the WJ is a "significant addition" to the available psychometric instruments. According to Kaufman (1990a), these comments apply as well to the WJ-R, although he expressed concern about interpreting many scales, each composed of few subtests. The WJ-R Cognitive battery is a well-standardized test developed on an interesting theory of intelligence. However, the test is not without shortcomings. Webster (1994) raises issues with the specific psychometric procedures used in developing test items. Data are lacking that show the efficacy of the WJ-R to predict, from a time-based perspective, actual functional levels of academic achievement and identify children at-risk-for-failure early in the educational process (Webster, 1994).

Other General Cognitive Measures

In addition to the major intelligence tests previously discussed, there are a number of other cognitive measures that are frequently used to assess the intelligence of both children and adolescents. These measures were developed based on a number of different theories and each of them offers a unique way of assessing the individual. This section of the text will provide general information on five cognitive tests for children: Peabody Picture Vocabulary Test-Revised (PPVT-R), Wechsler Preschool and Primary Scale of Intelligence-Revised (WPPSI-R), Detroit Tests of Learning Aptitude (DTLA-3), Matrix Analogies Test, and Differential Abilities Scales (DAS). The tests that were chosen for this section are by no means exhaustive. In fact, there are a number of tests that have not been discussed. For example, the Kaufman Brief Intelligence Test (K-BIT) is integrated in the case report at the end of this chapter; however, it was not presented in the group of cognitive tests that were chosen to be discussed in this chapter.

Peabody Picture Vocabulary Test-Revised (PPVT-R)

This brief test provides an approximate estimate of intelligence by measuring receptive vocabulary and replaces the original Peabody Picture Vocabulary Test (PPVT) published in 1959 (Dunn & Dunn, 1981). The 1981 version retains many of its predecessor's best features: it consists of two equivalent forms, allows for a verbal or nonverbal response, and is untimed. The examinee is shown plates with four pictures on each and is to point to the picture that best illustrates the meaning of the stimulus word spoken by the examiner. The PPVT-R is appropriate for individuals aged 2½ years through adult who can hear the stimulus word, see the drawings, and respond in some manner.

While the original PPVT was normed on a large but restricted sample, the PPVT-R norms were based on a nationwide data-gathering effort which, for children, was representative of the 1970 U.S. Census data with regard to sex, age, geographic region, occupational background, race and ethnic background, and urban-rural distributions. Because only 828 adults (ages 19 through 40 years) in contrast to 4,200 children, were included in the standardization, the manual suggests careful interpretation of scores for individuals above 18 years, 11 months old. Minority groups were included in the normative sample and are also included on the test plates. Sex- and ethnic-stereotyping, a problem with the original PPVT, has been virtually eliminated. The pictorial stimuli were redrawn to reflect a more appropriate racial, ethnic, and gender representation. Following the trend of other new or revised tests, the PPVT-R adopted conversion of raw scores to either percentile ranks, age equivalents, or standard score equivalents (mean = 100; standard deviation = 15).

The test manual reports moderate internal consistency (.61 to .88) and alternate form reliability estimates (.71 to .91) for the standardization sample. The degree of equivalence of the two forms was established for a subsample of 642 children. Coefficients of equivalence ranged from .73 to .91 (median = .82). Correlations of the PPVT-R with other intelligence composites typically range from .40 to .60 (Dunn & Dunn, 1986; Kaufman & Kaufman, 1983c; McCallum, 1985). These modest concurrent validity estimates suggest limited shared variance. Therefore, the PPVT-R should not be interpreted as equivalent to intelligence test scores.

Critique. As a test of hearing vocabulary, the PPVT-R is one of the most widely used instruments of its kind (Umberger, 1985). The PPVT-R is an easy-to-use test of receptive language, providing content that is current and that contains appropriate racial, ethnic, and gender representation. The national representative standardization for the educationally critical age range (2½ to 19 years) responds to requirements set by P.L. 94-242 (Wiig, 1985) and psychometric characteristics of this latest revision appear adequate to excellent (McCallum, 1985). It allows for flexibility in administration which lends itself to applicability to a number of exceptional populations (Umberger, 1985).

As a brief test, the PPVT-R is reliable, but it is not as reliable as one would expect for a test composed of 175 items. The reliability is greatly hindered by the element of chance that the test has within its design. Each item on the PPVT-R is a four-option multiple-choice question, meaning that guessers will be correct one out of four times. Another shortcoming of the PPVT-R is that its norms stop at 40 and that the adult norms are inferior to its superb norms for children and adolescents (Kaufman, 1990). Kaufman (1990a) also

cites a number of strengths of the PPVT-R that should be mentioned. For example, the PPVT-R's reliability and validity data for adolescents and adults are generally good, and the revisions were done thoroughly and with extreme care.

The test materials are well designed, making the test easy to administer and score. The test can be administered and scored quickly, allowing quick interpretation. Facilitation of interpretation has also been improved by providing the ability to convert raw scores to percentile ranks, age equivalents, and standard-score equivalents. In addition to being a useful assessment tool, the PPVT-R has a wide application as a research tool or as one test in a battery of tests on language competence (Umberger, 1985).

Wechsler Preschool and Primary Scale of Intelligence—Revised (WPPSI-R)

The WPPSI-R is an intelligence test for children aged 3 years, 0 months through 7 years, 3 months. The original version of the WPPSI was developed in 1967 for ages 4 to 6 ½ years, and the WPPSI-R was revised in 1989. Several changes were made to the revised version of the WPPSI-R. The norms were updated, the appeal of the content to young children was improved, and the age range was expanded.

The WPPSI-R is based on the same Wechsler-Bellevue theory of intelligence, emphasizing intelligence as a global capacity but having Verbal and Performance scales as two methods of assessing this global capacity (Kamphaus, 1993). The Verbal scale subtests include: Information, Comprehension, Arithmetic, Vocabulary, Similarities, and Sentences. The Performance scale subtests include: Object Assembly, Block Design, Mazes, Picture Completion, and Animal Pegs. Both the Sentences subtest and the Animal-Pegs subtest are supplemental tests and may be used in place of other subtests when deemed necessary.

Like the K-ABC and the Differential Abilities Scales (DAS), the WPPSI-R allows the examiner to "help" or "teach" the client on early items on the subtests to assure that the child understands what is expected of him or her. Providing this extra help is essential when working with reticent preschoolers (Kamphaus, 1993). Subtest scores have a mean of 10 and a standard deviation of 3. The overall Verbal, Performance, and Full Scale IQs have a mean of 100 and a standard deviation of 15. The examiner manual also provides interpretive tables that allow the examiner to determine individual strengths and weaknesses as well as the statistical significance and clinical rarity of Verbal and Performance score differences.

The WPPSI-R was standardized on 1,700 children from ages three years through seven years, 3 months. The standardization procedures followed the 1986 U.S. Census Bureau estimates. Stratification variables included gender, race, geographic region, parental occupation, and parental education.

The WPPSI-R appears to be a highly reliable measure. The internal consistency coefficients across age groups, for the Verbal, Performance, and Full Scale IQs, are .95, .92, and .96 respectively. The reliability coefficients for the individual subtests vary considerably, from an average internal consistency coefficient of .86 for Similarities to an average of .63 for Object Assembly. With a group of 175 children from the standardization sample, a test-retest investigation was conducted. The investigation yielded coefficients in the high .80s and low .90s. The test-retest coefficient for the Full Scale IQ is .91 (Kamphaus, 1993).

The WPPSI-R manual provides some information on validity; however, it provides no information on the predictive validity of the test. Various studies have shown that concurrent validity between the WPPSI-R and other tests is adequate. The correlation between the WPPSI and the WPPSI-R Full Scale IQs was reported at .87, and the correlation between WPPSI-R and WISC-III Performance, Verbal, and Full Scale IQs for a sample of 188 children was .73, .85, and .85 respectively. The correlations between he WPPSI-R and other well known cognitive measures is, on average, much lower. The WPPSI-R Full Scale IQ correlated .55 with the K-ABC Mental Processing Composite (Kamphaus, 1993) and .77 with the SB-IV. In general, the validity coefficients provide strong evidence for the construct validity of the WPPSI-R (Kamphaus, 1993).

Critique. The WPPSI-R is a thorough revision of the 1967 WPPSI and is for an expanded age range. It has new colorful materials, and item-types for very young children, as well as a new icebreaker subtest (Object Assembly) and a comprehensive manual (Kaufman, 1990a). The revision of the test has resulted in an instrument that is more attractive, and engaging, and has materials that are easier

to use (Buckhalt, 1991; Delugach, 1991). The normative sample is large, provides recent norms, and is representative of the 1986 U.S. Census data (Delugach, 1991; Kaufman, 1990a). The split-half reliability of the IQs and most subtests are exceptional, the factor analytic results for all age groups are excellent, and the concurrent validity of the battery is well supported by several excellent correlational studies (Delugach, 1991; Kaufman, 1990a). The manual provides a number of validity studies, factor-analytic results, research overviews, and state-of-the-art interpretive tables, which provide the examiner with a wealth of information. "The WPPSI-R is standing on a rock-solid psychometric foundation" (Kaufman, 1990a) .

In spite of its reported strengths, the WPPSI-R has numerous flaws. The WPPSI-R has an insufficient floor at the lowest age levels, which limits the test's ability to diagnose intellectual deficiency in young preschoolers (Delugach, 1991). The directions on some of the Performance subtests are not suitable for young children because they are not developmentally appropriate, and the heavy emphasis on response speed on some nonverbal test is inappropriate for young children who have not yet internalized the importance of working very quickly (Kaufman, 1990a). However, Delugach (1991) reports that if the directions are too difficult, the test provides procedures to ensure that the child understands the demands of the task.

The WPPSI-R is a useful assessment tool, but, like all others, it possesses certain weaknesses that limit its usefulness (Delugach, 1991). Examiners should be aware of the WPPSI-R's inherent strengths and weaknesses and keep them in mind during administration, scoring, and interpretation. The WPPSI-R may provide the examiner with useful information; however, "it does little to advance our basic understanding of the development and differentiation of intelligence or our understanding of the nature of individual differences in intelligence" (Buckhalt, 1991).

Detroit Tests of Learning Aptitude (DTLA-3)

Harry J. Baker and Bernice Leland recognized that the study of intra-individual strengths and weaknesses could be enhanced by the availability of a test battery composed of sort subtests that measured different abilities and that were standardized on the same population. In order to properly assess these intra-individual strengths and weaknesses, Baker and Leland developed the Detroit Tests of Learning Aptitude (DTLA) in 1935. The DTLA was comprised of 19 subtests and was appropriate for use with individuals between the ages of 4 and 19 years. A number of abilities could be assessed by the DTLA, including reasoning, verbal skills, time and space relationships, number, attention, and motor abilities.

Baker and Leland's original DTLA was used until it was revised in 1985 by Donald Hammill. The DTLA-2 was designed by Hammill to be used for individuals aged 6 years through 17 years 11months. The DTLA-2 included 11 subtests and 9 composites. The reviews of the DTLA-2 include both positive and negative evaluations. One of the primary criticisms was that there was not enough information provided on how the standardization sample was selected. An attempt was made to rectify the shortcoming of the DTLA-2 and incorporate many of the suggestions from the original reviews into the DTLA-3 (Hammill, 1991).

The DTLA-3, developed by Hammill in 1991, was designed to measure different, but interrelated, mental abilities for individuals ages 6 years through 17 years, 11 months. It is a battery of 11 subtests and has 16 composites that measure both general intelligence and discrete ability areas. Hammill and Bryant (1991) report that the DTLA-3 was greatly influenced by Spearman's two-factor theory (1927). This theory of "aptitude" consisted of a general factor "g" that is present in all intellectual pursuits, and specific factors that vary from task to task (McGhee, 1993).

The 11 subtests are used to form the 16 composite scores. The subtests are grouped into different combinations according to various hypothetical constructs that exist in current theories of intelligence and information-processing (McGhee, 1993). In general, the composite scores estimate general mental ability; however, they all do so in a somewhat different manner. The General Mental Ability Composite is formed by combining the standard scores of all 11 subtests, and thus, has been referred to as the best estimate of "g". The Optimal Level Composite is composed of the four largest standard scores that the individual earns. This individualized score is often referred to as the best estimate of a person's overall "potential." The Domain Composites may be divided into three areas; Linguistic, Attentional, and Motoric. Furthermore, there is a Verbal and Nonverbal Composite in the Linguistic domain, an Attention-Enhanced and Attention-Reduced Composite in

the Attentional Domain, and a Motor Enhanced and a Motor-Reduced composite in the Motoric Domain. Finally, there are the Theoretical Composites of the DTLA-3 on which the battery's subtests are constructed. The major theories upon which the subtests were developed include Horn and Cattell's (1966) fluid and crystallized intelligences, Das's (1973) simultaneous and successive processes, Jensen's (1980) associative and cognitive levels, and Wechsler's (1974, 1981, 1989) verbal and performance scales.

The DTLA-3 yields five types of scores: raw scores, subtest standard scores, composite quotients, percentiles, and age equivalents. Standard scores for the individual subtests have a mean of 10 and a standard deviation of 3 and the Composite Quotients have a mean of 100 and a standard deviation of 15. The individual subtest reliabilities range from .77 to .94 (median=.87) and the averaged alphas for the composites range from .89 to .96 (median=.94). To assess the DTLA-3's stability over time, the test-retest method was used with a sample of 34 children residing in Austin, Texas. The children, ages 6 through 16 years, were tested twice, with a two-week period between testings (Hammill, 1991). The results of this test-retest analysis indicate that individual subtest reliabilities range from .75 to .96 (median=.86) and composite reliabilities range from .81 to .96 (median=.89).

Critique. The DTLA-3 was designed to measure both general intelligence and discrete ability for children ages 6 years to 17 years 11 months. The DTLA-3 is not grounded in one specific theory but rather can be linked to a number of different theorists and their views on intelligence and achievement. This "eclectic" theorizing has resulted in the DTLA-3's numerous subtests, composites, and various combinations of the two that yield potentially important information about an individual's abilities.

Reliability and validity studies are encouraging but are based on specific and limited samples (VanLeirsburg, 1994). Additional research in this area would be beneficial. Furthermore, test-retest reliability data were collapsed across age levels, which makes it impossible to determine the stability of scores of the various age levels (Schmidt, 1994). The standardization sample was representative of the U.S. population but more information on socioeconomic level is needed (Schmidt, 1994). Also, there is no normative data reported for subjects with handicapping conditions and sample stratification for age was not equalized (VanLeirsburg, 1994).

The testing manual suggests that individual testing time may vary but that on average it takes 50 minutes to 2 hours to administer. Scoring and interpretation of the results is easy, yet it can be quite time-consuming without the aid of the accompanying computer program (VanLeirsburg, 1994). Despite apparent shortcomings, the DTLA-3 should be useful for eligibility or placement purposes and for research (Schmidt, 1994).

The Matrix Analogies Test

The Matrix Analogies Test (Naglieri, 1985) is composed of a set of figural matrices that can be used as a measure of general intelligence. There is little language involvement; therefore, the test is particularly well suited to assessing the intelligence of individuals with hearing impairments and language disabilities as well as the intelligence of children whose first language is not English.

The age range of the Matrix Analogies Test is from 5 through 18 years. There are two forms of the test: a group-administered form and the expanded form. The group-administered or short form consists of 34 multiple-choice items and the expanded form, which is individually administered, consists of 64 multiple-choice items. Raw scores on the expanded form can be converted into standard scores with a mean of 100 and a standard deviation of 15. These scores can then be converted into age equivalents and percentile ranks. On the short form of the test, the raw scores can be converted into percentiles, stanines, and age equivalents (Kamphaus, 1993).

The expanded form of the Matrix Analogies Test was standardized on a sample of 5,718 children in the early 1980s. The stratification of the sample matched U.S. Census statistics by race, sex, age, geographic region, community size, and socioeconomic status.

Internal consistency reliability of the expanded form ranges from .88 to .95. However, test-retest reliability of the total test score over a one-month interval was lower (.77). The validity of the expanded form has been evaluated primarily via correlations of other tests. Correlations with the WISC-R Full Scale IQ were .41 for a sample of 82 nonhandicapped children; .43 for Native American children, and .68 for hearing-impaired children (Kamphaus, 1993). According to Naglieri and

Prewett (1990), the trend across studies on the expanded form is for individuals to score about ten points lower on the Matrix Analogies test than on the Performance score of the WISC-R.

Critique. The Matrix Analogies Test—Expanded Form and the Matrix Analogies Test—Short Form can be useful tools in assessing the intelligence of children with communication or motor problems as well as the intelligence of children whose first language is not English. The test is user-friendly and easy to score and administer (Robinson, 1987). The Matrix Analogies Test is considerably more modern than many of its predecessors and its norming sample is more recent, larger, and more psychometrically sophisticated. The expanded form is a useful single-screener of intelligence for clinical or research use, based upon the mental processing of figural matrices by children (Kamphaus, 1993), while the short form may serve as a useful screening device (Robinson, 1987).

Differential Abilities Scales (DAS)

The DAS was developed by Elliott (1990a) and is an individually administered battery of cognitive and achievement tests for use with individuals aged 2½ through 17 years. The DAS Cognitive Battery has a preschool level and a school-age level. The preschool core consists of the following cognitive core subtests: Verbal Comprehension, Naming Vocabulary, Picture Similarities, Pattern Construction, Copying, and Early Number Concepts. The school-age cognitive core subtests include: Word Definitions, Similarities, Matrices, Sequential and Quantitative Reasoning, Recall of Designs, and Pattern Construction. The school-age level also includes reading-, mathematics-, and spelling-achievement tests that are referred to as "screeners." The same sample of subjects was used to develop the norms for the Cognitive and Achievement Batteries; therefore, intra- and inter-comparisons of the two domains is possible.

The DAS is not based on a specific theory of intelligence. Instead, the test's structure is based on tradition and statistical analysis. Nonetheless, the test is not theory-free, and, in fact, is based in part on "g" and the view of intelligence as hierarchical in nature (McGhee, 1993). Elliott (1990b) described his approach to the development of the DAS as "eclectic" and cited researchers such as Cattell, Horn, Das, Jensen, Thurstone, Vernon, and Spearman. Indeed, there are some clear-cut relationships between several DAS scales and theoretical constructs. For example, Horn's (1985, 1989) concepts of fluid and crystallized intelligence are measured quite well by the Nonverbal Reasoning and Verbal Ability scales, respectively. Elliott emphasizes Thurstone's ideas that the emphasis on intellectual assessment should be on the assessment and interpretation of distinct abilities (Kamphaus, 1993). Therefore, subtests were constructed to emphasize their unique variance, which should translate into unique abilities.

The cognitive portion of the DAS consists of core and diagnostic subtests designed to assess intelligence. The achievement portion measures skills in the areas of word reading, spelling, and mathematics. The core subtests are averaged to obtain the General Conceptual Ability (GCA) score and, depending on the age of the individual, additional composite scores, referred to as Cluster scores, are calculated (McGhee, 1993).

The individual Cognitive subtests have a mean of 50 and a standard deviation of 10. The GCA scores, Cluster scores, and Achievement scores, have a mean of 100 and a standard deviation of 15. Percentile ranks, age equivalents, and score comparisons are also available in the examiner's manual. Score comparisons provide a profile analysis and allow the examiner to ascertain information regarding aptitude-achievement discrepancies.

The norm sample of the DAS closely approximated U.S. census statistics estimated from 1986 to 1988, with the sample stratified by English-proficient, noninstitutionalized children from four U. S. geographic regions. At the preschool level, 175 children were included in each six-month age sample for ages 2 years, 6 months to 4 years, 11 months, with 200 children included at the 5-years to 5-years-11-months age-range. Children were divided equally at each age level for gender (Irvin, 1992). Exceptional children were also included in the standardization sample. Sex, race, geographic region, community size, and enrollment (for ages 2–5 years through 5–11years) in an educational program were controlled. Socioeconomic status was estimated using the average education level of the parents living with the child (Kamphaus, 1993).

The average internal-consistency estimates for the clusters at the school-age level are .88 for Verbal Ability, .90 for Nonverbal Reasoning Ability, and .92 for Spatial Ability. Internal consistency reliabilities of the subtests are also relatively

Table 4.5. Case Report Test Results

Wechsler Intelligence Scale for Children—Third Edition (WISC—III)

IQs

Verbal Scale 93±5 (32nd percentile)
Performance Scale 84±5 (14th percentile)
Full Scale 88±4 (21st percentile)

Factor Index Scores

Verbal Comprehension 95±5 (37th percentile)
Perceptual Organization 85±6 (16th percentile)
Freedom from Distractibility 101±8 (53rd percentile)
Processing Speed 91±7 (27th percentile)

Subtest Scaled Scores

	SCALED SCORE	PERCENTILE RANK
Information	8	25
Similarities	11-S	63
Arithmetic	8	25
Vocabulary	8	25
Comprehension	9	37
(Digit Span)	12-S	75
Picture Completion	3-W	1
Coding	9	37
Picture Arrangement	5	5
Block Design	9	37
Object Assembly	12-S	75
(Symbol Search)	7	16

Kaufman Brief Intelligence Test (K-B±T)

Composite IQ 93±6 (32nd percentile)

	SCALED SCORE	PERCENTILE RANK
Vocabulary	89	23
Matrices	98	45

Woodcock—Johnson Psycho Educational Battery-Revised: Tests of Cognitive Ability (WJ-R): Selected Subtests

SUBTEST/CLUSTER	STANDARD SCORE (± SEM)	PERCENTILE RANK
1. Memory for Names	108±3	69
8. Visual-Auditory Learning	97±5	41
Long-Term Retrieval (Tests 1 & 8)	103±4	59
4. Incomplete Words	90±3	26
11. Sound Blending	101±4	54
Auditory Processing (Tests 4 & 11)	96±7	40
5. Visual Closure	109±6	73
12. Picture Recognition	91±5	27
Visual Processing (Tests 5 & 12)	98±7	45
7. Analysis-Synthesis	103±4	59
14. Concept Formation	112±4	78
Fluid Reasoning (Tests 7 & 14)	108±4	71
Delayed Recall Memory for Names	69±7	02
Delayed Recall Visual-Auditory Learning	72±7	03

(continued)

Table 4.5. (Continued)

Kaufman Test of Educational Achievement-Comprehensive Form (K-TEA)

Reading Composite 103±4 (58th percentile)
Mathematics Composite 112±5 (79th percentile)
Battery Composite 104±3 (61st percentile)

	SCALED SCORE	PERCENTILE RANK	GRADE EQUIVALENT
Mathematics Applications	115	84	12.3
Reading Decoding	101	53	8.7
Spelling	92	30	6.4
Reading Comprehension	104	61	8.8
Mathematics Computation	107	68	9.3

strong, with only a few exceptions. The mean reliability coefficient for Recall of Objects, for example, is only .71, and for Recognition of Pictures, only .73 (Kamphaus, 1993).

Correlational research has shown good evidence of concurrent validity for the DAS (Kamphaus, 1993). With a sample of 27 children aged 7 to 14 years, the WISC-III Full Scale IQ correlated very highly with the DAS GCA score (.92), and the WISC-III Verbal IQ score correlated highly with the DAS Verbal Ability score (.87). The WISC-III Performance IQ correlated .78 with Nonverbal Reasoning and .82 with Spatial Ability. Additionally, the DAS Speed of Information Processing subtest score correlated .67 with the WISC-III Processing Speed Index score. The SB-IV Composite IQ also yielded strong correlations with the DAS GCA score, .88 for 9-and 10-year-olds and .85 for a sample of gifted children. The K-ABC Mental Processing Composite correlated with the DAS GCA yielded a correlation of .75 for 5- to 7-year-olds (Kamphaus, 1993).

Critique. In general, the professional reviews of the DAS seem to be quite positive. Sandoval (1992) believes that the DAS is one of the least obviously biased tests available today. The test development and the test results have resulted in a relatively culturally fair measure. However, its use with linguistically different children needs to be explored further (Sandoval, 1992). Sandoval does not provide clear evidence for support of his statement that the DAS is one of the least biased tests available. In fact, when one considers that the test is comprised of a number of verbal subtests it becomes very unclear how Sandoval could make such a claim. In fact, according to Bain (1991), the DAS appears to be useful in assessing both white and black students with learning problems; however, caution is recommended in using the DAS to predict achievement for Hispanic students because there is evidence that the test over-predicts achievement for this group based on group achievement results.

ILLUSTRATIVE CASE REPORT

An illustrative case report follows for Jared, an adolescent of 13 ½ years of age, who was referred for psychoeducational assessment because of school problems as well as emotional and behavior difficulties. The WISC-III, Kaufman Brief Intelligence Test (K-BIT), selected subtests from the Woodcock-Johnson Tests of Cognitive Abilities-Revised (WJ-R), Kaufman Test of Educational Achievement: Comprehensive Form (K-TEA), Rorschach Inkblot Test, Incomplete Sentence Test, Bender Gestalt Test, House-Tree-Person Test, and Kinetic Family Drawing Test were administered to Jared. The Child Behavior Checklist (CBCL) was also administered to Jared and his parents. (The client's name and pertinent identifying information have been changed to ensure anonymity.)

Referral and Background Information

Jared, a 13 ½-year-old male, was referred for evaluation by his parents, Mr. and Mrs. P. and Dr. Z., his psychiatrist. Jared's parents and Dr. Z would like to gain insight into Jared's current and previous level of cognitive and emotional functioning. Two months prior to this evaluation, on two separate occasions, Jared was placed in a psychiatric hospital for out-of-control behavior. After the hospitalizations, Jared was expelled from school for multiple suspensions and having a

weapon at school. He currently attends an alternative school for children with behavioral problems. School records, as well as parental reports, indicate that Jared's grades have been falling since 4th or 5th grade and that his problems in school appear to be both academic and emotional. More specifically, Jared has trouble focusing his attention and following directions, and he has temper tantrums. All of these troublesome behaviors occur at home and at school, and have resulted both in family conflict and concern.

Jared lives at home with his sister and both parents. He is the oldest of two children; his younger sister is 11 years old. Jared's birth history and early developmental history are unremarkable. He reached all developmental milestones within a normal time frame. It is reported that Jared has always had difficulty sleeping, even as an infant.

Jared attended preschool one to two mornings a week from age 2 to 5 years. His parents report that Jared cried when he first began school but that it only took him a short time to adjust. In the second grade, Jared was identified as gifted and was placed in the gifted program at his school. He reportedly did well in school and did not have any difficulties until the 4th grade when he had two different teachers. Mr. and Mrs. P. described that year at school, where the teachers alternated instruction, as problematic for Jared. Although Jared did well during his first few years in elementary school, his parents report that he has always had difficulty paying attention and sitting still for school work. His parents also related that as school became more difficult and advanced, Jared seemed to have more problems. Specifically, he had great difficulty with reading and organizational skills. School records, as well as parental and self-report, suggest that Jared both does well in, and enjoys, math. His parents also describe him as having a "great memory."

Jared's Middle School records indicate that he was quite disruptive in class. His grades prior to being expelled were primarily Ds and Fs. In comparison, Jared had approximately a B+ average, two years earlier, in the sixth grade. It appears that his grades did not begin to drop until the seventh grade, despite a reported history of difficulty with reading, focusing, and paying attention. When not in school, Jared enjoys skiing, surfing, watching television, and listening to music. He currently plays in a basketball league and he states that he really enjoys playing and that he looks forward to the games.

Appearance and Behavioral Characteristics

Jared is a handsome adolescent with big green eyes and short straight blonde hair. He was well groomed and dressed casually in stylish clothing. Jared appeared his stated age and his overall presentation was consistent with an independent adolescent. He made little-to-no eye contact, his posture was poor, and he did not converse easily with the examiners. Jared appeared to feel relatively uncomfortable during most of the testing and even though he complied with all requests, it was apparent that Jared retained his sense of privacy and minimal social involvement with the examiner. Jared provided one-word answers whenever possible and never initiated conversation. Jared participated in a considerable amount of testing and there were only two times when he seemed to feel a little more at ease, let his guard down somewhat, and conversed readily with the examiner. The first time that this occurred was during the third appointment when Jared described his relationship with his mother. While discussing their relationship, Jared sat up straight in his chair, made direct eye contact, and spoke with emotion. He stated that his relationship with his mother was strained and that he felt that she continually tries to interfere when he is talking to his father. During this discussion, Jared was able to express himself clearly and he communicated his feelings and thoughts in an age-appropriate manner. The second time that Jared opened up was during the home visit, where Jared was described as pleasant and able to interact well with the home-visit staff member.

Jared arrived for his appointments on time and was cooperative. He spoke softly and did not enunciate well, making it difficult to understand what he was saying at times. Although Jared seemed to be somewhat uninterested, he appeared to be trying his best. He was a little anxious, as evidenced by his excessive psychomotor activity and self-stimulating behavior. For example, Jared continually engaged in the following behaviors: touching his face, rubbing his arm, playing with his fingers, cracking his knuckles, cracking his neck, and tapping his fingers on the underside of his chair. These behaviors did not seem to distract him from what he was doing but rather seemed to soothe him emotionally and/or reduce his anxiety.

While solving problems and answering test questions, Jared spoke in a flat, monotone voice. The tone of his speech made it sound as if he

was bored and not trying; however, his scores and other behavioral observations suggest that he was, in fact, exerting full effort. When Jared felt that he did not know how do to something, or that he did not know the correct answer, he would become very frustrated. For example, he would get a look of disgust on his face or sometimes softly hit his fists on the table. In general, Jared did not respond well to feedback and encouragement from the examiner. In fact, he seemed to not recognize or care that he had received feedback.

Overall, Jared's affect was blunted, his responses were given in monotone and were brief. Although his affect was flat and he was not very communicative, he worked hard and was compliant. Jared was attentive and was not easily distracted. He appeared to be trying his best; however, when he felt that he was not doing well, he was very hard on himself. For example, he would say "I should have known that one" or he would roll his eyes and sigh as if he were exasperated with himself.

Tests Administered

Wechsler Intelligence Scale for Children—Third Edition (WISC-III)

Kaufman Brief Intelligence Test (K-BIT)

Woodcock-Johnson Psycho-Educational battery-Revised: Tests of Cognitive Ability—(WJ-R): Selected Subtests

Kaufman Test of Educational Achievement: Comprehensive Form (K-TEA)

Rorschach Inkblot Test

Incomplete Sentence Test

Bender Gestalt Visual-Motor Test

House-Tree-Person Test

Kinetic Family Drawing Test

Child Behavior Checklist (CBCL): Administered separately to Mr. and Mrs. P.; Jared responded to the Self-Report version.

Test Results (See Table 4.5) and Interpretation

Cognitive Functioning

Jared scored in the average range of intelligence on the Wechsler Intelligence Scale for Children-Third Edition (WISC-III), earning a Verbal IQ of 93±5, a Performance IQ of 84±5, and a Full Scale IQ of 88±4. The 9-point discrepancy between his Verbal and Performance IQs is not statistically significant, and indicates that he performs about equally well whether solving verbal problems and expressing his ideas orally or solving nonverbal items via the manipulation of concrete materials.

His overall performance on the WISC-III indicates that he is functioning at a little below average when compared to other children of his approximate age.

Jared was also administered the Kaufman Brief Intelligence Test (K-BIT). The K-BIT consists of two subtests, Vocabulary and Matrices, both of which have been shown to provide basic information about an individual's cognitive functioning. More specifically, the Vocabulary subtest measures verbal ability and crystallized knowledge while the Matrices subtest measures nonverbal ability and fluid reasoning. Jared's overall K-BIT Composite IQ was 93±6, 32nd percentile, (average), which is consistent with his WISC-III Composite IQ of 88.

Test results also indicate that Jared has a fluid reasoning strength, relative to his crystallized knowledge. On the K-BIT Vocabulary subtest (crystallized knowledge), Jared earned a standard score of 89±7, 23rd percentile, which is a little below average. In comparison, on the Matrices subtest (fluid reasoning), he earned a standard score of 98±7, 45th percentile, which is average. His fluid-reasoning strength is also evident on his performance on the WJ-R as well as the WISC-III. On the WJ-R, Jared earned his highest score on the fluid-reasoning cluster with a standard score of 108±4, 71st percentile. On the WISC-III, consistent with this fluid strength, Jared scored in the 75th percentile on Object Assembly and in the 63rd percentile.

On the WISC-III, Jared earned his lowest overall scores on two subtests on the Performance Scale: Picture Completion and Picture Arrangement. Picture Completion required him to look at a drawing of an object and/or individual that was somehow

incomplete, and to determine what important part of the picture was missing. Picture Arrangement, on the other hand, required him to look at a series of cards that when placed in the correct order, tell a sequential story about an event or situation. Both Picture Completion and Picture Arrangement involve drawings of people, places, and things and tend to involve human relationships. This type of content appears to be problematic for Jared and causes him discomfort. On these two subtests, it appears that Jared had difficulty solving the problems because he became emotionally overwhelmed with the material, and as a result he had great difficulty organizing his perceptions efficiently and accurately. Jared's performance on some of the subtests further substantiate this hypothesis. On the WJ-R, Jared earned a standard score of 109±6 (73rd percentile) on Visual Closure, a task that measures his ability to identify a drawing or picture that was altered in some way. For example, the picture may have been distorted, have missing lines or areas, or have a superimposed pattern. In addition to some type of distortion, the pictures or drawings are partially covered up by horizontal lines, making the pictures and drawings appear more distant and abstract. As a result of the abstraction, Jared was able to distance himself and process more effectively. In comparison, Jared earned a standard score of 91±5 (27th percentile) on Picture Recognition, a task that measures the ability to recognize a subset of previously presented pictures within a field of distracting pictures. On this task the pictures are concrete and straightforward, which leads Jared to distort them and perceive them less accurately, perhaps because he viewed them as threatening. Importantly, when reasoning tasks do not involve threatening content, Jared displays quite good ability. He earned a standard score of 108 (70th percentile) on the WJ-R Fluid Reasoning Scale, and performed at a similar level on a WISC-III puzzle-solving test (75th percentile). All of these tasks measure a child's ability to solve problems that are novel, and not dependent on schooling.

Jared's average to low average scores earned on the WISC-III, K-BIT, and WJ-R are in stark contrast to the scores that he reportedly earned on the WISC-R that was administered to him in 1989. The WISC-R scores that were reported indicated that Jared's Verbal IQ was 128, his Performance IQ was 133, and his Full Scale IQ was 135. However, these scores were based on that examiner's questionable use of a scoring system that involved eliminating some of Jared's scores on several pertinent subtests of the WISC-R and then calculating an estimate based on an incomplete set of subtests. There was no indication why these important subtests were eliminated, resulting in a prorated IQ; therefore, it is difficult, if not impossible, to determine if these results were valid or meaningful in any way.

To assess Jared's academic achievement abilities, he was administered the Kaufman Test of Educational Achievement-Comprehensive Form (K-TEA). Jared's achievement abilities ranged from the 30th percentile on Spelling to the 84th percentile on Mathematics Applications. In general, all of Jared's scores fell within the average range, although his excellent performance on Mathematics Applications was above average. In addition to the individual subtest scores, the K-TEA provides three composite scores of overall academic functioning. On the Reading Composite, Jared earned a standard score of 103±4, 58th percentile, grade equivalent 8.7, (average). On the Mathematics Composite he earned a standard score of 112±5, 79th percentile, grade equivalent 10.5, (above average). His Battery Composite standard score of 104±3, 61st percentile, was equivalent to grade 8.9 (average). These results indicate that Jared's academic abilities are generally average and that, based on his cognitive abilities, he is working up to his potential.

The most significant difficulty noted in this cognitive evaluation was Jared's deficient performance on two-delayed recall tasks. He scored in the 2nd and 3rd percentile on two subtests on the WJ-R that measure incidental retention of previously taught material. During the administration of these subtests, Jared appeared to be paying attention and he was very focused. On the WISC-III, Jared scored in the 75th percentile on the Digit-Span subtest which evaluates short-term memory and retrieval. Based on these results, attentional difficulties do not appear to be the cause of Jared's specific memory difficulties.

The results from the cognitive and achievement portions of this assessment suggest that Jared is a young adolescent of basically average intelligence. He has average academic achievement and inconsistent long-term memory. There is no indication that he has a learning disability or that he has any significant academic weaknesses. Instead, it appears that Jared's academic difficulties stem from emotional factors that inhibit and interfere with his ability to do his school work. More impor-

tantly, Jared's current emotional state is, in general, significantly impeding his cognitive functioning.

Personality Functioning

Projective personality testing indicates that Jared appears to approach his world in a careless and unsystematic way. He often spends an insufficient amount of time sorting out the important from the unimportant details of a situation, which leads him to make hasty decisions in an attempt to resolve issues. In general, this style of coping is often associated with anxiety and depression. Jared also has a tendency to approach new situations with a negative and oppositional attitude in an attempt to defend himself from environmental influences that he perceives to be potentially harmful.

Personality assessment also indicates that Jared becomes easily overwhelmed. His thoughts and feelings are not organized in a way to permit their controlled use, which tends to stimulate undeliberate and erratic behavior. In an effort to avoid complexity and ambiguity, he often becomes emotionally constricted, which eventually leads to emotional explosions and emotional lability.

Feelings significantly disrupt Jared's ability to perceive reality accurately. When Jared experiences either his own or other people's emotions he becomes overwhelmed and often reacts impulsively. In emotional situations, he is unable to think rationally and reflect upon the information he has received. Jared is much more likely to display emotion and action in such coping situations rather than reflect and think about what is occurring. This behavior often results in Jared losing his temper and ultimately getting into trouble for his actions. It also appears that arousal of his emotions significantly interferes with his work output. This hypothesis is supported by Jared's school records, which indicate that he has been able to do very little in school and that his grades are poor.

Test results also suggest that Jared has an unusually painful and critical introspective orientation. He has poor self-esteem and has great difficulty relating to others. Jared feels extremely uncomfortable around other people and has trouble empathizing with them; therefore, he limits and controls his emotional connections to others. Jared's tendency to view himself and others in a somewhat negative light is indicative of depression.

Currently, Jared is unwilling to exert himself intellectually. This unwillingness to put energy into cognitive activity appears to stem from both immaturity and oppositionality. Jared's approach to some of the testing was relatively immature. He revealed a lack of complexity and an immaturity in his thought processes. His approach to the testing indicates that he is very concrete. His concrete cognitive style was present on the cognitive portion of the testing as well. When asked to answer questions about what one should do in a variety of life situations (e.g., "If you saw a person fall on the street in front of you, what should you do?"), Jared gave simple responses that were not well thought-out. Concrete thinking is often associated with younger clients; however, developmentally, Jared should be beyond this stage. Additionally, Jared is unwilling to interact much with the world because of his stereotyped and negative view of it.

Diagnostic Summary

Jared appears to be a very troubled young man. He is easily overwhelmed with emotion and has extremely limited coping abilities. Jared's inability to deal with his feelings for himself, as well as his feelings toward others, is quite taxing on him. Jared does not seem to be able to handle his current situation effectively and he does not have any idea what he should do or to whom he can turn. Feeling backed into a corner and hopeless is difficult for any individual, but is especially trying for a young adolescent. To make matters worse, Jared's current emotional state is significantly impeding his cognitive abilities. He is unproductive and disruptive at school, and his emotional overload and lability have resulted in both disturbed interpersonal relationships as well as distorted thought processes.

Recommendations

The following recommendations have been made to assist Jared and his parents with Jared's academic, emotional, and behavioral difficulties. This assessment suggests that Jared's difficulties are the result of a complex set of variables and dynamics and should be addressed from a multimodal approach.

1. This evaluation suggests that Jared is a very disturbed young man who is in a state of crisis. Therefore, it is recommended that immediate and drastic interventions be made as soon as possible. Jared would benefit from placement in a long-term residential treatment center that would be able to provide him with full-time intensive therapy and treatment in a structured and safe environment.
2. Jared may also benefit from participating in individual therapy with a clinician who is able to provide him with a supportive and trusting relationship. Jared needs to work with a therapist who is trained to work with adolescents, depression, and Conduct Disorder.
3. Jared should continue taking his medication. Medication will help Jared keep his emotional lability under control, which will make it easier for him to focus on other aspects of his life.
4. Mr. and Mrs. P. may want to consider participating in family therapy with their younger daughter, in order to deal with the impact that Jared's behavior has had on the family. Often, younger siblings emulate their older brothers or sisters and it will be important for his sister to recognize and understand that Jared's behavior is not ideal and that there are more effective ways of behaving and coping with situations.
5. Jared would benefit from carrying a notebook with him for writing down important information that he needs to remember at a later time. It is sometimes difficult to get into the habit of making notes to oneself; however, Jared needs to be encouraged to do so because of his poor memory.

REFERENCES

American Psychological Association. (1990). Standards for educational and psychological tests and manuals. Washington, DC: Author.

Bain, S. K. (1991). Test Reviews: Differential Ability Scales. *Journal of Psychoeducational Assessment*, 9, 372–378.

Bogen, J. E. (1975). Some educational aspects of hemispheric specialization. *UCLA Educator, 17,* 24–32.

Bracken, B. A. (1985). A critical review of the Kaufman Assessment Battery for Children (K-ABC). *School Psychology Review, 14,* 21–36.

Brown, D. T. (1994). Review of the Kaufman Adolescent and Adult Intelligence Test (KAIT). *Journal of School Psychology, 32,* 85–99.

Buckhalt, J. A. (1991). A critical review of the Wechsler Preschool and Primary Scale of Intelligence Revised (WPPSI-R). *Journal of Psychoeducational Assessment, 9,* 271–279.

Canter, A. (1990). A new Binet, an old premise: A mismatch between technology and evolving practice. *Journal of Psychoeducational Assessment, 8,* 443–450.

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology, 54,* 1–22.

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

Cummings, R. E. (1985, Fall). Preferences of gifted students for selected teacher characteristics. *Gifted Child Quarterly, 29(4),* 160-163.

Das, J. P. (1973). Structure of cognitive abilities: Evidence for simultaneous and successive processing. *Journal of Educational Psychology, 65,* 103–108.

Das, J. P., Kirby, J. R., & Jarman, R. F. (1975). Simultaneous and successive synthesis: An alternative model for cognitive abilities. *Psychological Bulletin, 82,* 87–103.

Das, J. P., Kirby, J. R., & Jarman, R. F. (1979). *Simultaneous and successive processes.* New York: Academic Press.

Delugach, R. (1991). Test Review: Wechsler Preschool and Primary Scale of Intelligence-Revised. *Journal of Psychoeducational Assessment, 9,* 280–290.

Dumont, R., & Hagberg, C. (1994). Test Reviews: Kaufman Adolescent and Adult Intelligence Test (KAIT). *Journal of Psychoeducational Assessment, 12,* 190–196.

Dunn, L. M., & Dunn, L. M. (1981). *Peabody Picture Vocabulary Test-Revised.* Circle Pines, MN: American Guidance Serivice.

Elliott, C. D. (1990a). *Differential Ability Scales (DAS) administration and scoring manual.* San Antonio, TX: Psychological Corporation.

Esquirol, J. E. D. (1828). *Observations pour servir a l'histoire de l'idiotie.* [Observations used for the history of idiocy]. *Les Malidies Mentales.*

Flanagan, D. P., Alfonso, V. C., & Flanagan, R. (1994). A review of the Kaufman Adolescent and Adult Intelligence Test: An advancement in cognitive assessment? *School Psychology Review, 23,* 512–525.

Golden, C. J. (1981). The Luria-Nebraska Children's Battery: Theory and formulation. In: Hund, G. W. and Obrzut, J. E. (Eds.), *Neuropsychological Assessment of the School-age Child.* New York: Grune and Stratton.

Guilford, J. P. (1988). Some changes in the structure-of-intellect model. *Educational and Psychological Measurement, 48,* 1-4.

Guilford, J. P. (1989). Three faces of intellect. *American Psychologist, 14,* 459–479.

Guilford J. P. (1967). *The nature of human intelligence.* New York: McGraw-Hill.

Guilford, J. P., & Hoepfner, R. (1971). *The analysis of human intelligence.* New York: McGraw-Hill.

Hammill, D. D. (1991). *Interpretive Manual for Detroit Tests of Learning Aptitude: Third Edition.* Austin, TX: PRO-ED.

Hammill, D. D., & Bryant, B. R. (1991). *Interpretive Manual for Detroit Tests of Learning Aptitude-Primary: Second Edition.* Austin, TX: PRO-ED.

Hodapp, A. F. (1993). Correlation between Stanford-Binet IV and PPVT-R scores for young children. *Psychological Reports, 73,* 1152–1154.

Horn, J. L., & Cattell, R. B. (1967). Age difference in fluid and crystallized intelligence. *Acta Psychologica, 26,* 107–129.

Horn, J. L. (1985). Remodeling old model in intelligence. In B. B. Wolman (Ed.), *Handbook of intelligence: Theories, measurements, and applications* (pp. 267–300). New York: Wiley.

Horn, J. L. (1989). Cognitive diversity: A framework of learning. In P. L. Ackerman, R. J. Sternberg, & R. Glaser (Eds.), *Learning and individual differences* (pp. 61–116). New York: Freeman.

Horn, J. L., & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized intelligence. *Journal of Educational Psychology, 57,* 253–270.

Horn, J. L., & Hofer, S. M. (1992). Major abilities and development in the adult period. In R. J. Sternberg & C. A. Berg (Eds.), *Intellectual development* (pp. 44–99). Boston: Cambridge University Press.

Inhelder, B., & Piaget, J. (1958). *The growth of logical thinking from childhood to adolescence.* New York: Basic Books.

Irvin, M. G. (1992). Preschool assessment with the Differential Ability Scales (DAS). *Journal of Psychoeducational Assessment, 10,* 99–102.

Jensen, A. R. (1980). *Bias in mental testing.* New York: The Free Press.

Kamphaus, R. W. (1993). *Clinical assessment of children's intelligence.* Boston: Allyn & Bacon.

Kamphaus, R. W., Beres, K. A., Kaufman, A. S., & Kaufman, N. L. (1995). The Kaufman Assessment Battery for Children (K-ABC). In C. S. Newmark (Ed.), *Major psychological assessment instruments* (2nd ed.). Boston: Allyn & Bacon.

Kamphaus, R. W., & Reynolds, C. R. (1987). *Clinical and research applications of the K-ABC.* Circle Pines, MN: American Guidance Service.

Kaufman, A. S. (1990a). *Assessing adolescent and adult intelligence.* Boston: Allyn & Bacon.

Kaufman, A. S. (1994). *Intelligent testing with the WISC-III.* New York: John Wiley & Sons.

Kaufman, A. S., & Doppelt, J. E. (1976). Analysis of WISC-R standardization data in terms of the stratification variables. *Child Development, 47,* 165–171.

Kaufman, A. S., & Kamphaus, R. W. (1984). Factor analysis of the Kaufman Assessment Battery for Children (K-ABC) for ages 2½ through 12½ years. *Journal of Educational psychology, 76,* 623–637.

Kaufman, A. S., & Kamphaus, R. W. (1994). Factor analysis of the Kaufman Assessment Battery for Children (K-ABC) for ages 2½ through 12½ years. *Journal of Educational Psychology, 76,* 623-637.

Kaufman, A. S., & Kaufman, N. L. (1983c). *Interpretive manual for the Kaufman Assessment Battery for Children.* Circle Pines, MN: American Guidance Service.

Kaufman, A. S., & Kaufman, N. L. (1993). *Interpretive Manual for Kaufman Adolescent & Adult Intelligence Test.* Circle Pines, MN: American Guidance Service, Inc.

Kaufman, A. S., & Kaufman, N. L. In press. The Kaufman Adolescent and Adult Intelligence Test (KAIT). In D. P. Flanagan, J. L. Genshaft, & P.L Harrison (Eds.), *Beyond traditional intellectual assessment: Contemporary and emerging theories, tests, and issues.* New York: Guilford.

Keith, T. Z. (1985). Questioning the K-ABC: What does it measure? *Journal of Psychoeducational Assessment, 8,* 391–405.

Keith, T. Z., & Dunbar, S. B. (1984). Hierarchical factor analysis of the K-ABC: Testing alternate models. *Journal of Special Education, 18*(3), 367–375.

Kinsbourne, M. (Ed.). (1978). *Asymmetrical function of the brain.* Cambridge, MA: Cambridge University Press.

Levy, J., & Trevarthen, C. (1976). Metacontrol of hemispheric function in human split-brain patients. *Journal of Experimental Psychology: Human Perception and Performance, 2,* 299–312.

Luria, A. R. (1966a). *Higher cortical functions in man.* New York: Basic Books.

Luria, A. R. (1966b). *Human brain and psychological process.* New York: Harper & Row.

Luria, A. R. (1973). *The working brain: An introduction to neuro-psychology.* London: Penguin Books.

Luria, A. R. (1980). Higher cortical functions in man (2nd ed.). New York: Basic Books.

McCallum, S. (1985). Review of Peabody Picture Vocabulary Test-Revised. In O. K. Buros (Eds.), *Ninth Men-*

tal Measurements yearbook (pp. 1126–1128). Highland Park, NJ: Gryphon.

McCallum, R. S. (1990). Determining the factor structure of the Stanford-Binet: Fourth Edition-The right choice. *Journal of Psychoeducational Assessment, 8,* 436–442.

McCallum, R. S., & Merritt, F. M. (1983). Simultaneous-successive processing among college students. *Journal of Psychoeducational Assessment, 1,* 85–93.

McGhee, R. (1993). Fluid and crystallized intelligence: Confirmatory factor analysis of the Differential Ability Scales, Detroit Tests of Learning Aptitude-3, and Woodcock-Johnson Psycho-Educational Battery-Revised. In B. A. Bracken & R. S. McCallum (Eds.), *Journal of Psychoeducational Assessment monograph series, advances in psychoeducational assessment: Woodcock Johnson Psycho-Educational Battery-Revised* (pp. 39–53). Germantown, TN: Psychoeducational Corporation.

Merz, W. R. (1985) Test Review of Kaufman Assessment Battery for Children. In D. J. Keyser & R. C. Sweetland (Eds.), *Test Critiques* (pp. 393–405). Test Corporation of America.

Miller, T. L., & Reynolds, C. R. (1984). Special issue...The K-ABC. *Journal of Special Education, 8* (3), 207–448.

Naglieri, J. A. (1985). *Interpretive Manual for Matrix Analogies Test.* San Antonio, TX: The Psychological Corporation.

Naglieri, J. A., & Das, J. P. (1988). Planning-Arousal-Simultaneous-Successive (PASS): A model for assessment. *Journal of School Psychology, 26,* 35–48.

Naglieri, J. A., & Das, J .P. (1990). Planning, Attention, Simultaneous, and Successive (PASS) cognitive processes as a model for intelligence. *Journal of Psychoeducational Assessment, 8,* 303–337.

Naglieri, J. A., & Prewett, P. N. (1990). Nonverbal intelligence measures: A Selected review of instruments and their use. In C. R. Reynolds & R. W. Kamphaus (Eds.), *Handbook of psychological and educational assessment of Children: Intelligence and achievement* (pp. 348-376). New York: Gullford.

Obringer, S. J. (1988, November). *A survey of perceptions by school psychologists of the Stanford-Binet IV.* Paper presented at the meeting of the Mid-South Educational Research Association, Louisville, KY.

Perlman, M. D. (1986). *Toward an integration of a cognitive-dynamic view of personality: The relationship between defense mechanisms, cognitive style, attentional focus, and neuropsychological processing.* Unpublished doctoral dissertation, California School of Professional Psychology, San Diego.

Piaget, J. (1950). *The psychology of intelligence.* New York: Harcourt Brace.

Piaget, J. (1972). Intellectual evolution from adolescence to adulthood. *Human Development, 15,* 1–12.

Pintner, R. (1949). *Intelligence testing: Methods and results.* New York: Henry Holt.

Pintner, R., & Patterson, D. G. (1925). *A scale of performance.* New York: Appleton.

Reynolds, C. R. (1987). Playing IQ roulette with the Stanford-Binet, 4th edition. *Measurement and Evaluation in Counseling and Development, 20,* 139–141.

Reynolds, C. R., Kamphaus, R. W., & Rosenthal, B. L. (1988). Factor analysis of the Stanford-Binet Fourth Edition for ages 2 years through 23 years. *Measurement and Evaluation in Counseling and Development, 21,* 52–63.

Robinson, A. (1987). Review of Matrix Analogies Test. In D. J. Keyser & R. C. Sweetland (Eds.), *Test Critiques* (pp. 336–340). Test Corporation of America.

Sandoval, J. (1992). Test Reviews: Using the DAS with multicultural populations: Issues of test bias. *Journal of Psychoeducational Assessment, 10,* 88–91.

Sattler, J. M. (1988). *Assessment of children.* San Diego, CA: Jerome M. Sattler.

Schmidt, K. L. (1994). Review of Detroit Tests of Learning Aptitude-Third Edition. *Journal of Psychoeducational Assessment, 12,* 87–91.

Sequini E. (1907). *Idiocy: Its treatment by the physiological method.* New York: Bureau of Publications, Teachers College, Columbia University. (Original work published 1866)

Shapiro, D. (1965). *Neurotic styles.* New York: Basic Books.

Shaw, S. R., Swerdlik, M. E., & Laurent, J. (1993). Review of the WISC-III. In B. A. Bracken & R. S. McCallum (Eds.), *Journal of Psychoeducational Assessment monograph series, advances in psychoeducational assessment: Wechsler Intelligence Scale for Children-Third Edition* (pp. 151–160). Germantown, TN: Psychoeducational Corporation.

Sperry, R. W. (1968). Hemisphere deconnection and unity in conscious awareness. *American Psychologist, 23,* 723–733.

Sperry, R. W. (1974). Lateral specialization in the surgically separated hemispheres. In F. O Schmitt & F. G. Worden (Eds.), *The neurosciences: Third study program.* Cambridge, MA: MIT Press.

Spruill, J. (1987). Review of Stanford-Binet Intelligence Scale, Fourth Edition. In D. J. Keyser & R. C. Sweetland (Eds.), *Test Critiques* (pp. 544–559). Test Corporation of America.

Sternberg, R. J. (1993). Rocky's back again: A review of the WISC-III. In B. A. Bracken & R. S. McCallum (Eds.), *Journal of Psychoeducational Assessment monograph series, advances in psychoeducational assessment: Wechsler Intelligence Scale for Children-Third Edition* (pp. 161–164). Germantown, TN: Psychoeducational Corporation.

Thorndike, R. L., Hagen, E. P., & Sattler, J. M. (1986). *Technical manual, Stanford-Binet Intelligence Scale: Fourth Edition.* Chicago: Riverside Publishing.

Tuddenham, R. D. (1962). The nature and measurement of intelligence. In L. J. Postman (Ed.), *Psychology in the making* (pp. 469–525). New York: Knopf.

Umberger, F. G. (1985). Review of Peabody Picture Vocabulary Test-Revised. In D. J. Keyser & R. C. Sweetland (Eds.), *Test Critiques* (pp. 488–495). Test Corporation of America.

VanLeirsburg, P. (1994). Review of Detroit Tests of Learning Aptitude-3. In D. J. Keyser & R. C. Sweetland (Eds.), *Test Critiques* (pp. 219–225). Test Corporation of America.

Wada, J., Clarke, R., & Hamm, A. (1975). Cerebral hemisphere asymmetry in humans. *Achieves of Neurology, 37,* 234–246.

Webster, R. E. (1994). Review of Woodcock-Johnson Psycho-educational Battery-Revised. In D. J. Keyser, & R. C. Sweetland (Eds.), *Test Critiques* (pp. 804–815). Test Corporation of America.

Wechsler D. (1974). *Manual for the Wechsler Intelligence Scale for Children-Revised.* San Antonio: Psychological Corporation.

Wechsler, D. (1981). *Manual for the Wechsler Adult Intelligence Scale-Revised (WAIS-R).* San Antonio, TX: The Psychological Coporation.

Wechsler, D. (1989). *Manual for the Wechsler Preschool and Primary Scale of Intelligence-Revised (WPPSI-R).* San Antonio, TX: Psychological Corporation.

Wechsler, D. (1991). *Manual for the Wechsler Intelligence Scale for Children-Third Edition, (WISC-III).* San Antonio, TX: Psychological Corporation.

Wiig, E. H. (1985). Review of Peabody Picture Vocabulary Test-Revised. In O. K. Buross (Eds.). *Ninth Mental Measurements yearbook* (pp. 1126–1128). Highland Park, NJ: Gryphon.

Wolf, R. H. (1969). The emergence of Binet's conceptions and measurement of intelligence: A case history of the creative process. Part ii. *Journal of the History of the Behavioral Science, 5,* 207–237.

Woodcock, R. W. (1990). Theoretical foundations of the WJ-R Measures of Cognitive Ability. *Journal of Psychoeducational Assessment, 8,* 231–258.

Woodcock, R. W., & Mather, N. (1989). *WJ-R Tests of Cognitive Ability-Standard and Supplemental Batteries: Examiner's Manual.* In R. W. Woodcock & M. B. Johnson, *Woodcock-Johnson psycho-educational battery-revised.* Allen, TX: DLM Teaching Resources.

Yerkes, R. M. (1917). The Binet versus the point scale method of measuring intelligence. *Journal of Applied Psychology, 1,* (11) 1–122.

CHAPTER 5

ASSESSMENT OF ADULT INTELLIGENCE WITH THE WAIS-III

David S. Tulsky
Jianjun Zhu
Aurelio Prifitera

INTRODUCTION

Since the publication of the Wechsler-Bellevue Intelligence Scale for adults in 1939, this scale and its revisions and derivatives, including the Wechsler Adult Intelligence Scale (WAIS) (Wechsler, 1955) and the Wechsler Adult Intelligence Scale-Revised (WAIS-R) (Wechsler, 1981), have had a tremendous influence on the field of psychology (see Kaufman, 1990; Lindemann & Matarazzo, 1984). In studies where the frequency of using assessment instruments has been examined, the Wechsler scales repeatedly come out as one of the most often-used scales. For example, in a study conducted by Harrison, Kaufman, Hickman, and Kaufman (1988), 97 percent of the respondents routinely gave the WAIS-R. More recently, Watkins, Campbell, Neiberding, and Hallmark (1995) reported that 93 percent of the 410 psychologists they surveyed administer the WAIS-R at least occasionally. Other surveys have also found that the Wechsler scales are used on such a frequent basis (Lubin, Larson, & Matarazzo, 1984; Lubin, Larson, Matarazzo, & Seever, 1985; Piotrowski & Keller, 1989).

These scales and especially the development of the new Wechsler Adult Intelligence Scale, Third Edition (WAIS-III) (Wechsler, 1997a), will be the focus of this chapter.

DAVID WECHSLER AND THE WECHSLER INTELLIGENCE SCALES

David Wechsler began using scales of intellectual functioning in his work with the U.S. Army during World War I. Dr. Wechsler was in charge of performing individual testing on people who had failed the group-administered tests. From this experience, he learned which tasks could be used to measure intelligence and used them in his testing sessions. He realized that intelligence could and should be measured by a diverse set of tasks, some verbal and some perceptual; and he saw the need for a new intelligence test, constructed for adults, that emphasized verbal and nonverbal intelligence. This idea of measuring both verbal and performance intelligence (rather than just global intelligence) revolutionized the field of cognitive testing. Wechsler (1944) wrote:

> The most obviously useful feature of the Wechsler-Bellevue scales is their division into a Verbal and Performance part....Its a [*sic*] priori value is that it makes a possible comparison between a subject's facility in using words and symbols and his ability to manipulate objects, and to perceive visual patterns. In practice this division is substantiated by differences between posited abilities and various occupational aptitudes. Clerical workers and teachers, in general, do much better on verbal tests, whereas manual workers and mechanics do better on perfor-

> mance. The correlations are sufficiently high to be of value in vocational guidance, particularly with adolescents of high school age.
>
> Apart from their possible relation to vocational aptitudes, differences between verbal and performance test scores, particularly when large, have a special interest for the clinician because such discrepancies are frequently associated with certain types of mental pathology. (p. 146)

David Wechsler had been well trained in matters of intellectual functioning as well as in merging and integrating what would appear to be a set of diverse ideas about intelligence testing. At Columbia University, Dr. Wechsler spent years training with James McKeen Cattell, E. L. Thorndike, and R. S. Woodworth. He was also fortunate to have spent three months studying with Charles Spearman and Karl Pearson in London, and he took pride in being trained, first and foremost, as a psychometrician. Several of his mentors (Cattell, Thorndike, and Spearman) had strong beliefs about intelligence and intellectual testing, and Wechsler believed that "they were all right" and that he should merge these different viewpoints together into a theory and framework that everyone could accept (Shackelford, 1978).

This goal was more difficult than it might sound because two of his mentors, Thorndike and Spearman, were locked in one of the greatest debates about intelligence testing. Spearman (1904, 1927) believed that intelligence was mediated by a general "g" factor that was responsible for how one would perform on a variety of tasks. Thorndike interpreted the data differently, believing that intellect consisted of several distinct abilities (see Thorndike, Lay, & Dean, 1909). Wechsler had the difficult task of bridging the gap between the beliefs of these two individuals. Throughout his writing, Wechsler (1944) graciously paid tribute to the contributions of both of these great psychologists while not choosing "sides" in the debate.

Wechsler's Concept of Intelligence

Wechsler defined intelligence as "the capacity of the individual to act purposefully, to think rationally, and to deal more effectively with his environment" (Wechsler, 1944; p. 3). In this definition of intelligence, he tried to include elements from other leading theorists and researchers of the time (e.g., Thordike, Spearman, Thurstone; see proceedings from the 1921 symposium, Henmon et al., 1921 and Thorndike et al., 1921). Wechsler believed that a definition had to be accepted by ones' peers first and foremost in order to gain acceptance (Shackelford, 1978).

Congruent with Spearman's ideas, Wechsler believed that global intelligence was important and meaningful as it measured the individual's overall behavior. However, similar to Thorndike, he also believed it was made up of specific abilities, each of which was important and different from one another. Hence, he emphasized the importance of sampling a variety of intellectual tasks. Wechsler (1974) wrote:

> To the extent that tests are particular modes of communication, they may be regarded as different *languages*. These languages may be easier or harder for different subjects, but it cannot be assumed that one language is necessarily more valid than another. Intelligence can manifest itself in many forms, and an intelligence scale, to be effective as well as fair, must utilize as many different languages (tests) as possible (p. 5).

Bridging the ideas of Spearman and Thorndike, Wechsler (1939) developed a test that included a general intelligence measure (FSIQ) while, at the same time, emphasized that there were two broad types of abilities, Verbal and Performance, that should be analyzed separately to make inferences about an individual's intellectual functioning. The Full-Scale Intelligence Quotient (FSIQ) captures Spearman's idea about a general intelligence, which was characterized as a dominant "g" or general factor with much smaller, less influential "s" or specific factors to guide intelligence. Wechsler agreed with parts of Spearman's theory, namely that there was an overall intelligence. Wechsler even wrote that "Professor Spearman's generalized proof of the two factor theory of human abilities constitutes one of the greatest discoveries of psychology" (Wechsler, 1944; p. 6). Contrary to Spearman's view, however, Wechsler placed more emphasis on the importance of the specific factors and even printed tables so that examiners could review the differences between various types of abilities (e.g., Verbal-Performance discrepancies; Wechsler, 1944).

Thorndike's influence can be seen in Wechsler's writing as he discusses the importance of each subtest and the ability of the examiner to perform profile analyses (e.g., examining differences between subtests). The Wechsler-Bellevue (and all of the derivatives) contains subtests designed to measure

Table 5.1. WAIS: III Subtests Grouped According to Verbal and Performance IQ Scales

VERBAL	PERFORMANCE
Vocabulary	Picture Completion
Similarities	Digit Symbol–Coding
Arithmetic	Block Design
Digit Span	Matrix Reasoning
Information	Picture Arrangement
Comprehension	

Table 5.2. WAIS-III Subtests Grouped According to Indexes

VERBAL COMPREHENSION	PERCEPTUAL ORGANIZATION	WORKING MEMORY	PROCESSING SPEED
Vocabulary	Picture Completion	Arithmetic	Digit Symbol-Coding
Similarities	Block Design	Digit Span	Symbol Search
Information	Matrix Reasoning	Letter-Number Sequencing	

qualitatively different types of cognitive abilities like abstract and verbal reasoning (e.g., Similarities, Vocabulary), nonverbal reasoning (e.g., Block Design, Object Assembly), and practical intelligence (e.g., Picture Arrangement, Comprehension). Building a scale that was composed of multiple subtests, each of which could be grouped into different types of intelligence, would allow the scale to match Thorndike's ideas, while at the same time these abilities could be aggregated into a single "global" score, which would allow the scale to coincide with Spearman's concepts. Through the structure of the Wechsler-Bellevue, David Wechsler found a way to "walk the fine line" between a global and a multi-factorial model of intellectual functioning.

Despite the many abilities that the Wechsler tests measure, David Wechsler also believed that his scale was not a complete measure of intelligence and that there were some elements missing in his definition of intelligence. He reviewed factor-analytic studies on the Wechsler scales and knew that they only accounted for a percentage of the overall variance of intelligence. From these data, he thought that there must be something else: a group of attributes that contributed to this unexplained variance. Wechsler believed that these attributes, or nonintellective factors, as he called them, were not so much skills as they were traits and included such factors as planning and goal awareness, field dependence, persistence, and enthusiasm (Wechsler, 1950). He believed that these factors contribute to intelligent behavior. These were called the nonintellective aspects of intelligent behavior.

Introduction to the WAIS-III

The WAIS-III is an individually-administered test of intellectual ability for people aged 16–89 years. It is administered in 60–75 minutes and consists of 14 subtests. Like the previous versions, the WAIS-III yields three intelligence composite scores: a Verbal Intelligence Quotient (VIQ), a Performance Intelligence Quotient (PIQ), and a Full Scale Intelligence Quotient (FSIQ). The IQs have a mean of 100 and a standard deviation of 15. Table 5.1 shows the set of six Verbal and five Performance subtests that can be combined to yield Verbal, Performance, and Full-Scale IQ scores on the WAIS-III. A new Matrix Reasoning subtest has replaced Object Assembly (used in previous Wechsler editions) on the Performance and Full-Scale IQ score.

Object Assembly has been included as an optional subtest for the IQ scales. It can be used to replace a spoiled Performance subtest when deriving IQ scores or it can replace another Performance subtest during retesting to help reduce the practice effects. Also, for those who want to use the same subtests as on the WAIS-R, to calculate PIQ and FSIQ, Object Assembly can be substituted for Matrix Reasoning.

A different subset of 11 subtests can also be combined to obtain a set of four index scores. Table 5.2 lists these subtests and how they relate to

the four Index scores: Verbal Comprehension Index (VCI), Perceptual Organization Index (POI), Working Memory Index (WMI), and Processing Speed Index (PSI). These index scores consist of more refined domains of cognitive functioning than do the IQ scores. For practical reasons, the index scores were limited to 11 subtests, with three subtests each for the VCI, POI, and WMI, and two subtests for the PSI.

The subtests vary in content from tasks such as defining vocabulary words, stating abstract relations between two objects or concepts, repeating a string of digits, putting puzzles together, putting blocks together to match a pattern, and sequencing a set of pictures to tell a story. Descriptions of each subtest and what they are purported to assess are discussed in the literature (Matarazzo, 1972; Kaufman, 1991, 1994; Sattler, 1992).

The scoring of each subtest differs. Some are dichotomously scored, some have consistent partial credits (0, 1, 2) and some vary because of differential weighting and time bonuses among the items. On 10 of the subtests, the item order is based on difficulty, which we believe approximates a Guttman pattern (Guttman, 1944). There are discontinue rules (e.g., 3 consecutive scores of 0), that are built on the assumption that the examinee would receive scores of 0 on any items that would be administered beyond the discontinue rule. This serves to reduce administration time and to not tax an individual. Digit Symbol-Coding and Symbol Search differ from the other subtests in that they are timed subtests on which the examinee completes as many items as possible within a 120-second time limit.

Scaled scores are presented in a lookup table based on the sum of the item scores for each subtest by age group. The WAIS-III deviates from its predecessors by basing subtest scores on age corrected scaled scores rather than on the performance of a younger reference group made up of individuals between the ages of 20 and 34 years. The distribution of each subtest was normed to a scale with a mean of 10 and standard deviation of 3. The subtest scores are normed according to 13 age bands (ranging from 16 to 89 years). These age-corrected scaled scores would then be summed to develop composite IQ or Index Scores.

Goals of the WAIS-III Revision

The first goal of the revision, to update the norms, stems from the fact that the normative information for intelligence tests becomes outdated over time and IQ scores become inflated. Joseph Matarazzo (1972) and James Flynn (1984, 1987) have written about this phenomenon of shifts in IQ norms. Dr. Matarazzo wrote that "it is imperative that such [age] norms be periodically updated lest they be less than fully efficient for the re-examination of individuals living in a social-cultural-educational milieu potentially very different from the one which influenced the individuals constituting the norms for that same age group in an earlier era." (Matarazzo, 1972; p. 11). Flynn's systematic review of this issue has shown that IQ scores tend to become inflated over time (Flynn, 1984) with the average IQ score drifting upward. Individuals appear to gain approximately 3-5 IQ points over a 10-year period. Generally, the phenomenon is more prevalent in the performance scales than it is in the verbal scales.

Based on these findings, the WAIS-III contains a contemporary, representative sample from which the IQ norms have been "re-anchored" at 100. Comparisons between the WAIS-III and WAIS-R scores reveal how "outdated" the norms on the WAIS-R had become. *The WAIS-III–WMS-III Technical Manual* (The Psychological Corporation, 1997) reports data on 192 individuals who completed both the WAIS-R and the WAIS-III. Examinees took the two scales in two sessions, 2–12 weeks apart, in a counterbalanced order. Consistent with the a priori predictions, the average FSIQ and PIQ scores were higher for the WAIS-R than the WAIS-III and the VIQ scores were relatively unchanged. The average FSIQ score on the WAIS-R was 2.9 IQ points higher than the corresponding average score on the WAIS-III, and the WAIS-R PIQ score was 4.8 IQ points higher. This finding adds further support to the hypothesis that IQ inflation is truly occurring. This inflation rate, however, is slightly lower than that which would have been expected from previously reported values (Flynn, 1984). Based upon the so-called "Flynn effect" alone, the average FSIQ would be increasing at a constant rate each year (e.g., an increase of one-third to one-half points per year), so that the average FSIQ of the WAIS-R would have been expected to be as high as 106–109 IQ points.

There are several reasons that the WAIS-R and the WAIS-III differences might be lower than predicted (see Zhu & Tulsky, 1999). Simply adding a constant oversimplifies the relation between the two tests. Besides this overall "Flynn effect,"

many other factors, such as practice effect, design differences between the two tests, floor and ceiling effect, other psychometric factors, (Bracken, 1988; Kamphaus, 1993; Zhu & Tulsky, 1997), and the interaction among these factors may affect the score discrepancies across the two testings. For instance, there are some significant differences between the WAIS-R and the WAIS-III that may be accounting for some of the differences. Most salient, the replacing of Object Assembly with Matrix Reasoning and the de-emphasis of timed bonus points in the WAIS-III may explain the difference between the two measures. Additionally, careful effort was taken to ensure that a representative proportion of individuals across the entire range of ability was sampled on the WAIS-III. To prevent truncated norms, 29 examinees with mental retardation were added to the overall standardization sample to ensure that the correct proportion of examinees (approximately 2.3 percent) that had FSIQ scores below 70 were included in the sample (Tulsky & Zhu, 1997). This effort may shrink the difference between the WAIS-R and WAIS-III.

The second goal of the revision was to extend the age range. Individuals in the United States are living longer. Current estimates place the average life expectancy at birth at more than 78 years for women and 72 years for men (Rosenberg, Ventura, Maurer, Heuser, & Freedman, 1996; La Rue, 1992). However, the WAIS-R only has normative information for people up to 74 years of age, and hence, it is becoming less sufficient for estimating the intelligence of older adults. Previously, to compensate for this deficit, two independent research teams have conducted studies to extend the WAIS-R norms upward for an older adult population. Ryan, Paolo, & Brungardt (1990) developed norms for older adults using a sample of 130 people (60 individuals who were between the ages of 75 and 79 years and 70 who were 80 years old and up). Attempts were made to match the sampling stratification criteria of the WAIS-R as much as possible. Concurrently, in an independent project, researchers at the Mayo Clinic collected normative data on 512 individuals between 56 and 97 years of age (Ivnik, et al., 1992). They deviated from the WAIS-R scoring technique by developing "age-specific" raw-score-to-scale-score conversions rather than basing the conversion on the optimal functioning "reference" group. Using the 56–74 year-old sample as a reference point, the research group also spent a considerable amount of time investigating the similarities between the Mayo Older Adult Normative Studies (MOANS) norms and the WAIS-R standardization sample norms so that they could make their norms as similar as possible to the WAIS-R. For the WAIS-III, the goal was to extend the normative information up to 89 years of age, allowing for appropriate use of scores for individuals in this older age range.

A third goal was to improve the item content of the subtests. A number of items were outdated and needed replacement. Additionally, some examiners have criticized the WAIS-R for containing some items that appear to be biased against certain groups. Extensive bias analyses and reviews were conducted so that biased items could be removed and replaced in the new revision (Chen, Tulsky, & Tang, 1997).

The fourth goal of the project was to update the artwork and make the WAIS-III more attractive for examinees. The WAIS-R was published in 1981 using the styles from the original Wechsler-Bellevue. Not only was some of the artwork outdated and unattractive, but some of the visual stimuli were small, putting individuals with visual acuity problems at a disadvantage. Several steps were taken to make the WAIS-III stimuli more appropriate for examinees. The Picture-Completion items were redrawn, enlarged, and colorized and the Picture-Arrangement cards were redrawn, enlarged, and modernized. The Digit Symbol-Coding subtest features more space between the items and keys to help assist left-handed examinees who might otherwise block the key as they were working. Finally, the WAIS-III Object-Assembly layout shield was modified radically to include the subtest instructions, and it was constructed of heavy card stock so that it could stand up on the table. The puzzle pieces themselves have numbers printed on the back to assist the examiner in laying out the pieces.

The fifth goal was to enhance the clinical utility of the scale, and this was accomplished in several ways. First, additional index scores were included in the WAIS-III. Some researchers have written about the limitations of the IQ score (Kaplan, 1988; Lezak, 1988, 1995). Others have suggested that the scale should measure a wider spectrum of domains of cognitive functioning (Malec et al., 1992). To incorporate some of the advances in the field, when the Wechsler Intelligence Scale for Children-Third Edition (WISC-III; Wechsler, 1991) was published, new factor-based Index scores (e.g., Verbal Comprehension, Perceptual Organizational, Freedom from Distractibility, and

Processing Speed) were added in addition to the traditional IQ composite scores. The WAIS-III revision includes a similar alternate index-scoring system, in addition to the traditional IQ-scoring system. New optional subtests have been developed to assess abilities on a hypothesized 3rd factor (Working Memory) and a 4th factor (Processing Speed). Specifically, Letter-Number Sequencing was included to measure Working Memory, and a second subtest, Symbol Search, was designed to measure Processing Speed.

Additionally, some optional procedures, such as testing incidental learning after the Digit Symbol-Coding administration (Hart, Kwentus, Wade, & Hamer, 1987; Kaplan, Fein, Morris, & Delis, 1991), were added to the WAIS-III-standardization edition. These procedures were based on the "process approach" to interpretation that was advocated by Kaplan and others. They were designed to help the examiner determine the nature of errors committed on the standardized tests.

The *WAIS-III Administration and Scoring Manual* (Wechsler, 1997a) includes optional normative tables designed to assist the clinician in the interpretation of scores. Besides the critical values for statistical significance of discrepancy, base rates of discrepancies between scores are presented in the manual. Matarazzo & Herman (1985) were the first to publish such tables based on the WAIS-R standardization sample and they demonstrated that VIQ-PIQ difference scores could be statistically significant but not clinically meaningful. Statistically significant scores would suggest that the difference score was "real" or that it was significantly different from 0. The base rates show how frequently such differences do occur in the population and even though someone might be better at one skill (Verbal or Performance) than the other skill, it might occur in a large percentage of the general population. These base-rate tables, therefore, allow the clinician to interpret the score based on the frequency at which such discrepancies occur.

Significant effort was made to enhance the measurement of the WAIS-III in individuals with very low or impaired intellectual functioning and other clinically relevant groups (e.g., people with mental retardation, people with neuropsychological impairment). With the WAIS-R, a 70–74-year-old person who cannot answer one item correctly can still receive a VIQ score of 60 and a PIQ score of 61 points! This was likely a result of the subtests having a restricted floor, the normative sample possibly not containing enough individuals whose true score extended that low, and the subtest scaled scores not extending more than 3 standard deviations below average. The floor of the WAIS-III extends lower than its predecessors, extending down to 45 for FSIQ, 47 for PIQ, and 48 for VIQ. To help validate that accurate scores were being obtained for people with low intellectual functioning, data on 62 people with moderate mental retardation and 46 people with mild mental retardation were obtained. The original diagnosis for each examinee was made using DSM-IV criteria (which included an appropriate score on an IQ test (other than the WAIS-III) and impairment in adaptive functioning. Roughly 83 percent of IQ scores in the mild group had WAIS-III IQ scores between 53 and 70 and 82 percent of the WAIS-III scores for the examinees in the moderate group had IQ scores between 45 and 52 (Tulsky & Zhu, 1997).

The sixth goal was to decrease the emphasis on timed performance. One criticism of the WAIS-R has been that some of the subtests are too dependent upon quick performance (Kaufman, 1990). For instance, on the Object Assembly subtest of the WAIS-R, in which subjects put puzzle pieces together, an examinee may earn up to 12 raw-score points (e.g., 29 percent additional raw-score points) as time-bonus points for speedy performance. This could result in a difference between 7 and 10 subtest scaled-score points. Hence, another objective was to reduce the contribution of speed and bonus points to the Performance IQ score wherever it is possible. To help achieve this goal, a new untimed performance subtest, Matrix Reasoning, was included.

The seventh goal was to enhance the measurement of fluid reasoning. Several recent theories of cognitive functioning have emphasized the importance of measuring fluid reasoning, or the ability to perform abstract mental operations (Sternberg, 1995). Matrix-reasoning tasks are considered typical of this type of ability, hence, the addition of this subtest to the WAIS-III.

Eighth, the theoretical structure of the WAIS-III was strengthened. Contemporary research has pointed out that intelligence encompasses more than what is measured by VIQ and PIQ scores (Carroll, 1993; Carroll, 1997). Reviews of factor-analytic work on the Wechsler scales have suggested that there are either three domains of cognitive functioning (Cohen, 1952a, 1952b, 1957a, 1957b, 1959; Leckliter, Matarazzo, & Silverstein, 1986) or, in the children's version after an optional Symbol Search subtest was included, that there are four domains of

cognitive functioning (Wechsler, 1991; Roid, Prifitera, & Weiss, 1993). Current theories of Working Memory (e.g., Baddeley, 1986; Kyllonen, 1987; Kyllonen & Christal, 1990) and Information Processing (e.g., Kyllonen, 1987) were used in developing new additional subtests on the WAIS-III. These subtests help expand the domains of cognitive functioning that are measured by the WAIS-III.

Ninth, the WAIS-III is linked with other tests such as Wechsler Individual Achievement Test (The Psychological Corporation, 1992) and Wechsler Memory Scale-Third Edition (WMS-III) (Wechsler, 1997b) to help the clinician interpret scores and patterns of scores. Significantly, the standardization sample was co-normed with the WMS-III. This linkage allows clinicians to examine IQ and memory relationships and discrepancy scores. Moreover, the linkage assists them in the interpretation of additional domains of cognitive functioning that include both intelligence and memory assessment.

Finally, extensive work has been performed to validate the new instrument and to demonstrate comparability between the WAIS-III and WAIS-R. Correlations between the WAIS-III and the WAIS-R, WISC-III, and the Stanford-Binet, 4th Edition, demonstrate that the WAIS-III is correlated with other instruments measuring intellectual functioning (The Psychological Corporation, 1997). The correlations between FSIQ on the WAIS-III and the general composite scores of these other instruments range from .88 to .93. The correlation of FSIQ with the *Raven's Standard Progressive Matrices* (SPM) (Raven, 1976), a nonverbal task of abstract ability, is lower, (r=.64); however, as expected, SPM has higher correlations with PIQ (r=.79) and the Matrix Reasoning subtest on the WAIS-III (r=.81).

The WAIS-III was also tested in a series of clinical validity studies with more than 600 individuals with neuropsychological impairment (e.g., Alzheimer's dementia, traumatic brain injury), psychiatric diagnosis (e.g., schizophrenia, depression), learning disabilities, mental retardation, and hearing impairment or deafness. From these studies, different patterns of performance tended to occur (especially among the index scores) and they provided an initial demonstration of the construct validity and clinical utility of the WAIS-III. A detailed description of these studies has been reported in *The WAIS-III–WMS-III Technical Manual* (The Psychological Corporation, 1997).

Development of the New WAIS-III Subtests

To enhance the measure of fluid reasoning, working memory, and processing speed, the WAIS-III includes three new subtests: Matrix Reasoning, Symbol Search, and Letter-Number Sequencing. The development of these new subtests will be described in the following sections.

Matrix Reasoning

In the WAIS-III, the new Matrix Reasoning subtest replaces Object Assembly[1] as a standard subtest and contributes to PIQ, FSIQ, and POI scores. As stated earlier, this subtest was added because it has long been recognized that matrix analogy tasks are good measures of "fluid" intelligence (Sternberg, 1995) and reliable estimates of general cognitive/intellectual ability or "g" (Brody, 1992; Raven, Raven, & Court, 1991). Studies have shown that IQ indices on matrix analogy tests are highly correlated with the IQ scores of the Wechsler scales (Desai, 1955; Hall, 1957; Levine & Iscoe, 1954; Watson & Klett, 1974). Research also demonstrates that, in general, matrix analogy tasks correlated higher with performance subtests than with verbal subtests of the Wechsler intelligence scales. In addition, matrix reasoning tasks are considered to be relatively culture-fair and language-free, requiring no hand manipulation and having no time limits. These features make it an appealing measure of PIQ, particularly with older adults and minorities. Such a measure also allows for contrasts with other nonverbal reasoning tasks, such as Block Design. When performance on Block Design is low, for example, the hypothesis that a person's score may have been affected because he or she responds slowly on a timed test can be evaluated by comparison with an untimed reasoning test. Such contrasts allow for more meaningful interpretation of test scores and performance.

The Matrix Reasoning subtest was developed after careful theory and content review of the existing literature. It contains 26 items: 3 basal items and 23 regular items. Four types of items were designed to provide a reliable measure of visual information-processing and abstract reasoning skills. These four types of matrices are continuous and discrete pattern completion, classification, analogy reasoning, and serial reasoning. They are commonly seen in existing matrix-analogy tasks

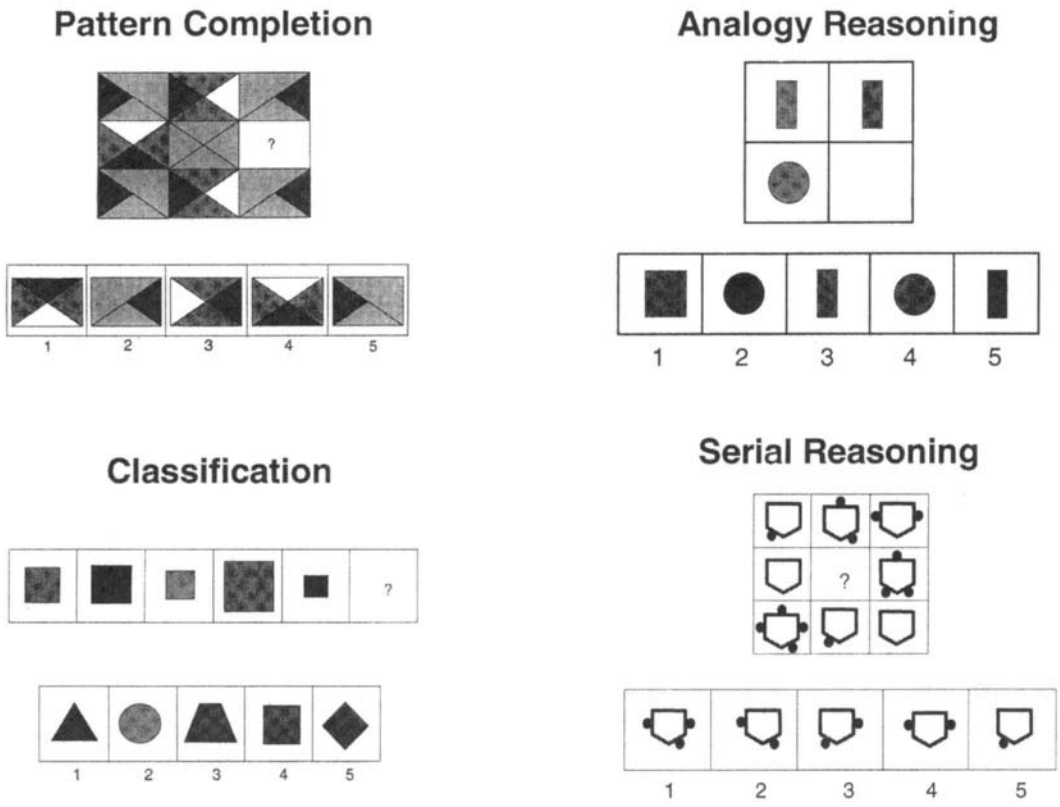


Figure 5.1. WAIS-III Matrix Reasoning Simulated Item

such as Raven's (1976) *Standard Progressive Matrices* and Cattell's (1973) *Culture Fair.* Figure 5.1 provides some examples of each type of item.

In addition to the type of matrices included in the subtest, content coverage was also influenced by two other dimensions. The first dimension includes the features and types of stimuli that can be manipulated during the problem-solving process. Attributes of the stimulus, such as color, pattern, shape, size, position, direction, and the number of attributes included in an item, were manipulated or controlled for each item. The second dimension involves the mental tasks performed during the problem-solving process, such as folding, rotating, mirroring, switching, cutting, adding, and flipping. A number of these tasks were carefully selected for each item. A progression of difficulty was developed by adding more stimuli and mental tasks from these two dimensions. The test format is multiple choice. For each item, the four foils among the five-choice answers were very carefully designed to enhance item-discrimination ability. There is no time limit for this test, but data from the WAIS-III standardization suggest that most individuals will provide answers within 10 to 30 seconds.

The reliability coefficients across the different age groups range from .84 to .94, with an average of .90, which is much higher than the Object Assembly subtest (.70) that it replaces. Moreover, the Matrix-Reasoning subtest minimizes speed and motor responses, and for the majority of examinees in the standardization sample, it takes less time to complete than the Object Assembly subtest, thus reducing overall test-administration time with most examinees. Data analysis indicated that the Matrix Reasoning subtest correlates the highest with Block Design (.60), and in factor analysis, loads on a factor made up by subtests measuring Perceptual Organization. Results of two validity

studies using samples of 26 nonclinical adults and 22 adults with schizophrenia found that the WAIS-III Matrix Reasoning subtest correlates at .81 and .79 with Raven's Progressive Matrices, respectively.

There is a legitimate concern that, because this subtest is untimed, there is a potential for the administration time of this subtest to become quite lengthy. However, the benefits of having a performance subtest measuring abstract, fluid ability independent of time outweigh the potential problems. As mentioned previously, examiners now have a subtest that can be contrasted to the other WAIS-III subtests (e.g., Block Design) that place a high emphasis on timing and bonus points. Moreover, Tulsky and Chen (1998) using the WAIS-III standardization sample estimated t that examinees tend to complete the Matrix Reasoning subtest quickly, generally in seven minutes or less. These estimates indicate that the median time for the subtest is 6.4 minutes, with 90 percent of examinees completing the subtest in 11.9 minutes. Comparatively, the estimated median time to complete Object Assembly is 10.7 minutes. Therefore, when contrasted with the Object Assembly subtest that it replaces, Matrix Reasoning is much shorter.

At the item level, the data show a similar trend. Almost 75 percent of the items were completed within 15 seconds and more than 90 percent were completed within 30 seconds. This supports the theory that, in general, examinees will respond quickly to these items. Occasionally, however, there will be individuals who take longer to answer the items. Based upon the data obtained from the 2,450 examinees who completed the standardization sample, it seems that additional time will not increase scores. Of those examinees who took longer than 60 seconds per item, the responses were wrong two-thirds of the time. This rate would be higher if the guessing factor was considered. This finding can be used to help guide examiners when administering the test. If an examinee has performed quite well on the scale and then takes additional time to solve the items as difficulty increases, the examiner should grant such leeway. Alternatively, if the examinee has low and inhibited output and tends to ruminate on items without any perceived benefit, the examiner should encourage the examinee to respond after 30 seconds or so, and definitely move him or her along after 45 to 60 seconds.

Symbol Search

The WAIS-III Symbol-Search subtest is designed to measure an individual's speed at processing new information. In this task, the examinee is presented with a series of paired groups, each pair consisting of a target group and a search group. The examinee's task is to decide whether either of the target symbols is in the search group, a group of five search-symbols.

A similar task was developed and included in the WISC-III as a supplemental subtest contributing to the 4th factor, Processing Speed (Kaufman, 1991; Wechsler, 1991; Roid, Prifitera, & Weiss, 1993; Carroll, 1993; Kamphaus, Bension, Hutchinson, & Platt, 1994). The purpose of including this subtest in the WAIS-III is to enhance the measure of processing speed of the instrument and to bring out the four-factor structure that was found on the WISC-III.

During the development of the WAIS-III Symbol Search, the following guidelines were used. First, to minimize the potential involvement of verbal encoding, only nonsense symbols were used. Second, because some nonsense symbols can be verbally coded more easily than others, the difficulty of each item was carefully evaluated across all age groups to make sure that there were no significant differences in difficulty across all items. Third, since the tasks of Symbol Search are mainly visual discrimination and visuo-perceptual scanning (Sattler, 1992), the difficulty of the test items affects the factor-loading of this subtest. If the items are too difficult, the test will tend to load more on the perceptual organization or working-memory factors rather than the speed-of-information processing factor. Therefore, the range of item difficulty is set at .80–1.00.

The test-retest reliability is .79 for the overall test-retest sample (n = 394), with a range from .74 to .82. Factor analysis suggests that the WAIS-III Symbol Search, along with Digit Symbol loads highest on the Processing Speed Index. Correlation analysis also suggests that this test correlates the highest with Digit Symbol-coding (.65).

In the WAIS-R, the Digit Symbol-coding subtest contributed the most unique variance to the scale. With the addition of Symbol Search in WAIS-III, a new dimension of functioning can now be measured. This new area of functioning appears to be sensitive to a variety of clinical conditions, such as Parkinson's Disease, Huntington's Disease, and Learning Disabilities (to name a few). Also,

Table 5.3. Reliability Coefficients of Object Assembly (OA) and Matrix Reasoning (MR)

Subtest	16–17	18–19	20–24	25–29	30–34	35–44	45–54	55–64	65–69	70–74	75–79	80–84	85–89	Average
OA	.73	.70	.73	.71	.75	.71	.78	.72	.77	.68	.59	.64	.50	.70
MR	.87	.89	.88	.91	.88	.91	.89	.93	.94	.91	.90	.89	.84	.90

Note: From *WAIS-III–WMS–III Technical Manual*. Copyright 1997 by The Psychological Corporation. Reproduced by permission. All rights reserved.

because Symbol Search requires less motor skill than Digit Symbol-coding, contrasting the two subtests can provide useful clinical information on the extent of motor involvement on low scores.

Letter-Number Sequencing

This is a new subtest designed to measure working memory. It was based on the work of James Gold and his colleagues at the University of Maryland (Gold, Carpenter, Randolph, Goldberg, & Weinberger, 1997). In this test, participants were presented with a mixed list of numbers and letters. Their task is to repeat the list by saying the numbers first in ascending order and then the letters in alphabetical order.

The reliability coefficients of the subtest are fairly good, ranging from .75 to .88, with an average of .79. Data-analysis results suggested that this test correlates the highest with other working memory measures, .55 with Arithmetic, and .57 with Digit Span. Factor analysis suggested that it loads substantially on working memory, together with Arithmetic and Digit Span.

The Content of the IQ scores

In developing the IQ scores, the decision to make Object Assembly optional may be considered problematic and controversial for several reasons. Object Assembly has been a core subtest on the Wechsler scales since their inception. Therefore, many clinicians are familiar with the performance on this subtest in various clinical populations. Also, years of research on previous Wechsler editions provide empirical support for use and interpretation of this subtest. Matrix Reasoning does not have this historical and empirical base within the Wechsler clinical and research literature.

Matrix Reasoning was designed to help assess nonverbal, fluid reasoning in an untimed manner. When work on the WAIS-III began, Matrix Reasoning was going to be an optional subtest that the examiner could use if he or she had questions about an individual's nonverbal ability and wanted to measure it independently from speeded or timed tasks or both.

The decision to replace Object Assembly with Matrix Reasoning in the IQ scores was instituted for a number of reasons. First, the statistical properties of Matrix Reasoning are far superior to those of Object Assembly. As shown in Table 5.3, the reliability of Matrix Reasoning is much higher than the reliability of Object Assembly. For Matrix Reasoning, the average of the split-half reliability coefficients is .90, which is significantly higher than the .70 average of the coefficients that was obtained on the Object Assembly subtest. Substituting Object Assembly with Matrix Reasoning allows for a smaller standard error of measurement and tighter confidence intervals in the determination of the Performance IQ. Furthermore, as can be seen in Table 5.3, the reliability coefficients of Object Assembly for adults 75 years or older are fairly low, making the measurement error too high to obtain a valid assessment of skills. This low reliability may be due, in part, to the incorporation of bonus points for quick performance on the Object Assembly subtest. Older adults generally perform at a slower pace, and this fact alone makes the Object Assembly subtest more problematic. Also, in the majority of cases, the administration time for Matrix Reasoning is less than the time needed for Object Assembly.

In deciding to replace Object Assembly with Matrix Reasoning on the IQ scores, however, a series of analyses were performed to determine if such a replacement would affect the nature of the Performance IQ scale. In one case, two "alternate" Performance sums of scaled (PSS) scores were developed and compared. The first score (PSS_1) was developed by summing scores on the following subtests: Picture Completion, Block Design, Picture Arrangement, Digit Symbol-Coding, and Object Assembly.

The second score (PSS_2) was the sum of scores on Picture Completion, Block Design, Picture

Arrangement, Digit Symbol-Coding, and Matrix Reasoning. Both summed scores were then converted to a Wechsler score with a mean of 100 and a standard deviation of 15.

Differences between these two perceptual sums of scaled scores indicated that 14 (out of the 2,450) individuals (or 1.7 percent) had difference scores of more than 0.67 SD (standard deviation) units. More important was the question of how many individuals would have fallen outside of the confidence interval of the PSS_1 score. Only 2 people out of the 2,450 examinees would have had a difference that significant. This indicated that there was not too great a change in the composite scores. Providing additional evidence that such a change would improve the IQ scores, the reliability of PSS_2 was slightly higher (r=.94 for PSS_2 versus r=.93 for PSS_1) and the standard error of estimate is slightly smaller (SEE [standard error of the estimate]=3.41 for PSS_2 versus SEE=3.79 for PSS_1).

The Domains of Intelligence: From Factor Analytic Studies to the Development of Index Scores

Background

Wechsler believed that his intelligence scales measure two domains, Verbal IQ and Performance IQ. However, evidence began to accrue after the release of the Wechsler-Bellevue that the scales could be broken down even further. For example, Balinsky (1941) performed the first factor analysis on the scale and suggested that there might be three distinct factors. In the 1950s, factor analytic studies reaffirmed this notion, demonstrating that there was at least one additional domain of functioning that was distinct from Verbal and Performance domains (see Cohen, 1952a, 1952b, 1957a, 1957b). This work showed that a third, small, yet discrete factor seemed related to the Digit Span, Arithmetic, and possibly, the Digit Symbol subtests. Though it had been given different names, Jacob Cohen's label, "Freedom from Distractibility" (Cohen, 1952a, 1952b), became the dominant label. This was due, in part, to the use of this label by Alan Kaufman in his initial interpretive book, *Intelligent Testing with the WISC-R* (Kaufman, 1979) and later by the inclusion of this label in the WISC-III factor-index scores (Wechsler, 1991). Most important is that examiners began using such factor scores in clinical settings.

The developers of the WISC-III sought to enhance the measurement of this additional factor and developed a new subtest, Symbol Search (Prifitera, Weiss, & Saklofske, 1998; Wechsler, 1991). Surprisingly, they found that this new subtest was more related to Coding, not Arithmetic or Digit Span. In factor-analytic studies, they found that four factors seemed to emerge from the analyses and they labeled the new domains of functioning Freedom From Distractibility and Processing Speed. The interpretation of a four-factor solution is not without controversy (see Sattler, 1992, for criticism of the four-factor model). Nevertheless, the four-factor model has been replicated in additional studies (Roid & Worrall, 1996; Blaha & Wallbrown, 1996; Donders, 1997). Furthermore, the additional factors of Processing Speed and Attention seem clinically relevant and are psychologically meaningful (Prifitera & Dersh, 1993; Kaufman, 1994).

Naming the Third Factor

Following the work of Cohen (1952a, 1952b) and Kaufman (1979) the third factor continued to be called Freedom From Distractibility) in the WISC-III manual, and normative information was provided for this factor (Wechsler, 1991; Roid et al., 1993). Cohen's original term Freedom From Distractibility had become entrenched in the psychological community.

In the revision of his classic text, Kaufman (1994, p. 212) criticized this name and wrote that it was a mistake not to "split with tradition" and change the label of this factor years ago. He also pointed out that this factor should have been called "by a proper cognitive name" when The Psychological Corporation published the WISC-III. His criticism appears valid; years of research indicate that the 3rd factor is more than distractibility alone (Wielkiewicz, 1990). Also, this label may lead to improper interpretation of this factor as diagnostic for attention problems, which is not necessarily the case.

Several other WISC-R researchers have echoed this concern. Some have directly stated that labeling this 3rd factor as Freedom From Distractibility is an oversimplification (Stewart & Moely, 1983; Owenby & Matthews, 1985). In a review paper, Wielkiewicz (1990) concluded that low scores on this factor of the WISC-R are not diagnostic of any

single childhood disorder and he argues against this traditional label in favor of either short-term or working memory. Other researchers, while not directly critical of the label, have demonstrated that the Digit Span and Arithmetic subtests of the WAIS-R are related to the Attention and Concentration subtests of the WMS (Larrabee, Kane, & Schuck, 1983) or to other independent measures of attention (Sherman, Strauss, Spellacy, & Hunter, 1996).

These studies and reviews seem to suggest that the subtests that make up this factor (e.g., Arithmetic, Digit Span, and sometimes, Digit Symbol) make up a higher-order cognitive ability, such as working memory. The name Freedom From Distractibility is really a misnomer that implies that this 3rd factor is nothing more than a WISC-III or WAIS-R validity measure, used solely to test a hypothesis that an obtained score underestimates an individual's "true" score. It may also imply that it is a direct measure of attention disorders, which is an oversimplification.

Attention, an alternate label, is made up of several high-level functions like focusing, encoding, sustaining, and shifting (Mirsky,1989; Mirsky, Anthony, Duncan, Ahearn, & Kellam, 1991), and selective attention is an even more complicated system that involves the selection of some stimuli for higher levels of processing as well as the inhibition of other signals for those high levels of processing (Posner, 1988).

Working Memory, still another label, involves the storage of information, the manipulation of information, and the storage of products. It requires individuals to track multiple tasks while actively processing information (Baddeley, 1986). Digit Span and Arithmetic tasks have been considered tasks involving working memory (Sternberg, 1993; Kyllonen & Christal, 1987). For the WAIS-III, the label Working Memory was adopted because it is conceptualized as a key process in the acquisition of new information.

Working Memory

Working memory is a term that denotes a person's processing capacity. The concept of working memory has replaced (or updated) the concept of short-term memory. Newell and his colleagues coined the term "working memory" and conceptualized it as a "computational workspace" (Newell, 1973; Newell & Simon, 1972). They viewed this "workspace" as being a "more active part of the human processing system" as opposed to the traditional term, short-term memory, that is the passive storage buffer. Hence, the concepts of working memory and short-term memory are similar because both have been thought of as a place where incoming information is stored temporarily and both are limited in capacity. However, the two concepts differ in one key aspect: short term memory is a "passive" form of memory and working memory is an "active" form. Traditional short-term memory is thought of as a passive storage area for information while it either becomes encoded into long term-memory or is forgotten. Working memory, on the other hand, is an area where incoming information is stored temporarily. It is also the place where calculations and transformation-processing occurs. Furthermore, as Baddeley and Hitch (1974) point out, this component also stores the products or output of these calculations and transformations (as well as the original information).

For the WAIS-III, the definition advanced by Kyllonen & Christal (1987) was employed. Working memory can be defined as the portion of memory that is in a highly active and accessible state whenever information is being processed. This includes the memory that is involved when an individual is simply attending to information (Kyllonen & Christal, 1987).

Recent literature has suggested that working memory is a key component to learning (Kyllonen, 1987; Kyllonen & Christal, 1989; Kyllonen & Christal, 1990; Woltz, 1988). Individuals with greater working memory will be capable of processing and encoding more material than individuals with a smaller working-memory capacity, thus accounting for individual differences in attention and learning capacities. Some cognitive psychologists have come to believe that working memory is an important predictor of individual differences in learning, ability, and fluid reasoning (Sternberg, 1993; Kyllonen & Christal, 1989).

The measurement of working memory dates back to early experiments conducted by Baddely and Hitch (1974). Traditionally, this construct has been measured by presenting a large amount of information (which the person has to retain in memory), requiring the person to first process (or transform) this information and then to retain the end product. Tasks tend to increase in complexity (e.g., the system is more likely to become "overloaded") as the test progresses. Individual differ-

Table 5.4. WAIS-III Exploratory Factor Pattern Loadings for Four-Factor Solutions, Overall Standardization Sample

	VERBAL COMPREHENSION	PERCEPTUAL ORGANIZATION	WORKING MEMORY	PROCESSING SPEED
Vocabulary	0.89	−0.10	0.05	0.06
Similarities	0.76	0.10	−0.03	0.03
Information	0.81	0.03	0.06	−0.04
Comprehension	0.80	0.07	−0.01	−0.03
Picture Completion	0.10	0.56	−0.13	0.17
Block Design	−0.02	0.71	0.04	0.03
Matrix Reasoning	0.05	0.61	0.21	−0.09
Picture Arrangement	0.27	0.47	−0.09	0.06
Arithmetic	0.22	0.15	0.51	−0.04
Digit Span	0.00	−0.06	0.71	0.03
Letter-Number Sequencing	0.01	0.02	0.62	0.13
Digit Symbol-Coding	0.02	−0.03	0.08	0.68
Symbol Search	−0.01	0.16	0.07	0.63

Note: From *WAIS-III–WMS–III Technical Manual*. Copyright 1997 by The Psychological Corporation. Reproduced by permission. All rights reserved.

ences become apparent when the number of errors that the person makes are tallied and analyzed. The fewer the errors (especially as the tasks increase in complexity), the greater the working memory.

As described in the *WAIS-III–WMS-III Technical Manual* (The Psychological Corporation, 1997), the 3rd-factor score is conceptualized as one that seems to tap a dimension of cognitive processing that is more than a simple validity measure. Actually "true" executive processes such as working memory or attention seem to be more relevant. Significantly, these processes would help determine how much an individual can process, and ultimately, learn. With this conceptualization, the working-memory factor took on a far more important role in guiding development efforts of the WAIS-III.

Factor Analysis of the WAIS-III

To determine the factor structure of the WAIS-III, several exploratory analyses were conducted in different ways, using different data sets, using subsets of the data, using different sets of variables, and using different extraction techniques and rotational techniques. Overall, the primary factor loading for each subtest remained relatively consistent from analyses to analyses. A few of these key analyses are reported in the *WAIS-III–WMS-III Technical Manual* (The Psychological Corporation, 1997) and one of the analyses is reprinted in Table 5.4. Fairly consistently, the Vocabulary, Similarities, Information, and Comprehension subtests all had their highest loading on one factor (called the Verbal Comprehension Index); the Picture Completion, Block Design, and Matrix Reasoning, subtests had the highest loading on a different factor (called the Perceptual Organization Index[2]); the Arithmetic[3] Digit Span, and Letter-Number Sequencing subtests had the highest loading on a third factor (called the Working Memory Index); and Digit Symbol-Coding and Symbol Search had the highest loading on a fourth factor (called the Processing Speed Index).

To test the stability of these results and the appropriateness of this factor structure in different ethnic groups, the exploratory analyses were conducted separately by ethnic group (AfricanAmerican, Hispanic, and white) using the standardization sample. The sample sizes for the three analyses were: African-American examinees, $N = 279$; Hispanic examinees, $N = 181$; white examinees, $N = 1925$. The exploratory factor-pattern loadings are listed in Table 5.5 for the African-American, Table 5.6 for Hispanic, and Table 5.7 for white examinees. For all three groups, the results are similar to those presented in the *WAIS-III–WMS-III Technical Manual*. Although there are a couple of variables with split loadings (e.g., Arithmetic is split between Verbal Comprehension and Working Memory for the group of African-American examinees and between Verbal Comprehension and Perceptual Organization for the group of Hispanic examinees), the patterns are extremely similar between these groups.

Table 5.5. WAIS-III Exploratory Factor-Pattern Loadings for Four-Factor Solutions, African-American Examinees[a]

	VERBAL COMPREHENSION	PERCEPTUAL ORGANIZATION	WORKING MEMORY	PROCESSING SPEED
Vocabulary	0.85	–0.01	0.03	0.09
Similarities	0.75	0.05	–0.03	0.11
Information	0.82	–0.01	0.09	–0.07
Comprehension	0.77	0.09	–0.06	–0.01
Picture Completion	0.02	0.57	–0.09	0.13
Block Design	0.05	0.52	0.15	0.12
Matrix Reasoning	–0.02	0.56	0.34	–0.03
Picture Arrangement	0.22	0.51	0.00	–0.02
Arithmetic	0.38	–0.02	0.51	0.09
Digit Span	0.05	0.06	0.67	0.00
Letter–Number Sequencing	–0.02	0.07	0.68	0.22
Digit Symbol–Coding	0.05	0.01	–0.01	0.76
Symbol Search	0.01	0.12	0.13	0.66

[a]$N = 279$

Table 5.6. WAIS-III Exploratory Factor-Pattern Loadings for Four--Factor Solutions, Hispanic Examinees[a]

	VERBAL COMPREHENSION	PERCEPTUAL ORGANIZATION	WORKING MEMORY	PROCESSING SPEED
Vocabulary	0.80	–0.15	0.26	0.08
Similarities	0.73	0.04	–0.01	0.07
Information	0.76	0.11	–0.07	–0.01
Comprehension	0.70	0.14	0.02	0.03
Picture Completion	0.17	0.39	–0.01	0.10
Block Design	–0.04	0.72	0.11	0.01
Matrix Reasoning	0.14	0.54	0.05	0.14
Picture Arrangement	0.36	0.37	0.00	0.04
Arithmetic	0.25	0.31	0.32	0.10
Digit Span	–0.09	0.11	0.67	0.10
Letter–Number Sequencing	0.18	0.02	0.55	–0.04
Digit Symbol–Coding	0.01	–0.01	–0.07	0.81
Symbol Search	0.00	0.03	0.14	0.73

[a]$N = 181$

Determining the Number of Subtests to Include on Each Index

Only three subtests were included in the Verbal Comprehension and Perceptual Organization Indexes, leaving the Comprehension and Picture Arrangement subtests as supplemental to the Index scores. The decision to exclude the Comprehension subtest on the Verbal Comprehension Index and the Picture Arrangement subtest on the Perceptual Organization Index was based on practical and empirical considerations. Practically, by not including these two subtests, the examiner can save a significant amount of time. Both Picture Arrangement and Comprehension can be lengthy subtests to administer and if reliable data can be obtained from fewer subtests, the clinician can save time and possibly administer other tests to answer specific clinical hypotheses. Also, by including a maximum of three subtests on each index, the four indexes are more balanced and equally weighted. Furthermore, Picture Arrangement tends to have split loadings between the Verbal and Performance factors, so it is a less "pure" task of perceptual organization than the other scales.

Empirical evidence indicated that there was some redundancy between the subtests and that the overall VCI and POI index scores typically do not

Table 5.7. WAIS-III Exploratory Factor-Pattern Loadings for Four-Factor Solutions, White Examinees[a]

	VERBAL COMPREHENSION	PERCEPTUAL ORGANIZATION	WORKING MEMORY	PROCESSING SPEED
Vocabulary	0.92	−0.11	0.04	0.05
Similarities	0.74	0.09	−0.01	0.03
Information	0.81	0.02	0.07	−0.03
Comprehension	0.80	0.06	−0.01	−0.02
Picture Completion	0.12	0.49	−0.10	0.21
Block Design	−0.03	0.68	0.08	0.07
Matrix Reasoning	0.07	0.60	0.22	−0.06
Picture Arrangement	0.27	0.41	−0.03	0.09
Arithmetic	0.22	0.20	0.47	−0.01
Digit Span	0.02	0.01	0.66	0.04
Letter-Number Sequencing	0.01	0.01	0.63	0.10
Digit Symbol-Coding	0.04	−0.05	0.07	0.70
Symbol Search	−0.02	0.15	0.06	0.69

Note: Data and Table Copyright 1998 by The Psychological Corporation. All rights reserved.
[a]N = 1, 925

Table 5.8. *R*-Squre of Different Combinations of Verbal Comprehension Subtests

NUMBER OF INDEPENDENT VARIABLES	R-SQUARED	SUBTESTS			
1	0.84	VOC			
1	0.79	INF			
1	0.78	SIM			
1	0.78	COM			
2	0.93	VOC	COM		
2	0.93	VOC	SIM		
2	0.92	VOC	INF		
2	0.92	SIM	INF		
2	0.92	INF	COM		
2	0.92	SIM	COM		
3	0.98	SIM	INF	COM	
3	0.97	VOC	SIM	COM	
3	0.97	VOC	SIM	INF	
3	0.97	VOC	INF	COM	
4	1.00	VOC	SIM	INF	COM

Note: VOC = Vacabulary; INF = Information; SIM = Similarities; COM = Comprehension.
Data and Table Copyright 1998 by The Psychological Corporation. All rights reserved.

differ if either three or four subtests are included. To determine how much incremental validity is lost by omitting a subtest, a procedure similar to that employed by Glenn Smith and his colleagues at the Mayo Clinic as part of the Mayo Older Adult Normative Study (Smith et al., 1994) was used in developing the WAIS-III.

The first step was to obtain a sum of scaled scores for each examinee in the WAIS-III standardization sample on all of the subtests that loaded on the Verbal Comprehension Index (VCI) and another sum of scaled scores for those that loaded on the Perceptual Organization Index (POI). For the VCI, Vocabulary, Similarities, Information, and Comprehension were summed. For the POI, Picture Completion, Block Design, Matrix Reasoning, and Picture Arrangement were summed. Then, in a series of separate regression analyses, these two total scores were "predicted" by using their part scores. For example, the Verbal

Table 5.9. *R*-square of Different Combinations of Performance Subtests

NUMBER OF INDEPENDENT	R-SQUARED	SUBTESTS			
1	0.67	BD			
1	0.66	MR			
1	0.61	PA			
1	0.61	PC			
2	0.86	BD	PA		
2	0.85	MR	PC		
2	0.84	MR	PA		
2	0.84	BD	PC		
2	0.83	BD	MR		
2	0.82	PC	PA		
3	0.95	MR	PC	PA	
3	0.95	BD	PC	PA	
3	0.94	BD	MR	PA	
3	0.93	BD	MR	PC	
4	1.00	BD	MR	PC	PA

Note: BD = Block Design; MR = Matrix Reasoning; PA = Picture Arrangement; PC = Picture Completion.

sum was predicted in 15 different analyses. The sums of scaled scores served as the dependent variable in regression analysis using different subsets of subtests as the independent variables. The first wave examined how well a single subtests (e.g., Vocabulary) could predict the total score. The second and third waves examined how well, two of the four subtests (e.g., Vocabulary and Similarities), or three of the four subtests (e.g., Vocabulary, Similarities, and Information) could predict the total scores. As reported in Table 5.8, approximately 97 percent of the variance of the sum of scaled scores can be accounted for by including three of the four Verbal Comprehnesion subtests, and as Table 5.9 shows, approximately 93 or 94 percent of the variance of the sum of scaled scores could be accounted for with three of the four Perceptual Organization subtests. The results suggest that any of the four Verbal Comprehnsion scales and any of the four Perceptual Organization scales could have been reported from the index with approximately the same results. Comprehension and Picture Arrangement were the logical subtests to omit because of the length of time needed to administer each of them and, in the case of the latter subtest, the split loadings obtained between the verbal-comprehension and perceptual-organization factors.

The next step was to analyze whether these "shortened" indexes would perform roughly the same as the indexes that consisted of all four subtests in a sample of "normally functioning" adults. Again, the 2,450 examinees from the standardization sample were used for these analyses. To test the effect of omitting Comprehension, two sums of scaled scores were obtained: one by summing four subtests (Vocabulary, Similarities, Information, and Comprehension) and the other by summing three subtests (Vocabulary, Similarities, and Information). Both of these sums of scaled scores were transformed to standardized scales with a mean of 100 and a standard deviation of 15.

Significant differences ($p < .05$) between these two verbal-comprehension sums of scaled scores were obtained, and only 31 of the 2,450 examinees had differences of more than 0.5 SD units. Moreover, only three of the 2,450 examinees had scores on the three-subtest sum of scaled scores that "fell" outside of the 90 percent confidence interval of the index scores that were based on four subtests. The standard error of estimate and the reliability of these two sums of scaled scores were roughly identical.

For the two perceptual-organization sums of scaled scores, similar procedures were performed with similar results. For the perceptual-organiation subtests, two sums of scaled scores were created (one by summing Picture Completion, Block Design, Matrix Reasoning, and Picture Arrangement, and the other by omitting Picture Arrangement and summing the three subtests). As before,

these scores were standardized and then transformed to standardized scales with a mean of 100 and a standard deviation of 15.

Differences between these two perceptual sums of scaled scores indicated that 41 of the 2,450 individuals had difference scores of more than 0.67 SD units. In terms of examining how many of these individuals would have fallen outside of the confidence interval of the index score based on four subtests, only 13 of the 2,450 would have had a difference that was significant. As with the verbal sums of scaled scores, there was not a significant change in the standard error of estimate or the reliabilities of these two scores.

These results supported the conclusion that, for the vast majority of examinees, there would not be significant differences between their index scores based on three subtests and index scores based on their longer counterparts. This is not to say that it is not valuable to administer the two additional subtests. Certainly, it is more desirable to obtain the additional information provided by the Comprehension and Picture Arrangement subtests. This is especially true if there were significant and unusually large differences between Comprehension and the other Verbal Comprehension subtests or between Picture Arrangement and the other Perceptual Organization subtests. Data analysis suggests that, for example, in 3.4 percent of the standardization sample, the Comprehension subtest was at least three points lower than the mean of the verbal subtests[4] and in 4 percent of cases, it was at least three points higher than the verbal mean. Similarly, in 6.4 percent of the standardization sample, the Picture Arrangement subtest was at least three points lower than the mean of the performance subtests[5] and in 8.1% of cases, it was at least three points higher than the mean of the performance subtests. So, by keeping these subtests out, one might miss important information about the relative strengths and weaknesses of some individuals.

Nevertheless, the time required to administer these two additional subtests may not justify the additional information obtained in the majority of cases. Hence, it was decided to construct the Index scores the way they were. Strengths and weaknesses on the Comprehension and Picture Arrangement subtests could always be obtained through profile analysis.

Technical Characteristics of the WAIS-III

Changes in Normative Information

The WAIS-III standardization sample contains 2,450 adults, and covers an age range from 16 to 89 years of age. Extending the upper age to 89 years to adjust for the longer life span of the U.S. population is a significant improvement to the scale. The normative sample was stratified on several demographic variables (e.g., sex, ethnicity, educational level, and region of the country) using the newest census data. The WAIS-III sample includes

Table 5.10. Demographic Characteristics of the WAIS-III Standardization Sample: Percentages by Age and Occupational Level

OCCUPATION	AGE 16–19	20–24	25–34	35–44	45–54	55–64	65+	TOTAL PERCENT
Executive	0.0	0.0	0.3	0.0	1.1	2.3	0.6	0.5
Manager	0.3	7.3	4.1	6.0	6.1	4.0	1.6	3.2
Supervisor	1.2	2.8	4.9	3.8	2.8	3.4	.3	2.2
Professional or Tech Specialist	2.9	12.8	21.6	28.8	19.0	11.3	4.8	11.8
Marketing or sales	4.6	10.6	7.2	8.1	15.6	9.6	3.2	6.8
Administrative support and clerical specialist	4.4	12.3	13.5	15.2	16.2	8.0	3.8	8.6
Farming, Forestry, Fishing, & Related	0.3	0.6	0.0	0.5	0.5	0.0	0.6	0.4
Precision Production, Craft, & Repair	0.9	1.1	6.1	8.2	2.8	4.0	1.4	3.0
Operator, Fabricator, & Laborer	10.5	18.4	17.1	13.6	11.2	8.5	3.0	10.0
Homemaker, Retired, Not in Labor Force	74.1	29	11.1	4.9	14.6	6.8	5.6	20.5
Total Percentage	100	100	100	100	100	100	100	100
N	343	179	346	184	179	176	690	2097

Table 5.11. Demographic Characteristics of the U.S. Population: Percentages by Age and Occupational Level

Total Labor Force Statistics, 1996
(Numbers in thousands)

OCCUPATION	AGE 16–19	20–24	25–34	35–44	45–54	55–64	65+	TOTAL PERCENT
Executive	0.0	0.0	0.2	0.4	0.7	0.4	0.1	0.3
Manager	0.5	3.0	7.5	10.0	11.0	7.7	2.1	6.9
Supervisor	1.1	3.0	5.2	6.1	5.9	3.9	0.7	4.2
Professional or Tech Specialist	1.7	9.8	18.1	18.9	19.0	11.1	2.0	13.3
Marketing or sales	10.3	8.4	6.0	5.1	5.4	4.6	1.4	5.4
Administrative support and clerical specialist	19.0	24.5	20.5	19.0	18.1	14.2	3.2	16.7
Farming, Forestry, Fishing, & Related	2.1	1.6	1.3	1.0	0.8	0.8	0.2	1.0
Precision Production, Craft, and Repair	2.0	5.8	7.7	8.1	6.6	4.6	0.6	5.6
Operator, Fabricator, & Laborer	14.5	21.0	17.8	16.6	14.9	11.5	1.9	14.0
Homemaker, Retired, Not in Labor Force	48.9	22.9	15.6	14.8	17.5	41.2	87.7	32.7
Total Percentage	100	100	100	100	100	100	100	100
N	14,350	17,317	40,486	43,445	32,477	21,146	31,369	200,590

Employment Status of the Civilian Noninstitutional Population, 1996
(Numbers in thousands)

CIVILIAN NONINSTITUTIONAL POPULATION	CIVILIAN LABOR FORCE				NOT IN LABOR FORCE
	TOTAL	PERCENT OF POPULATION	EMPLOYED	UNEMPLOYED	
200,590	133,943	66.8	126,708	7,236	66,647

many more older adults and minority groups than samples that had been collected for previous versions of the Wechsler adult scales. In the WAIS-R, for instance, 216 people were minorities (or "Nonwhite" as they were labeled in the WAIS-R manual) which roughly reflected the percentage of minorities in the U.S. based upon the 1970 U.S. census report (Wechsler, 1981). The number has become significantly outdated as the population of the United States has changed. Hence, the sample collected for the WAIS-III reflects the changes that have occurred in the U.S. population over the last 25 years.

Another difference between the WAIS-III and the previous editions is the exclusion of occupational status as a demographic-stratification variable. Occupational status has been replaced by educational level, which is highly correlated with occupational level, and may be used as a predictor of socio-economic status. However, some may find the occupational status of the WAIS-III standardization sample of interest; it is not reported in the WAIS-III manual but it is shown in Table 5.10. Occupation levels were grouped into 10 categories. Nine of these are the categories that have been suggested by the National Industry-Occupational Matrix of the Bureau of Labor Statistics. The remaining category included people outside of the work force (people who were retired, homemakers, and or not in the labor force). Table 5.11 lists the population figures for occupational level. These percentages were based upon data from the U.S. Department of Labor (1996). In general, the data obtained for the WAIS-III sample reflects the occupational level of the U.S. population.

Additional Normative Information

The WAIS-III provides additional normative information for optional procedures and for special clinical analysis (e.g., profile analyses and subtest scatter, IQ and factor-index discrepancy scores, memory-ability discrepancy scores, and ability-achievement discrepancy scores). To facilitate interpretation of testing results, The WAIS-III not only provides critical values for determining statistical significance of a given discrepancy, but also

the "base rate" for evaluating whether the discrepancy is clinically meaningful. Previously, this type of normative data was only available in journal articles and related literature that were published well after the test was printed. These tables are provided in the WAIS-III manual, which should be convenient for the clinician using the test.

Reliability

The overall split-half internal consistency coefficients are from .94 to .98 for IQ scales, from .88 to .96 for factor indexes, from .82 to .93 for Verbal subtests, and from .70 to .90 for Performance subtests. The test-retest stability was evaluated using a large sample containing 394 cases, and the stability coefficients were provided for four age-bands as well as for the overall sample. The overall stability coefficients are from .91 to .96 for IQ scales, from .85 to .95 for factor-index scales, from .75 to .94 for Verbal subtests, and from .69 to .86 for Performance subtests. Interrater reliability coefficients are also in the .90s-range for the three Verbal subtests (Vocabulary, Similarities, and Comprehension) that require more judgment in scoring. These reliability coefficients are either improved from or equally as good as WAIS-III predecessors.

Correlation with Other Wechsler Intelligence Scales

The WAIS-III is highly correlated and highly consistent with the WISC-III. The WAIS-III and the WISC-III measure similar constructs and produce similar results. The correlation coefficients between the WAIS-III and WISC-III IQ scores are .88, .78, and .88 for VIQ, PIQ, and FSIQ, respectively. The correlations between index scores are also very high, ranging from .74 to .87. The means of the WAIS-III IQ scores were from 0.4 to 0.7 points higher than the corresponding means of the WISC-III IQ scores. The classification consistency is 95 percent or higher when a 95 percent-confidence interval was used. Similarly, the WAIS-III is also highly correlated and consistent with the WAIS-R. The correlation coefficients between the WAIS-III and WAIS-R IQ scores are .94, .86, and .93 for VIQ, PIQ, and FSIQ, respectively. The mean WAIS-III scores are about 1.2, 4.8, and 2.9 points lower than the corresponding WAIS-R VIQ, PIQ, and FSIQ scores, respectively. This validity ensures the meaningful transition and comparison between the WAIS-III and the WISC-III or the WAIS-R.

It is important to point out that high consistency between the WAIS-III and other Wechsler intelligence scales does not mean that the majority of individuals will obtain "identical" scores across two different Wechsler intelligence scales. When examining individuals, the majority of examinees will obtain different scores across two different tests. This may be because of many factors, such as Flynn effect, practice effect, design differences between the two tests, effects of having a restricted floor or ceiling, and other psychometric factors (Bracken, 1988; Kamphaus, 1993; Zhu & Tulsky, 1997). Furthermore, there is likely to be an interaction among these factors. Therefore, clinicians should take the confidence intervals into account when comparing the testing results of WAIS-III and other Wechsler intelligence scales. Score discrepancy should be evaluated on the basis of both statistics and clinical meaningfulness. A true score discrepancy should be statistically significant and clinically meaningful (rare) (Matarazzo & Herman, 1985).

The age-overlapping between the WAIS-III and WISC-III makes it possible for test users who work with adolescents and young adults to test their clients with either tests or to compare their performance on the two tests (Sattler, 1992). A commonly asked question is: When assessing a 16-year-old, which test is more appropriate, WISC-III or WAIS-III? The answer to this question is: It depends on the ability of the 16-year-old. Because the WISC-III has better floor than the WAIS-III, its score should be more reliable for individuals with low abilities; on the other hand, because the WAIS-III has better ceiling than the WISC-III, its score is more reliable for individuals with high ability. For individuals with average ability, the testing results should be very stable across the two tests.

Clinical Group Studies

Whenever a test is revised, there is always a question of whether the patterns of scores are consistent with the scores that would have been obtained if the previous version had been used. Often, these studies are conducted by clinicians and researchers, and the results of the studies are available only in professional journals. For the

WAIS-III, more than 600 individuals who have been diagnosed with a variety of clinical conditions participated at the time of the standardization; the *WAIS-III–WMS-III Technical Manual* reports the results of these small validation studies. These conditions include: Alzheimer's disease, Huntington's disease, Parkinson's disease, traumatic brain injury, temporal lobe epilepsy, chronic alcohol abuse, Korsakoff's syndrome, schizophrenia, mental retardation, learning disability, attention-deficit hyperactivity disorder, and deaf and hearing impairment.

Similar to many previous clinical studies with its predecessors, these WAIS-III clinical studies demonstrated the clinical utility of the instrument. It is important to note, however, that the clinical studies reported in the technical manual are not conclusive and provide only initial construct validity. They should not be used to provide "normative" information about the typical functioning of individuals with these clinical conditions.

The samples used in these clinical studies may not be representative because there were not "tight" inclusion and exclusion criteria. Most of the data were collected by clinicians who had busy clinical practices, and often data on some of the groups were collected in different clinics, diagnostic centers, or hospitals. Often, the different sites used different diagnostic procedures, criteria, and data collection methods. Moreover, some samples, such as Parkinson's disease and lobectomy groups, were relatively small, which increased the likelihood of sampling error.

With those cautionary points mentioned, these studies, nonetheless, provide information to the clinician and researcher. As with many of the previous clinical studies using a Wechsler scale in similar clinical groups, the WAIS-III often replicated the clinical findings that have been published in the literature. For instance, in a clinical study, 108 adolescents and adults diagnosed with mental retardation (62 mild and 46 moderate), using DSM-IV and American Association of Mental Retardation (AAMR) criteria, were tested using the WAIS-III. The results showed that the participants exhibited relatively flat-score profiles and that 99 percent of the sample obtained IQ scores 2 to 3 SDs below the mean. These results are very consistent with previous findings by Atkinson (1992), Craft and Kronenberger (1979), and Spruill (1991) for adult participants and by Wechsler (1991) for children. Further analysis showed that roughly 83 percent of the participants in the mild group had IQ scores between 53 and 70, and that 82 percent of the examinees in the moderate group had IQ scores between 45 and 52 (Tulsky & Zhu, 1997). These results suggest that the WAIS-III not only has sensitivity in identifying individuals with cognitive functioning that is 2 SDs below the mean, but the WAIS-III also has specificity in that it can separate individuals who function at mild and moderate levels.

In another study, the WAIS-III was administered to a sample of 30 adolescents and adults diagnosed with Attention-deficit hyperactivity disorder (ADHD) according to clinical interviews, DSM-IV diagnostic criteria, and the Brown Attention-Deficit Disorder Scales (Brown, 1996). The results at the IQ score level suggested that, this sample performed similarly to the standardization sample. The mean FSIQ was at the average range and there was no significant difference between the VIQ and PIQ. When the factor-index scores were evaluated, however, marked results were found. Their mean WMI score is about 8.3 points lower than their mean VCI score, and their mean PSI score is about 7.5 points lower than their mean PSI score. About 30 percent of the sample with ADHD had WMI scores at least 1 SD lower than their VCI scores, whereas 13 percent of the WAIS-III standardization sample obtained such discrepancies. About 26 percent of the ADHD sample had PSI scores at least 1 SD lower than their POI scores whereas, 14 percent of the WAIS-III standardization sample had such discrepancies. For the difference between the higher score of either the VCI or POI and the lower score of either the WMI or PSI, 61.3 percent of the sample obtained differences of 1 SD, and 16.1 percent obtained differences of 2 SDs or more; only 30.5 percent and 3.5 percent of the WAIS-III standardization sample had such differences for the VCI- or POI-score differences, and the WMIs or PSI-score differences, respectively. These results are comparable to the findings by Brown (1996) using a larger adolescent and adult ADHD sample, and the WAIS-R. They are also very consistent with the findings by Wechsler (1991), Prifitera and Dersh (1992), and the research group lead by Biederman et al. (1993).

The study using the traumatic brain injury (TBI) sample further demonstrated the clinical utilities of the new factor-index scores. The WAIS-III was administered to 22 adults who had experienced a moderate-to-severe single-closed head injury. Consistent with the previous findings, the TBI sample exhibited some overall impairment. Their

IQ scores were all at the low-average-range and no significant differences were found between the mean VIQ and PIQ scores. When the factor scores were compared, however, the relative strengths and weaknesses were obvious. The mean PSI score (73.4) of the TBI sample was significantly lower than the POI scores (92.1) and other factor-index scores. Further analysis showed that about 77 percent of the traumatic-brain-injury sample had a PSI score that was at least 1 SD lower than their POI score, while it was only 14 percent for the standardization sample.

Although it is apparent that evaluating the factor-index scores alone is usually not conclusive for clinical diagnosis, the factor scores certainly can provide extra information that will facilitate the diagnostic processes. Understanding the strengths and weaknesses can also assist in the interpretive process and intervention planning.

Interpretive Considerations

Included in the *WAIS-III–WMS-III Technical Manual* is a chapter devoted to basic issues in interpreting WAIS-III scores. This chapter provides descriptions of IQ and Factor Indexes, suggestions for basic interpretive consideration, and procedures for discrepancy analysis. The suggestions for interpretive considerations should not be used as a "cook book" or comprehensive guideline for interpretation. Clinical interpretation is a very complicated hypothesis-testing process that varies from situation to situation. Therefore, no single approach will work for all scenarios.

Since the WAIS-III continues the tradition of the Wechsler intelligence scales, many interpretation strategies, methods, and procedures that were developed by experienced clinicians and researchers for its predecessors should still be valid and useful for interpreting its results. Test users should refer to Kaufman (1990, 1994) and Sattler (1992) for detailed introductions and discussions of these interpretation strategies, methods, and procedures. Additionally, in response to progress in the field of cognitive assessment, the WAIS-III provides new factor-index scores that measure more refined cognitive domains, and these factor indexes have proven useful and informative in clinical diagnosis (Tulsky, Zhu, & Vasquez, 1998). Clinicians should evaluate the additional information provided by these factor indexes when interpreting the traditional IQ scores.

While detailed discussion of interpretation strategies, methods, and procedures is beyond the scope of this chapter, the authors would like to suggest a few basic interpretive considerations that may help readers understand the nature of clinical interpretation with the WAIS-III.

First, testing results should never be interpreted in isolation. Instead, interpretation must be made within the context of an individual's current mental status, social environment, and life history. As suggested by the *WAIS-III–WMS-III Technical Manual*, when interpreting the WAIS-III results, clinicians should consider four broad sources of information: medical and psychological history, direct behavioral observations, quantitative test scores, and qualitative aspects of test performance.

Second, testing is different from assessment (Matarazzo, 1990; Prifitera, Weiss, & Saklofske, 1998; Robertson & Woody, 1997). Psychological testing is a data-collection process in which an individual's behaviors are sampled and observed systematically under standardized conditions. Psychological assessment is a complicated problem-solving process that usually begins with psychological testing. Therefore, obtaining some test scores is just the beginning of assessment, not the end.

Third, interpretation is the process of "making sense" out of the test results. It includes a very complicated, multi-level process where hypotheses are systematically formed and tested using test scores and other clinical information. Interpretation integrates data collected through testing (such as quantitative test scores, qualitative aspects of test performance, and direct behavioral observations) with information about a person's medical and psychosocial history and weaves them together into meaningful information. Each test score may be used as a piece of evidence supporting certain conclusions. Each piece of information is like a puzzle piece. Clinicians must first gather all puzzle pieces and then put all of them together in a meaningful way before any conclusions can be made. With this analogy in mind, it will be clear that even though identifying one puzzle piece is usually not sufficient to solve the whole puzzle, it is a necessary and important step. It is similar to the physician who measures a patient's body temperature and blood pressure as just two steps along the way in reaching a diagnosis. Temperature and blood pressure are universally-performed procedures, however, neither of them, in isolation, are conclusive for a final diagnosis. Similarly, scores

Table 5.12. An Example of Age-Corrected and Reference-Group-Based Scaled Scores from a Hypothetical 85-Year-Old Examinee's Subtests

SCORES	PC	CD	BD	MR	PA	SS	OA	V	S	A	DS	I	C	LNS
Raw Subtest	14	33	23	7	4	14	18	33	16	10	14	13	16	6
Age-Corrected Scaled	10	10	10	10	10	10	10	10	10	10	10	10	10	10
Reference-Group Scaled	5	4	6	5	4	3	6	9	7	8	8	9	8	5

Note: PC = Picture Completion; BD = Block Design; MR = Matrix Reasoning; PA = Picture Arrangement; SS = Symbol Search; OA = Object Assembly; V = Vocabulary; S = Similarities; A = Arithmetic; DS = Digit Span; I = Information; C = Comprehension; LNS = Letter-Number Sequencing.

on an intelligence test must be combined with scores on other tests, the examinee's demographic information, such as socioeconomic status, life history, educational background, and other extratest information before any clinical decision can be made.

Basic Interpretation of the WAIS-III

Wechsler Scores

The WAIS-III uses a scoring metric that will be familiar to users of other Wechsler tests. Subtest raw-scores are transformed to subtest scaled-scores with a mean of 10 and a standard deviation of 3. A subtest scaled-score of 10 indicates that the individual is performing at the average level of a given group. Scores of 7 and 13 would reflect performance that is 1 SD below and above the mean, respectively, while scaled scores of 4 and 16 would reflect performance that is 2 SDs from the mean.

The WAIS-III differs from its predecessors in that the scaled scores are now age-corrected. On the WAIS-R, a reference group of subjects (ages 20–34 years) is used to convert raw scores to scaled scores. By doing this, the subtest scores are compared to the level of performance of a relatively young reference group. Since some of the skills measured by the Wechsler adult scales decline with age, these subtest scores will reflect this decline when examinees are compared with a normative group much younger then themselves. In previous editions, the composite scores were the unit that was adjusted for age (e.g., Verbal, Performance, and Full Scale IQ scores are computed separately by age group).

In the WAIS-III, the correction for age was made at the initial transformation to scaled subtest scores. This change was made in order to prevent older subjects from receiving very low scaled-scores on some (most) of the subtests because they are being compared with examinees their own age rather than examinees who are much younger. As an example of how profound this decline can be, an example of converting raw scores to scaled scores for an 85-year-old is presented in Table 5.12. In the first line of the table, the subtest raw-scores are presented. The raw scores for this example were selected so that they corresponded with an average performance (e.g., scaled score of 10) of an 85- to 89-year-old examinee, and would emphasize the point of how different scores could look. The second row in the Table 5.12 shows the age-corrected scaled scores (e.g., 10) that correspond to the raw-score points. The "reference" group's scaled scores are presented in the third row. As shown, the perceptual subtests (e.g., Matrix Reasoning or Picture Completion) and processing-speed subtests (e.g., Symbol Search or Digit Symbol) show significantly lower scaled-scores when the individual is compared to a younger-aged reference group rather than compared 85-year-olds. In the WAIS-III, a reference group comparison (at the subtest level) can still be made, but this is now an optional procedure and these reference-based scaled scores do not feed into the formula to calculate the IQ score.

Instead, it is the age-corrected scaled scores that are summed and transformed to yield composite scores. The WAIS-III IQ and Index scores have retained the common metric of a mean of 100 and a standard deviation of 15 for evaluating level of performance. A score of 100 on any of these measures defines the average performance of individuals within the same age group. Scores of 85 and 115 correspond to 1 SD below and above the mean, respectively, whereas scores of 70 and 130 are 2 SDs below and above the mean. About 68 percent of all examinees obtain scores between 85 and 115, about 95 percent score in the 70–130 range, and

nearly all examinees obtain scores between 55 and 145 (3 SDs on either side of the mean).

Scores should be reported in terms of confidence intervals so that the actual score is evaluated in light of the reliability of that test score. Confidence intervals assist the examiner in test interpretation by delineating a range of scores in which the examinee's "true" score most likely falls, and remind the examiner that the observed score contains measurement error.

Level of Performance

The level of performance refers to the rank that is obtained by an individual in comparison to the performance by an appropriate normative group. Clinical decisions can then be made if the level of performance of the individual is significantly lower than the normative group. Alternatively to this normative approach, clinical decisions can also be made if a specific score is lower than the individual's other scores (relative weaknesses). In nonclinical settings (e.g., industrial and occupational settings), the emphasis on level of performance shifts slightly, as more weight is placed on competency and the patterns of a person's strengths and weaknesses without necessarily implying any type of impairment. As described in the *WAIS-III–WMS-III Technical Manual*, test results can be described in a manner similar to the following example:

> Relative to individuals of comparable age [or, alternatively, of a reference group of younger adults], this individual is currently functioning in the [_____ range of functioning] on a standardized measure of [IQ or Index name]. (p. 185)

IQ and index scores are estimates of overall functioning in an area that should always be evaluated. As composite scores, they should be interpreted within the context of the subtests that contribute to the overall IQ scale or index score. The IQ and index scores are much more reliable measures then the subtest scores, and, in general, these are the first scores to examine when one begins to review WAIS-III data. Sometimes, the VIQ or PIQ scores and the various index scores are discrepant from one another, indicating that the examinee has some areas of functioning that are stronger or weaker than other areas of functioning.

Alternatively, sometimes the subtests that make up the IQ and index scores are substantially different from one another. It is important to realize that when two component subtest-scores are substantially different from one another, with one unusually high and the other unusually low, it will push the index score toward the arithmetic mean and thus toward the average range. Such an average score reflects a dramatically different pattern of abilities than does an average index score obtained from two subtest scores that are both in the average range. It is common practice for examiners to closely examine profiles in an ipsative fashion (e.g., examine the subtests against the examinee's own anchor point rather than against the subtests of a norm-referenced group) to see which scores show relative strengths and which show relative weaknesses. This technique is called profile analysis.

Profile Analysis and Cluster Interpretation

In clinical practice, clinicians compare the examinee's performance on the 11 (WAIS-R) or 13 (WISC-III) subtests to see if any "patterns" emerge from which they can make inferences about an examinee. Glutting, McDermott, & Konold (1997), reported that there are more than 75 different patterns of subtest variation. Some have suggested various ipsative analyses (see Kaufman, 1994; Sattler, 1992) while others have stressed using a normative approach (McDermott, Fantuzzo, Glutting, Watkins, & Baggaley, 1992; McDermott, Fantuzzo, & Glutting; 1990) to analyze patterns of scores. McDermott et al. (1992) have even used cluster-analytic techniques to develop different subtest taxonomies as an alternative to profile analyses (McDermott, Glutting, Jones, Watkins, & Kush, 1989; McDermott, Glutting, Jones, & Noonan 1989). Unfortunately, by taking a strictly normative approach, the examiner may miss some information about an individual's strengths and weaknesses. In fact, it is very common for an individual to function at different ability levels in different cognitive areas. By examining deviations from the individual's average level of functioning (e.g., significant and unusual differences between subtests and the average of all subtests), the examiner may see a pattern and generate additional hypotheses. Furthermore, in the WAIS-III, the examiner can use the frequency data of deviations from a mean that were obtained in the WAIS-III standardization study to decide how rare the obtained difference is in a nor-

mative sample and interpret such normative information within the context of other facts and life history data that the examiner has accumulated about the individual.

IQ-Score and Index-Score Discrepancies

In Wechsler's (1939) initial work on his first intelligence scale, he placed most of the emphasis on the FSIQ score and believed that an examinee's FSIQ score is always an average of the person's performance on all of the subtests (Wechsler, 1944). Nevertheless, Wechsler did realize that it was still important, at times, to view the VIQ and PIQ scores separately. He thought that this procedure would usually be reserved for the occasion of testing a person with special disabilities (Wechsler, 1944).

Since the publication of the Wechsler-Bellevue scale, the practice of interpreting VIQ-PIQ differences has become a common method of determining when to modify the interpretation of an FSIQ score and to examine the VIQ and PIQ scores separately. In the WAIS-R, Wechsler (1981) included a table to show the minimum differences between the VIQ and PIQ scores required for significance at the .15 and .05 levels of confidence for each age group. "Rules of thumb" abounded, and generally, a difference score of 12–15 points became the marker at which examiners started inferring that the examinee had a clinically relevant deficit. Matarazzo and Herman (1985) documented that the frequency of 12- to 15-point differences were much more common than examiners had previously believed and they demonstrated the need for examining statistical significance as well as clinical meaningfulness (base rates). In other words, they differentiated between a statistically significant difference (which suggests that the examinee is better at one skill than another) and a clinically meaningful difference (which indicates that the obtained-difference score is of such a high magnitude that it does not occur very frequently). This latter finding may suggest that the examinee has a true clinical deficit in an area, however, this can only be concluded after the examiner has reviewed the other variables (e.g., other test scores, psychosocial history, educational level) and found results to support such an interpretation. In addition to the VIQ- and PIQ-score differences, the WAIS-III includes normative discrepancy information on all possible pairs of index scores. A variety of detailed interpretation schemes has been suggested to explain meaningful differences (e,g., Kaufman, 1990,1994; Sattler, 1992).

The *WAIS-III–WMS-III Technical Manual* also presents frequency-of-score differences by ability level (The Psychological Corporation, 1997). Unfortunately, there is no presentation of these frequency data by other demographic information (e.g., educational level). These alternative tables may prove to be more useful because variables such as previous level of education would not be affected by a neuropsychological disorder or condition, whereas current overall ability may be lowered by the neuropsychological deficit.

Interpretation in a Neuropsychological Setting

Neuropsychology is a highly specialized approach to the understanding of individual differences (Hynd & Semrud-Clikeman, 1990). It is the measurement and analysis of the cognitive, behavioral, and emotional consequences of brain damage or dysfunction (see the neuropsychology section, this volume). Often, the WAIS-III is used to gauge the individual's current overall ability and will play a part in helping the neuropsychologist detect gross intellectual deterioration. The IQ scores generated by the Wechsler scales of intelligence are typically very sensitive to generalized impairment. However, these same IQ scores are also relatively insensitive to very focalized lesions of the brain (Matarazzo, 1972; Hynd & Semrud-Clikeman, 1990; Chelune, Ferguson, & Moehle, 1986). Instead, other tests that measure more distinct cognitive functions are used to supplement Wechsler IQ scores to detect specific deficits.

Since the emphasis of the evaluation typically focuses on specific abilities, the examiner may place more weight on the measurement of a person's ability in various functional areas than on an overall IQ score. Various researchers have identified between five and seven major functional areas, including intelligence, language, spatial or perceptual ability, sensorimotor functioning, attention, memory, emotional or adaptive functioning, psychomotor speed, and learning (see Lezak, 1995; Larrabee & Curtiss; 1995; Smith et al., 1992 for a comprehensive review). With the new WAIS-III, the index scores that break down Verbal and Performance IQs into somewhat more specified scores than those obtained by the Verbal and Performance IQ scores, should be an asset to

the neuropsychologist. Moreover, subtest-level interpretation may also be appropriate for assessing specific abilities.

Since, the neuropsychologist is attempting to detect some of the cognitive consequences of brain damage, he or she must: (a) compare an individual's current score to his or her estimated (or known) premorbid level of functioning or use demographically-corrected scores, or both, to factor in the effects attributable to various demographic variables, (b) factor in any effects due to previous testing, and (c) examine test scores to determine strengths and weaknesses between various cognitive and memory functions. The remainder of this chapter will examine these various areas of interpretation.

Predicting Premorbid Functioning. A difficult task faced by any psychologist is determining if an individual's current test scores reflect a drop in performance from the same individual's previous ability before an accident occurred or illness began (Franzen, Burgess, & Smith-Seemiller, 1997). This process can help the neuropsychologist make a determination about whether the individual has sustained loss in functioning from the accident or illness as compared with his or her previous ability. Wechsler was the first person to propose that there was a "deterioration index" that could be derived by comparing the performance on so called "hold subtests" of the Wechsler scales (e.g., those subtests' scores that were found not to decline with the age of the examinee) to the "don't hold subtests" (e.g., those subtests' scores in which performance was not expected to remain stable over time and would ultimately deteriorate with the age of the examinee) (Wechsler, 1944). However, basing the assessment of premorbid function on "hold" tests can underestimate premorbid IQ by as much as a full standard deviation (Larrabee, Largen, & Levin, 1985; Larrabee, 1998).

Alternative techniques include using scores that are obtained on vocabulary or reading tests because the skills they reflect were believed to be relatively independent of general loss of functioning and could therefore be used as an index of premorbid functioning. Yates (1956) was the first person to hypothesize that, using the WAIS Vocabulary score, one could estimate premorbid functioning because it is relatively independent of age-related declines in performance. Follow-up research by Russell (1972) and Swiercinsky & Warnock (1977) showed that individuals with brain damage do much poorer than the general population on Vocabulary, a finding that contradicted Yates' hypothesis.

The more recent focus has been on using reading tests as an indicator of premorbid functioning (Nelson, 1982; Nelson & McKenna, 1975; Nelson & O'Connell, 1978). Nelson and O'Connell (1978) introduced the National Adult Reading Test (subsequently named the New Adult Reading Test [NART]), which was a reading test using irregularly pronounced words. They developed a regression-based formula for estimating WAIS IQ scores from the scores on the NART reading test and concluded that the predictions based on NART scores are fairly accurate. Subsequent revisions have included an alternative NART for American participants (AMNART) (Grober & Sliwinski, 1991), an alternative revision of the NART for American examinees (NART-R) (Blair & Spreen; 1989), and a reading subtest from the Wide Range Achievement Test-Revised (Kareken, Gur, & Saykin, 1995).

Alternate methods of determining premorbid functioning utilize the relationship between Wechsler IQ scores and demographic variables such as age, education, sex, race, and occupational level. For a detailed discussion about the relation between demographic variables and IQ scores, see a review by Heaton, Ryan, Grant, and Matthews (1996), and studies by Heaton, Grant, and Matthews (1986), and Kaufman, McLean, and Reynolds (1988). In general, two methodologies have prevailed. Some have used the correlations to develop prediction equations (e.g., Wilson et al., 1978; Barona, Reynolds, & Chastain, 1984), while others have developed independent norms that use more focused reference-groups against which the examiner can compare scores.

Capitalizing on the high correlations between demographic variables and IQ scores, researchers began performing regression analyses on the WAIS (Wilson et al., 1978) and the WAIS-R (Barona, Reynolds, & Chastain, 1984) standardization samples in an effort to develop formulas to calculate premorbid IQ. Wilson et. al. (1978) constructed prediction formulas based on five demographic variables (age, sex, race, education, and occupation) for Full Scale, Verbal, and Performance IQs. They also found that education and race were the most powerful predictors in each equation. Once the WAIS-R was published, Barona et al. (1984) replicated this work and constructed equations consisting of the following

predictor variables: age, sex, race, geographic region of residence, occupation, and education). Sweet, Moberg, and Tovian (1990) reviewed these two prediction formulas and concluded that there was not strong support for using the Barona index over the Wilson index and that, at best, "modest success" in terms of adequate classifications may be achieved by both formulas. The authors conclude that these formulas may be useful in research or when used in conjunction with past records, but they should not be used "in isolation with individual patients" (Sweet et al., 1990; p. 44).

A variant of this approach that should be mentioned is a "combined approach" that would use scores (e.g., Vocabulary, achievement NART scores) as a concurrent measure of ability along with demographic variables to develop a "better" formula to predict premorbid IQ (Krull, Scott, & Sherer, 1995; Vanderploeg & Shinka, 1995). To support this methodology, the advocates of this technique stress that the amount of variance that is accounted for through multiple regression-analyses increases when the regression model includes a concurrent measure of reading and demographic variables as predictors. A word of caution should be offered, however. To the extent that a potential disorder or disability may affect current functioning, the correlation between a concurrent measure and premorbid ability will decrease, and such a methodology may be less accurate than prediction using only demographic information.

Heaton and his colleagues (Heaton, Grant, & Matthews, 1991; Heaton, 1992) proposed an alternate way to interpret IQ scores in light of demographic variables. They conducted a study with 553 neuropsychologically normal adults that investigated the relationship between neuropsychological-test scores and demographic characteristics (Heaton, Grant, & Matthews, 1986). Some scores were highly correlated with age; others were related to other demographic variables like educational level. Moreover, these demographics affected diagnostic accuracy of neuropsychological tests.

As a result, Heaton and his colleagues obtained new normative information on several measures that are commonly used in neuropsychology. They developed and published new normative information for the WAIS (Heaton et al., 1991) and for the WAIS-R (Heaton, 1992), corrected for age, education and sex. The neuropsychologist could then evaluate an individual's performance and compare how he or she performed relative to a person of similar age, ethnicity, background, gender, and education. This score could be compared and contrasted with the traditional IQ and the examiner would have a pretty good idea of how the average individual coming from a certain culture and age would have performed. In a separate study, Malec, Ivnik, Smith and their colleagues at the Mayo Clinic (Malec et al., 1992) developed age- and education-corrected WAIS-R scores for examinees older than 74 years of age. While this methodology is different than the others proposed above (i.e., it isn't used to predict premorbid IQ directly), the clinician uses a systematic technique to evaluate an obtained score. By comparing the overall ability score with a demographically corrected score, the examiner can judge if the score seems to reflect a deficit, given the individual's background and socio-economic status. If so, then there is a greater probability of a neuropsychological deficit.

Throughout the development of the WAIS-III and WMS-III, there was a significant effort to develop techniques to assist the neuropsychologist. Though not included with the publication of the WAIS-III and WMS-III, research studies and development work on two of the techniques described above were included in the research design. First, an additional 437 examinees completed the WAIS-III while the WAIS-III and WMS-III were standardized. This "educational-level oversample" was collected to ensure that a minimum of 30 individuals within four educational levels were tested in each age group with the WAIS-III. This ensured that there were enough examinees at each educational level so that age-by-education levels could be created. Second, a new-word reading test was developed and co-administered with the WAIS-III and WMS-III. It was completed by 1,250 individuals. The test uses words that are phonetically difficult to decode and would probably require previous learning. Similar to the results presented by Vanderploeg and Shimka (1995), regression analyses demonstrate that this reading test adds more incremental validity in predicting IQ and Memory Scores than do equations that just include demographic variables in predicting IQ scores. Moreover, by being co-normed directly with the WAIS-III and WMS-III, the reading test should provide invaluable information to the clinician who is trying to determine premorbid IQ. It is unfortunate that these techniques were not included in the released versions of the tests.

A recent study by Smith-Seemiller, Franzen, Burgess, and Prieto (1997) suggests that such techniques have been slow to integrate into clinical practice. Smith-Seemiller and colleagues conducted a survey using a sample of some of the doctorate-level members of the National Academy of Neuropsychology. They discovered that despite all of the research that is being performed in this area, relatively few neuropsychologists are applying these techniques in their evaluations with patients. Instead, the vast majority of clinicians tend to rely solely on self-report data that is obtained in a clincal interview, and some also utilize an individual's vocational status to make rough predictions about premorbid functioning. In the *WAIS-III–WMS-III Technical Manual*, it was emphasized that good practice means that all scores should be evaluated in light of someone's life history, socioeconomic status, and medical and psychosocial history. It is unclear whether these examiners were not practicing in this fashion or whether they simply did not value the actuarial- or regression-based approaches that were surveyed.

LINKS BETWEEN THE WAIS-III AND OTHER MEASURES

Differences Between the WAIS-III and the WMS-III

Perhaps the most important development in the revision of the WAIS-R was to codevelop and conorm the WAIS-III and the WMS-III. Because the scales were co-normed, examiners can directly compare IQ and memory differences, which may lead to additional power in detecting when and what type of deficits occur. Discrepancies between intelligence and memory are sometimes used to evaluate memory impairment. With this approach, learning and memory are assumed to be underlying components of general intellectual ability and, as such, to be significantly related to the examinee's performance on tests of intellectual functioning. In fact, the examinee's IQ scores are often used as an estimate of the individual's actual memory ability. Several researchers have advanced the theory that when memory scores are significantly lower than IQ scores, the discrepancy is suggestive of a focal memory impairment (Milner, 1975; Prigatano, 1974; Quadfasel & Pruyser, 1955). This is especially true when the difference between the IQ and memory scores exceeds what one might expect when comparing an individual's performance to the normative sample. For instance, if the base rate of occurrence of a large IQ-Memory discrepancy is very low, then this discrepancy score would have clinical utility. By overlapping the samples so that everyone who was part of the WMS-III sample was also part of the WAIS-III sample, more accurate base rates of IQ-memory discrepancies may be obtained. The interpretation and treatment of memory deficits is beyond the scope of this chapter and the curious reader should refer to Larrabee (this volume) for more information about memory testing in the neuropsychological evaluation or to Sohlberg, White, Evans, and Mateer (1992) and Mateer, Kerns, and Eso (1996) for a review and presentation of treatment methods.

Differences Between the WAIS-III and Measures of Achievement

In educational settings, the IQ scores of the Wechsler intelligence tests had been widely used in the comparison of students' general ability level and their level of achievement. As observed by Gridley and Roid (1998), the main purpose of comparing ability and achievement is to evaluate the discrepancy between expected and observed achievement. Since the enactment of the Education for All Handicapped Children Act of 1975, the more recent Individuals with Disabilities Act (IDEA), 1990; and the reauthorization of IDEA (1997), the comparison of intellectual ability to academic achievement has become a key step in determining the presence of specific learning disabilities. Nevertheless, there are pros and cons about this methodology (Gridley & Roid, 1998).

To help clinicians analyze the discrepancy between expected and observed achievement, the WISC-III was linked to the Wechsler Individual Achievement Test (WIAT) (The Psychological Corporation, 1992). Using FSIQ as the measure of ability, one can predict what an individual's achievement scores should be at a given level of IQ. When an individual does not achieve this predicted level (e.g., a lower-than-predicted score on the WIAT) or exceeds the predicted level (e.g., a high score on the WIAT), the examiner should examine the test scores more closely. Critical values required for a given ability-achievement discrepancy score to be significant at the .05 and .15 levels were included in the

WAIS-III–WMS-III Technical Manual. More importantly, this technical manual also presents the frequencies of the ability-achievement discrepancy scores obtained by the standardization sample. Flanagan and Alfonso (1993a, 1993b) developed similar tables using VIQ and PIQ as measures of ability. This additional normative information has increased the clinical utility of the WISC-III in educational settings.

The tables reported in the WIAT Manual look at the relationship between the WAIS-R and the WIAT (for 17- to 19-year-olds). For the WAIS-III, a validity study was conducted in order to evaluate the relationship between the WAIS-III and the WIAT. A linking sample of 142 normal adults 16–19 years of age was used. The correlation coefficients are from .53 to .81 between the WAIS-III IQs and the WIAT composite scores, and from .37 to .82 between the WAIS-III IQs and the WIAT subtest scores. These results are comparable to those reported previously (Wechsler, 1991; The Psychological Corporation, 1992). Using similar tables as those reported in the WIAT manual (The Psychological Corporation, 1992), the technical manual reports abililty-achievement discrepancy scores for both a simple difference and a regression method. Discrepancy scores are reported as both statistically significant values as well as base rate frequency data. Following the WIAT tradition, the critical values required for a given ability-achievement discrepancy score to be significant at the .05 and .15 levels and the frequencies of the ability-achievement discrepancy scores obtained by the linking sample, were provided in the technical manual for both simple difference and regression methods.

When using the simple-difference method, the examiner subtracts the IQ scores directly from the achievement scores and evaluates significance and meaningfulness in a two-step process. First, the user should determine whether a given ability-achievement discrepancy is statistically significant. If it is, then the examiner should determine how frequently such a discrepancy had occurred in the linking sample.

When using the predicted-achievement method, the steps are a little more complicated. First, the examiner should find the predicted achievement scores using the examinee's IQ scores as a guide. The second step is to find the discrepancy score by subtracting the observed-achievement score from the predicted-achievement score. Third, the statistical significance of the difference should be decided and if it is statistically significant, then the examiner should determine if the discrepancy is rare by using the base-rate data from the linking sample.

In general, the predicted-achievement method is preferred for the ability-achievement analysis because it takes into account the measurement errors and the relationship between the measures of ability and achievement. Although it is easy to use, the simple-difference method assumes perfect correlation between the measures of ability and achievement and overlooks the measurement errors (Braden & Weiss, 1988).

NOTES

1. Object Assembly is still included in the WAIS-III but is considered an optional subtest; (see Wechsler, 1997a; p. 6).

2. Picture Arrangement generally had a split loading between the Perceptual Organization and Verbal Comprehension Indexes.

3. Generally, Arithmetic had a primary loading on this factor. In some of the analyses, however, it had split loadings with the Verbal Comprehension Index. Occasionally, it would have a split loading between the Working Memory and Perceptual Organization Indexes.

4. The mean of the verbal subtests was calculated using the six subtests that contribute to the Verbal IQ score (e.g., Vocabulary, Similarities, Arithmetic, Digit Span, Information, and Comprehension).

5. The mean of the performance subtests was calculated using the five subtests that contribute to the Performance IQ score (e.g., Picture Completion, Digit Symbol-Coding, Block Design, Matrix Reasoning, and Picture Arrangement).

REFERENCES

Atkinson, L. (1992). Mental retardation and WAIS-R scatter analysis. *Journal of Intellectual Disability Research, 36,* 443–448.

Baddeley, A. D. (1986). *Working Memory.* Oxford, England: Oxford University Press.

Baddeley, A. D., & Hitch, G. (1974). Working memory. In G. H. Bower (Ed.), *The Psychology of Learning and Motivation: Advances In Research and Theory* (Vol. 8, pp. 47–90). San Diego, CA: Academic Press.

Balinsky, B. (1941). An analysis of the mental factors in various age groups from nine to sixty. *Genetic Psychology Monographs, 23,* 191–234.

Barona, A., Reynolds, C. R., & Chastain, R. (1984). A demographically based index of premorbid intelligence for the WAIS-R. *Journal of Consulting and Clinical Psychology, 52,* 885–887.

Biederman, J., Faraone, S. V., Spencer, T., Wilens, T., Norman, D., Lapey, K. A., Mick, E., Lehman, B. K., & Doyle, A. (1993). Patterns of psychiatric comorbidity, cognition, and psychosocial functioning in adults with attention deficit hyperactivity disorder. *American Journal of Psychiatry, 150*(12), 1792–1798.

Blaha, J., & Wallbrown, F. H. (1996). Hierarchical factor structure of the Wechsler Intelligence Scale for Children-III. *Psychological Assessment, 8*(2), 214–218.

Blair, J. R., & Spreen, O. (1989). Predicting premorbid IQ: A revision of the National Adult Reading Test. *The Clinical Neuropsychologist, 3,* 129–136.

Bracken, B. (1988). Ten psychometric reasons why similar tests produce dissimilar results. *Journal of School Psychology, 26,* 155–166.

Braden, J. P., & Weiss, L. (1988). Effects of simple difference versus regression discrepancy methods: An empirical study. *Journal of School Psychology, 26,* 133–142.

Brody, N. (1992). *Intelligence* (2nd ed.). San Diego, CA: Academic Press.

Brown, T. E. (1996). *Brown Attention-Deficit Disorder Scales.* San Antonio, TX: The Psychological Corporation.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies.* New York: Cambridge University Press.

Carroll, J. B. (1997). The three-stratum theory of cognitive abilities. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 122–130). New York: Guilford Press.

Chelune, G. J., Ferguson, W., & Moehle, K. (1986). The role of standard cognitive and personality tests in neuropsychological assessment. In T. Incagnoli, G. Goldstein, & C. J. Golden (Eds.), *Clinical Application of Neuropsychological Test Batteries.* New York: Plenum Press.

Chen T., Tulsky D., & Tang H. (1997). Innovative approaches to removing bias in the WAIS-III. In D. Tulsky & J. Zhu (Co-Chairs), *Methodological considerations and innovations in the development of the WAIS-III.* Symposium conducted at the 105th annual American Psychological Association Convention, Chicago.

Cohen, J. (1952a). A factor-analytically based rationale for the Wechsler–Bellevue. *Journal of Consulting Psychology, 16,* 272–277.

Cohen, J. (1952b). Factors underlying Wechsler-Bellevue performance of three neuropsychiatric groups. *Journal of Abnormal and School Psychology, 47,* 359–364.

Cohen, J. (1957a). The factorial structure of the WAIS between early adulthood and old age. *Journal of Consulting Psychology, 21,* 283–290.

Cohen, J. (1957b). A factor-analytically based rationale for the Wechsler Adult Intelligence Scale. *Journal of Consulting Psychology, 6,* 451–457.

Cohen, J. (1959). The factorial structure of the WISC at ages 7-6, 10-5, and 13-6. *Journal of Consulting Psychology, 23,* 285–299.

Craft, N. P., & Kronenberger, E. J. (1979). Comparability of WISC-R and WAIS IQ scores in educable mentally handicapped adolescents. *Psychology in the Schools, 16*(4), 502–504.

Desai, M. M. (1955). The relationship of the Wechsler-Bellevue verbal scale and the Progressive Matrices Test. *Journal of Consulting Psychology, 19,* 60.

Donders, J. (1997). Sensitivity of the WISC–III to head injury with children with traumatic brain injury. *Assessment, 4,* 107–109. Education for All Handicapped Children Act, 20 U.S.C. §1400 *et. seq* (1975).

Flanagan, D. P., & Alfonso, V. C. (1993a). Differences required for significance between Wechsler verbal and performance IQs and the WIAT subtests and composites: The predicted-achievement method. *Psychology in the Schools, 30,* 125–132.

Flanagan, D. P., & Alfonso, V. C. (1993b). WIAT subtest and composite predicted-achievement values based on WISC-III verbal and performance IQs. *Psychology in the Schools, 30,* 310–320.

Flynn, J. R. (1984). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin, 95,* 29–51.

Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin, 101,* 171–191.

Franzen, M. D., Burgess, E. J., & Smith-Seemiller, L. (1997). Methods of estimating premorbid functioning. *Archives of Clinical Neuropsychology, 12*(8), 711–738.

Glutting, J. J., McDermott, P. A., & Konold, T. R. (1997). Ontology, structure, and diagnostic benefits of a normative subtest taxonomy from the WISC-III Standardization sample. In D. P. Flanagan, J. L. Grenshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 349–372). New York: Guilford Press.

Gold, J. M., Carpenter, C., Randolph, C., Goldberg, T. E., & Weinberger, D. R. (1997). Auditory working memory and Wisconsin Card Sorting Test performance in schizophrenia. *Archives of General Psychiatry, 54,* 159–165.

Gridley, B. E., & Roid, G. H. (1998). The use of the WISC-III with achievement tests. In A. Prifitera & D. Saklofske (Eds.), *WISC-III clinical use and*

interpretation: Scientist-practitioner perspectives (pp. 249–288). New York: Academic Press.

Grober, E., & Sliwinski, M. (1991). Development and validation of a model for estimating premorbid verbal intelligence in the elderly. *Journal of Clinical and Experimental Neuropsychology, 13,* 933–949.

Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review, 9,* 139–150.

Hall, J. C. (1957). Correlation of a modified form of Raven's Progressive Matrices (1938) with the Wechsler Adult Intelligence Scale. *Journal of Consulting Psychology, 21,* 23–26.

Harrison, P. L., Kaufman, A. S., Hickman, J. A., & Kaufman, N. L. (1988). A survey of tests used for adult assessment. *Journal of Psychoeducational Assessment, 6,* 188–198.

Hart, R. P., Kwentus, J. A., Wade, J. B., & Hamer, R. M. (1987). Digit symbol performance in mild dementia and depression. *Journal of Consulting and Clinical Psychology, 55*(2), 236–238.

Heaton, R. K. (1992). *Comprehensive norms for an expanded Halstead-Reitan Battery: A supplement for the WAIS-R.* Odessa, FL: Psychological Assessment Resources.

Heaton, R. K., Grant, I., & Matthews, C. G. (1986). Differences in neuropsychological test performance associated with age, education and sex. In I. Grant & K. M. Adams (Eds.), *Neuropsychological assessment in neuropsychiatric disroders: Clinical methods and empirical findings.* (pp. 100–120). New York: Oxford University Press.

Heaton, R. K., Grant, I., & Matthews, C. G. (1991). *Comprehensive norms for an expanded Halstead-Reitan Battery: Demographic corrections, research findings, and clinical applications.* Odessa, FL: Psychological Assessment Resources.

Heaton, R.K., Ryan, L., Grant, I., & Matthews, C. G. (1996). Demographic influences on neuropsychological test performance. In I. Grant & K. M. Adams (Eds.), *Neuropsychological assessment of neuropsychiatric disorders* (pp. 141–163). New York: Oxford University Press.

Henmon, V. A. C., Peterson, J., Thurstone, L. L., Woodrow, H., Dearborn, W. F., & Haggerty, M. E. (1921). Intelligence and its measurement: A symposium. *Journal of Educational Psychology, 12*(4), 195–216.

Hynd, G. W., & Semrud-Clikeman, M. (1990). Neuropsychological Assessment. In A. S. Kaufman (Ed.), *Assessing adolescent and adult intelligence* (pp. 638–695). Boston: Allyn & Bacon.

Individuals with Disabilities Education Act (IDEA). (1990). 20 U. S. C. 1400 *et. seq.* Individuals with Disabilities Education Act Amendments of 1997.

Ivnik, R. J., Malec, J. F., Smith, G. E., Tangalos, E. G., Petersen, R. C., Kokmen, E., & Kurland, L. T. (1992). Mayo's older adult normative studies: WAIS–R norms for ages 56-97. *The Clinical Neuropsychologist, 6* (Suppl.), 1–30.

Kamphaus, R. W. (1993). *Clinical assessment of children's intelligence.* Needham Heights, MA: Allyn & Bacon.

Kamphaus, R. W., Bension, J., Hutchinson, S., & Platt, L. O. (1994). Identification of factor models for the WISC-III. *Educational and Psychological Measurement, 54,* 174–186.

Kaplan, E. (1988). A process approach to neuropsychological assessment. In T. J. Boll & B. K. Bryant (Eds.), *Clinical neuropsychology and brain function: Research, measurement, and practice* (pp. 129–167). Washington, DC: American Psychological Association.

Kaplan, E., Fein, D., Morris, R., & Delis, D. C. (1991). *WAIS–R as a Neuropsychological Instrument manual.* San Antonio, TX: The Psychological Corporation.

Kareken, D. A., Gur, R. C., & Saykin, A. J. (1995). Reading on the Wide Range Achievement Test-Revised and parental education as predictors of IQ: Comparison with the Barona formula. *Archives of Clinical Neuropsychology, 10,* 147–157.

Kaufman, A. S. (1979). *Intelligent testing with the WISC-R.* New York: Wiley.

Kaufman, A. S. (1990). *Assessing adolescent and adult intelligence.* Boston: Allyn & Bacon.

Kaufman, A. S. (1991). King WISC the third assumes the throne. *Journal of School Psychology, 11,* 345–354.

Kaufman, A. S. (1994). *Intelligent testing with the WISC-III.* New York: Wiley.

Kaufman, A. S., McLean, J.E., & Reynolds, C. R. (1988). Sex, race, residence, region, and education differences on the 11 WAIS-R subtests. *Journal of Clinical Psychology, 44*(2), 231–248.

Krull, K. R., Scott, J. G., & Sherer, M. (1995). Estimation of premorbid intelligence from combined performance and demographic variables. *The Clinical Neuropsychologist, 9,* 83–88.

Kyllonen, P. C. (1987). Theory-based cognitive assessment. In J. Zeidner (Ed.), *Human productivity enhancement: Organizations, personnel, and decision making.* (Vol. 2, pp. 338–381). New York: Praeger.

Kyllonen, P. C., & Christal, R. E. (1987). Cognitive modeling of learning disabilities: A status report of LAMP. In R. Dillon & J. W. Pelligrino (Eds.), *Testing: Theoretical and applied issues.* New York: Freeman.

Kyllonen, P. C., & Christal, R. E. (1989). Cognitive modeling of learning abilities: A status report of LAMP. In R. Dillon & J. W. Pelligrino (Eds.),

Testing: Theoretical and applied issues. New York: Freeman.

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability (little more than) working-memory?! *Intelligence, 14,* 389–433.

La Rue, A. (1992). *Aging and neuropsychological assessment.* New York: Plenum Press.

Larrabee, G. J., & Curtiss, G. (1995). Construct validity of various verbal and visual memory tests. *Journal of Clinical and Experimental Neuropsychology, 17,* 536–547.

Larrabee, G. J., Kane, R. L., & Schuck, J. R. (1983). Factor analysis of the WAIS and Wechsler Memory Scale: An analysis of the construct validity of the Wechsler Memory Scale. *Journal of Clinical Neuropsychology, 5,* 159–168.

Larrabee, G. J., Largen, J. W., & Levin, H. S. (1985). Sensitivity of age-decline resistant ("hold") WAIS subtests to Alzheimer's disease. *Journal of Clinical and Experimental Neuropsychology, 7,* 497–504.

Leckliter, I. N., Matarazzo, J. D., & Silverstein, A. B. (1986). A literature review of factor analytic studies of WAIS-R. *Journal of Clinical Psychology, 42,* 332–342.

Levine, B., & Iscoe, I. (1954). A comparison of Raven's Progressive Matrices (1938) with a short form of the Wechsler–Bellevue. *Journal of Consulting Psychology, 18,* 10.

Lezak, M. D. (1988). IQ: R. I. P. *Journal of Clinical and Experimental Neuropsychology, 10,* 351–361.

Lezak, M. D. (1995). *Neuropsychological assessment* (3rd ed.). New York: Oxford University Press.

Lindeman, J. E., & Matarazzo, J. D. (1984). Intellectual assessment of adults. In G. Goldstein & M. Herson (Eds.), *Handbook of Psychological Assessment* (pp. 77–99). New York: Pergamon Press.

Lubin, B., Larson, R. M., & Matarazzo, J. D. (1984). Patterns of psychological test usage in the United States: 1935–1982. *American Psychologist, 39,* 451–454.

Lubin, B., Larson, R. M., Matarazzo, J. D., & Seever, M.F. (1985). Psychological test usage patterns in five professional settings. *American Psychologist, 40,* 857–861.

Malec, J. F., Ivnik, R. J., Smith, G. E., Tangalos, E. G., Petersen, R. C., Kokmen, E., & Kurland, L. T. (1992). Mayo's older adult normative studies: Utility of corrections for age and education for the WAIS-R. *The Clinical Neuropsychologist, 6* (Suppl.), 31–47.

Matarazzo, J. D. (1972). *Wechsler's measurement and appraisal of adult intelligence* (5th ed.). Baltimore: Williams & Wilkins.

Matarazzo, J. D. (1990). Psychological assessment versus psychological testing: Validation from Binet to the school, clinic, and courtroom. *American Psychologist, 45*(9), 999–1017.

Matarazzo, J. D., & Herman, D. O. (1985). Clinical uses of the WAIS-R: Base rates of differences between VIQ and PIQ in the WAIS-R standardization sample. In B. B. Wolman (Ed.), *Handbook of intelligence: Theories, measurements, and applications* (pp. 899–932). New York: Wiley.

Mateer, C. A., Kerns, K. A., & Eso, K. L. (1996). Management of attention and memory disorders following traumatic brain injury. *Journal of Learning Disabilities, 29(6),* 618–632.

McDermott, P. A., Fantuzzo, J. W., & Glutting, J. J. (1990). Just say no to subtest analysis: A critique on Wechsler theory and practice. *Journal of Psychoeducational Assessment, 8,* 290–302.

McDermott, P. A., Fantuzzo, J. W., Glutting, J. J., Watkins, M. W., & Baggaley, A. R. (1992). Illusions of meaning in the ipsative assessment of children's ability. *The Journal of Special Education, 25*(4), 504–526.

McDermott, P., Glutting, J. J., Jones, J. N., & Noonan, J. V. (1989). Typology and prevailing composition of core profiles in the WAIS-R standardixation sample. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 1,* 118–125.

McDermott, P., Glutting, J. J., Jones, J. N., Watkins, M. W., & Kush, J. (1989). Core profile types in the WISC-R national sample: Structure membership, and applications. *Psychological Assessment, 1,* 292-299.

Milner, B. (1975). Psychological aspects of focal epilepsy and its neurosurgical management. *Advances in Neurology, 8,* 299–321.

Mirsky, A. F. (1989). The neuropsychology of attention: Elements of a complex behavior. In E. Perecman (Ed.), *Integrating theory and practice in clinical neuropsychology,* (pp. 75–91). Hillsdale, NJ: Erlbaum.

Mirsky, A. F., Anthony, B. J., Duncan, C. C., Ahearn, M. B., & Kellam, S. G. (1991). Analysis of the elements of attention: A neuropsychological approach. *Neuropsychology Review, 2,* 109–145.

Nelson, H. E. (1982). *National Adult Reading Test (NART) manual.* Windsor, UK: NFER-Nelson.

Nelson, H. E., & McKenna, P. (1975). The use of current reading ability in the assessment of dementia. *British Journal of Social and Clinical Psychology, 14,* 259–267.

Nelson, H. E., & O'Connell, A. (1978). Dementia: The estimation of premorbid intelligence levels using the new adult reading test. *Cortex, 14,* 234–244.

Newell, A. (1973). Production systems. In W. G. Chase (Ed.), *Visual information processing.* New York: Academic Press.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Owenby, R. L., & Matthews, C. G. (1985). On the meaning of the WISC–R third factor: Relations to selected neuropsychological measures. *Journal of Consulting and Clinical Psychology, 53,* 531–534.

Piotrowski, C., & Keller, J. W. (1989). Psychological testing in outpatient mental health facilities in 1975. *Professional Psychology: Research and Practice, 20*(6), 423–425.

Posner, M. I. (1988). Structures and functions of selective attention. In M. Dennis, E. Kaplan, M. I. Posner, D. G. Stein, & R. F. Thompson (Eds.), *Clinical neuropsychology and brain function: Research, measurement, and practice*. Washington, DC: American Psychological Association.

Prifitera, A., & Dersh, J. (1992). Base rates of the WISC-III diagnostic subtest patterns among normal, learning-disabled, and ADHD samples. [WISC-III Monograph[. *Journal of Psychoeducational Assessment,* 43–55.

Prifitera, A., Weiss, L. G., & Saklofske, D. H. (1998). The WISC–III in context. In A. Prifitera & D. H. Saklofske (Eds.), *WISC-III clinical use and interpretation: Scientist-practitioner perspectives*. San Diego, CA: Academic Press.

Prigatano, G. P. (1974, May). *Memory deficit in head injured patients.* Paper presented at the meeting of the Southwestern Psychological Association, El Paso, TX.

The Psychological Corporation. (1992). *Wechsler Individual Achievement Test.* San Antonio, TX: Author.

The Psychological Corporation. (1997). *WAIS-III–WMS-III Technical Manual.* San Antonio, TX: Author.

Quadfasel, A. F., & Pruyser, P. W. (1955). Cognitive deficit in patients with psychomotor epilepsy. *Epilepsia, 4,* 80–90.

Raven, J. C. (1976). *Standard Progressive Matrices.* Oxford, England: Oxford Psychologists Press.

Raven, J., Raven, J. C., & Court, J. H. (1991). *Manual for Raven's Progressive Matrices and Vocabulary Scales*. Oxford, England: Oxford Psychologists Press.

Robertson, M. H., & Woody, R. H. (1997). *Theories and methods for practice of clinical psychology.* Madison, CT: International Universities Press.

Roid, G. H., Prifitera, A., & Weiss, L. G. (1993). Replication of the WISC-III factor structure in an independent sample [WISC-III Monograph]. *Journal of Psychoeducational Assessment,* 6–20.

Roid, G. H., & Worrall, W. (1996, August). *Equivalence of factor structure in the U.S. and Canada editions of WISC-III.* Paper presented at the meeting of the American Psychological Association, Toronto, Canada.

Rosenberg, H. M., Ventura, S. J., Maurer, J. D., Heuser, R. L., & Freedman, M. A. (1996). Births and deaths: United States, 1995. *Monthly Vital Statistics Report, 45*(3, Suppl. 2). (Preliminary Data from the Centers for Disease Control and Prevention/National Center for Health Statistics).

Russell, E. (1972). WAIS factor analysis with brain-damaged subjects using criterion measures. *Journal of Consulting and Clinical Psychology, 39,* 133–139.

Ryan, J. J., Paolo, A. M., & Brungardt, T. M. (1990). Standardization of the Wechsler Adult Intelligence Scale-Revised for persons 75 years and older. *Psychological Assessment, 2,* 404–411.

Sattler, J. M. (1992). *Assessment of children: WISC-III and WPPSI-R supplement.* San Diego, CA: Author.

Shackelford, W. (Producer). (1978). *A conversation with David Wechsler* [Videotape]. (Available from Jeffrey Norton Publishers, On The Green, Guilford, CT 06437-2612)

Sherman, E. M. S., Strauss, E., Spellacy, F., & Hunter, M. (1996). Construct validity of WAIS–R factors: Neuropsychological test correlations in adults referred for evaluation of possible head injury. *Psychological Assessment, 7,* 440–444.

Smith, G. E., Ivnik, R. J., Malec, J. F., Kokmen, E., Tangalos, E. G., & Kurland, L. T. (1992). Mayo's older Americans normative studies (MOANS): Factor structure of a core battery. *Psychological Assessment, 4*(3), 382–390.

Smith, G. E., Ivnik, R. J., Malec, J. F., Petersen, R. C., Kokmen, E., & Tangalos, E. G. (1994). Mayo cognitive factors scales: Derivation of a short battery and norms for factor scores. *Neuropsychology, 8*(2), 194–202.

Smith-Seemiller, L., Franzen, M. D., Burgess, E. J., & Prieto, L. R. (1997). Neuropsychologists' practice patterns in assessing premorbid intelligence *Archives of Clinical Neuropsychology, 12(8),* 739–744.

Sohlberg, M. M., White, O., Evans, E., Mateer, C. (1992). An investigation of the effects of prospective memory training. *Brain Injury, 6(2),* 139–154.

Spearman, C. E. (1904). "General intelligence," objectively determined and measured. *American Journal of Psychology, 15,* 201–293.

Spearman, C. E. (1927). *The abilities of man.* New York: Macmillian.

Spruill, J. (1991). A comparison of the Wechsler Adult Intelligence Scale-Revised with the Stanford-Binet Intelligence Scale (4th ed.) for mentally retarded adults. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 3*(1), 1–3.

Sternberg, R. J. (1993). Rocky's back again: A review of WISC-III. [WISC-III Monograph]. *Journal of Psychoeducational Assessment,* 161–164.

Sternberg, R. J. (1995). *In search of the human mind.* Orlando, FL: Harcourt Brace College Publishers.

Stewart, K. J., & Moely, B. E. (1983). The WISC-R third factor: What does it mean? *Journal of Consulting and Clinical Psychology, 51*(6), 940–941.

Sweet, J. J., Moberg, P. J., & Tovian, S. M. (1990). Evaluation of Wechsler Adult Intelligence Scale-Revised premorbid IQ formulas in clinical populations. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 2,* 41–44.

Swiercinsky, D. P., & Warnock, J. K. (1977). Comparison of the neuropsychological key and discriminant analysis approaches in predicting cerebral damage and localization. *Journal of Consulting and Clinical Psychology, 45,* 808–814.

Thorndike, E. L., Lay, W., & Dean, P. R. (1909). The relation of accuracy in sensory discrimination to general intelligence. *American Journal of Psychology, 20,* 364–369.

Thorndike, E. L., Terman, L. M., Freeman, F. N., Colvin, S. S., Pintner, R., Ruml, B., & Pressey, S. L. (1921). Intelligence and its measurement: A symposium. *The Journal of Educational Psychology, 12*(3), 123–147.

Tulsky, D. S., & Chen, H. (1998, Fall). *Assessment Focus Newsletter.* San Antonio, TX: The Psychological Corporation.

Tulsky, D., & Zhu, J. (1997, August). Lowering the floor of the WAIS–III. In D. Tulsky & J. Zhu (Co-chairs), *Methodological considerations and innovations in the development of the WAIS-III.* Symposium conducted at the 105th annual American Psychological Association Convention, Chicago.

Tulsky, D., Zhu, J., & Vasquez, C. (1998, February). The clinical utility of WAIS-III index scores in patients with neuropsychological disorders. In D. Tulsky & M. Ledbetter (Co-chairs), *Patterns of the WAIS-III and WMS-III performance in several samples of individuals with neurological disorders.* Symposium conducted at the 26th annual International Neuropsychological Society Convention, Honolulu, HI.

United States Department of Labor. (1996). [Current population survey]. Unpublished raw data. Washington, DC: Bureau of Labor Statistics.

Vanderploeg, R. D., & Schinka, J. A. (1995). Predicting WAIS-R IQ premorbid ability: Combining subtest performance and demographic variable predictors. *Archives of Clinical Neuropsychology, 10,* 225–239.

Watkins, C. E., Campbell, V. L., Nieberding, R., & Hallmark, R. (1995). Contemporary practice of psychological assessment by clinical psychologists. *Professional Psychology: Research and Practice, 26*(1), 54–60.

Watson, C. G., & Klett, W. G. (1974). Are nonverbal IQ tests adequate substitutes for WAIS ? *Journal of Clinical Psychology, 30,* 55–57.

Wechsler, D. (1939). *Wechsler-Bellevue Intelligence Scale.* New York: The Psychological Corporation.

Wechsler, D. (1944). *The measurement of adult intelligence* (3rd ed.). Baltimore: Williams & Wilkins.

Wechsler, D. (1950). Cognitive, conative, and non-intellective intelligence. *American Psychologist, 5,* 78–83.

Wechsler, D. (1955). *Wechsler Adult Intelligence Scale.* New York: The Psychological Corporation.

Wechsler, D. (1974). *Wechsler Intelligence Scale for Children-Revised.* San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1975). Intelligence defined and undefined: A relativistic appraisal. *American Psychologist, 30,* 135–139.

Wechsler, D. (1981). *Wechsler Adult Intelligence Scale-Revised.* San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1991). *Wechsler Intelligence Scale for Children-Third Edition.* San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1997a). *WAIS-III Administration And Scoring Manual.* San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1997b). *Wechsler Memory Scale-Third Edition.* San Antonio, TX: The Psychological Corporation.

Wielkiewicz, R. M. (1990). Interpreting low scores on the WISC-R third factor: It's more than distractibility. *Psychological Assessment, 2,* 91–97.

Wilson, R. S., Rosenbaum, G., Brown, G., Rourke, D., Whitman, D., & Grissell, J. (1978). An index of premorbid intelligence. *Journal of Consulting and Clinical Psychology, 46,* 1554–1555.

Woltz, D. J. (1988). An investigation of the role of working memory in procedural skill acquisition. *Journal of Experimental Psychology: General, 117,* 319–331.

Yates, A. (1956). The use of vocabulary in the measurement of intellectual deterioration—A review. *Journal of Mental Science, 102,* 409–440.

Zhu, J., & Tulsky, D. (1997). The consistency and discrepancy between the WISC-III and WAIS-III. In D. Tulsky & J. Zhu (Co-Chairs), *Methodological considerations and innovations in the development of the WAIS–III.* Symposium conducted at the 105th annual American Psychological Association Convention, Chicago.

Zhu, J., & Tulsky, D. S. (1999). Can IQ gain be accurately quantified by a simple difference formula? *Perceptual and Motor Skills, 88,* 1255-1260.

This Page Intentionally Left Blank

CHAPTER 6

GROUP INTELLIGENCE TESTS

Robert W. Motta
Jamie M. Joseph

The terms "intelligence test" and "IQ test" are used synonymously in this chapter. This is done more for reasons of convenience than for accuracy, for it is clear that the term "intelligence" implies a far wider range of abilities and adaptive skills than does a single IQ score. Whether referred to as IQ or intelligence, group tests of intellectual ability are used extensively in the United States and throughout the world. Over the last 30 to 40 years countries such as Belgium, France, the Netherlands, and Norway have regularly tested all young people entering military service; while countries such as Australia, Canada, former East Germany, Great Britain, and New Zealand have regularly conducted large-scale group intelligence testing of school children (Flynn, 1987).

EARLY HISTORICAL PERSPECTIVE

Written examinations of academic capabilities are a fairly recent development in the United States, but use of assessment procedures to evaluate the capabilities of groups of individuals has a long history. Records indicate that as early 2357 B.C. Chinese emperors were employing examinations of military officers. In 1115 BC civil service examinations were first used in China (Aiken, 1976). From 500 BC to 100 A.D. the Greeks employed tests for military proficiency and for college admissions. In 1200 AD the first oral examination for the Ph.D. degree as well as public and private exams for the Master of Law degree were held at the University of Bologna. It was not until the 1860s that United States' schools and colleges began using written examinations. Modern U.S. college-admissions testing dates back to 1900, when the College Entrance Examination Board was founded as a membership association of colleges and universities (Garber & Austin, 1982).

Interest in the study of individual differences was stimulated by Charles Darwin's work on the origin of species. Sir Francis Galton, the cousin of Charles Darwin, was interested in the hereditary nature of genius and published Hereditary Genius in 1869. Galton also devised a number of sensory-motor tests and developed procedures for studying individual differences. As an outgrowth of his efforts to measure individual variation, Galton in 1888 described a method of "co-relations," which is the basis of modern correlational procedures (Aiken, 1976). During this time, experimental psychologists in Germany, including Fechner, Wundt, and Ebbinhaus demonstrated that, as could physical phenomena, psychological events could be quantitatively assessed.

James M. Cattell, who studied in Germany for his Ph.D. degree, became acquainted with Galton's methods of sensory-motor assessment and attempted to relate them to what he called "mental tests" in the late 1800s. Theorists such as Spearman, Thorndike, Thurstone, Cattell, and Guilford, and more practically oriented psychologists such as Terman, Wechsler, Bayley, and Gishelli made substantial contributions to the assessment of intelligence. Yet, Alfred Binet and Theodore Simon are

credited with constructing the first mental test that was effective in predicting academic achievement in school.

DEVELOPMENT OF INDIVIDUAL AND GROUP INTELLIGENCE MEASURES

In the early 20th century Binet and Simon were commissioned by the Minister of Public Instruction in Paris to identify those children who were not able to benefit from regular public education. Binet used 30 school-related questions of increasing difficulty that were reputed to assess the ability to judge, understand, and reason. A later revision of this test in 1908 contained far more items than the original, and these were grouped into age levels from 3 to 13 years. It was in this later revision that the concept of *mental age* was introduced. Three years later Binet and Simon extended their test to the adult level.

Around 1915, Otis in the United States and Burt in England were experimenting with group intelligence tests for children. Not only was it seen as economical to assess children in large groups, but group tests could be administered by teachers who did not require the extensive training needed to administer Binet testing. These group tests tapped many of the processes of the individual tests, including the comprehension of relations (e.g., analogies), classification, vocabulary, problem solving, common knowledge, and so on (Vernon, 1978). The advent of World War I in 1917 served as a powerful stimulus for wide-scale group intelligence testing of young adults in the United States. The work of Arthur Otis, Lewis Terman, and Robert Yerkes, who was then president of the American Psychological Association, resulted in the development of the Army Alpha Intelligence Test. This test was used to screen large numbers of recruits for WWI so that they could be placed in service positions for which they were most suited. The Army Alpha, designed for literate recruits, and the Army Beta, for the less literate, were used in the screening of 1,726,966 men in 35 camps. Testing of military recruits by means of various group-administered assessment devices continues to this day.

Assessment of young men for military service and placement of these individuals in positions for which they were suited eventually led to civilian use of group tests of intelligence and ability. Following World War I several variations of the Army Alpha were used in hundreds of schools to assess academic capabilities. These group-administered, objectively scored tests were viewed by many as being superior to standard methods of teacher evaluation and grades. By 1923 use of group-intelligence testing devices had expanded to the point where 37 different group tests of intelligence were identified (Pintner, 1923). Thus, by the early part of the 1900s, group intelligence tests had firmly taken hold and established their utility in the identification of those who could benefit from academic instruction of various kinds.

DEFINITIONS OF INTELLIGENCE

Whether measured by group-administered or individually administered tests, the definition of intelligence has varied considerably over time. The fact that these tests correlate with academic performance does not help in clarifying what intelligence is. This ambiguity has led many to take the position that "intelligence is what intelligence tests measure." An overview of these definitions (Sattler, 1988) is presented below; however, space limitations prevent little more than a brief mentioning of them. What should be noted in the material that follows is the wide array of definitions and the equally diverse theoretical models of intelligence.

Binet and Simon (1916) focused on a set of qualities such as judgment, common sense, initiative, and adaptation; while Wechsler (1958) stressed that intelligence implies purpose, rationality, and ability to deal effectively with the environment. Factor-analytic and statistical theories of intelligences such as those of Spearman (1923) and Vernon (1950), proposed a general theory of intelligence, whereas others like Thorndike (1927) and Thurstone (1938) viewed intelligence as being composed of many independent faculties. Thurstone enumerated at least eight primary mental factors, including verbal, perceptual-speed, inductive-reasoning, number, rote-memory, deductive-reasoning, word-fluency, and space or visualization. Thorndike described three kinds of intelligence: social, concrete, and abstract. Spearman (1923) proposed a two-factor theory which emphasized a general factor (g) and one or more specific factors (s). The concept of a "g" factor has had a tremendous impact on early and current conceptualizations of intelligence.

Guilford (1967) proposed that three classes of variables must be considered when attempting to

define intelligence, and these include: the activities or operations performed (operations), the material or content on which the operations are performed (content), and the product that is the result of the operations (products). Vernon (1950) put forth a hierarchical approach to intelligence emphasizing the "g" factor. Listed under the "g" factor were verbal-educational and spatial-mechanical group factors, and these were further broken down into minor group factors. R. B. Cattell (1963) and Horn (1985) suggested two types of intelligence: fluid, which referred to capacity and which was independent of experience, and crystallized, which was learned knowledge. Campione and Brown (1978) stressed an information-processing approach to intelligence. Sternberg (1986) saw intelligence as consisting of three dimensions: the componential dimension, which related to internal mental mechanisms; the experiential dimension, which related to both the external and internal worlds; and the contextual dimension which related to the external world of the individual. Sternberg defined intelligence as "the mental activity involved in purposive adaptation to, shaping of, and selection of real-world environments relevant to one's life" (p. 33).

Das (1973) proposed a non-hierarchical simultaneous-successive information-processing model as a way of categorizing cognitive ability. Simultaneous processing occurs in an integrated, usually semi-spatial form; successive processing is sequence dependent and temporally based. Jensen (1973) attempted to demarcate two separate but partially interdependent mental functions: associative ability, which is represented by memory and serial-learning tasks, and cognitive ability, which is represented by conceptual-reasoning tasks. Gardner, H. (1983) viewed intelligence in terms of problem solving and finding or creating problems, and he suggested the assessment of a number of competencies for solving problems.

Some have argued that traditional views of intelligence are too restrictive and that what is measured on group IQ tests does not relate to how one functions in the "real world" (e.g., McClelland, 1973; Neisser, 1976). Sternberg, Wagner, Williams, and Horvath (1995), for example, stated, "Even the most charitable view of the relation between intelligence tests scores and real-world performance leads to the conclusion that the majoritiy of variance in real-world performance is not accounted for by intelligence test scores" (p. 913). Sternberg et al. (1995) suggested use of "practical intelligence" measures that would assess an individual's ability to problem solve and to know how to proceed in real-life situations. Traditional group intelligence tests are said to assess functioning that is more related to school performance than to on-the-job performance or to solving problems of daily living. As might be expected, others (e.g., Barrett and Depinet, 1991) dismiss the utility of practical intelligence in favor of the more traditional (real) mesures.

CRITICISMS OF THE TESTS

Although group measures of intelligence had been widely accepted following World War I, a number of controversies began to emerge with regard to the abilities of different groups of people. Publication of results of the Army Alpha from World War I revealed that there were considerable differences among scores when scores were classified according to the recruits' national or racial origin. Differences were noted between the mean scores of recruits of Anglo-American or northwestern European descent and those descended from southern and eastern European backgrounds; and between American whites and those of African-American heritage. Some argued that the mental capability of the average white U.S. Army recruit was equivalent to that of a 12-year-old child. Cronbach (1975) and Haney (1981) provide detailed descriptions of the controversies that erupted in the 1920s. In addition to the scholarly debates in academic circles with regard to the use of group intelligence tests, Lippman (1922a, 1922b; 1923) led a press attack on the value of intelligence testing as a whole. Rebuttals were provided by Freeman (1922), Terman (1922a, 1922b), Brigham (1923), and Yerkes (1923).

Adding to the controversy over observed differences in intellectual capability as a function of national origin and ethnic group, was the issue of whether intelligence arose primarily through hereditary factors or environmental influences. At the time when group-intelligence measures were initially developed, it was assumed that differences in intelligence were largely because of genes. Doubts were cast upon the genetic position when it was found that those scoring poorly on the Army Alpha were usually from relatively low socio-economic backgrounds and from areas where there were scant educational resources.

In support of the environmental side of the controversy, Gordon (1923) conducted a study with gypsy and canal-boat children in England who received little, if any, formal education and found that these children, up to the age of six, scored in the average range on intelligence tests. After that their mental ages failed to progress and their IQs declined, showing the negative impact of a lack of schooling. These children did not show a decline on non-verbal performance tests, and this supported the position that environment, and specifically education, played a major role in IQ. Similar results were obtained in the United States with children living in isolated rural communities or mountainous regions of Kentucky (Hirsch, 1928). Further research starting in the late 1920s and 1930s continued to suggest that IQ scores could be raised significantly when children were placed in enriched environments. In 1937 Newman, Freeman, and Holzinger published a study involving identical twins who were separated and reared apart shortly after birth. Despite having the same genes, IQ differences as much as 24 points were found between a few pairs whose environments were highly divergent.

The heredity-versus-environment debate continued for many years and resulted in a scholarly survey by R.S. Woodworth in 1941, concluding that heredity and environment both contribute to a given IQ score. In a somewhat similar vein, D.O. Hebb (1949) argued for two kinds of intelligence. One type of intelligence was seen as being acquired genetically; the other represented an interaction of genetically-based potential in interaction with environmental stimulation. Despite these developments, proponents of testing in the early- to mid-1900s remained unwilling to concede that environmental issues played a major role in IQ scores.

In a 1967 publication, A.R. Jensen stressed the importance of environmental influences on intellectual development. However, in 1969, when reviewing what appeared to be a failure of Head Start programs, he indicated that it was "not an unreasonable hypothesis" that genetic factors were involved in the average "Negro-white difference" (Vernon, 1978; p. 266). At this time many psychologists saw intelligence as largely malleable and responsive to a variety of environmental alterations despite a possible genetic base. Jensen (1969) put forward an alternative hypothesis which could be subject to testing. Later his views were solidified in support of genes as the major determinant of IQ (Jensen, 1973). What followed from Jensen's support of the possibility that genes contributed more to intelligence than did environment and that IQ scores were not as malleable as once believed, was a fire storm of criticism of Jensen, his work, and of anyone else who supported his position in whole or part. Thus, Hernstein (1973), Shockley (1971), and Eysenck (1971) were similarly pilloried for their views on genes and IQ. Vernon (1979) notes that part of the vitriol that followed Jensen's publication may have been due to an antiestablishment Zeitgeist which was in vogue at the time of Jensen's writing. Shuey (1958, 1966), who had also argued for the importance of genetic factors in assessing the intelligence of AfroAmericans, raised less of an uproar because of the earlier period in which that work came to print.

Since Jensen's work and the subsequent controversies in response to his writings, there have been numerous objections to the use of intelligence tests. Many of these criticisms have centered around the potential negative impact that tests can have on certain groups. Several court battles have been fought over the use of tests for the categorization and placement of children within special-education classes. In some instances tests for the purpose of educational placement have been rejected by the courts only to be reinstituted in later court decisions (Reshley, Kicklighter, & McKee, 1988). Since 1970, a number of states have banned the use of intelligence tests within schools or have considered such bans. In some instances, a ban or moratorium has been specifically directed to group intelligence testing. Those who favor testing assert that intelligence tests can be beneficial to individuals who may have good ability but are handicapped by poor past performance or a poor academic record. Gordon and Terrell (1981), who have strongly criticized the misuse of standardized tests, nevertheless, state: "To argue that standardized testing should be done away with or radically changed simply because ethnic minorities and disadvantaged groups do not earn as high scores as do middle-class whites is an untenable position" (p. 1170). These authors do suggest, however, that the wholesale use of standardized tests be greatly reduced. Gordon and Terrell propose the development of alternate devices and procedures that would be "process sensitive instruments designed to elicit data descriptive of the functional and conditional aspects of learner behavior" (p. 1170).

It would be naive to suppose, however, that if tests were developed that could do all the things Gordon and Terrell and other critics suggest, these tests would then be above criticism. Hargadon (1981), for example, states:

> As a subject that invites debate and controversy, tests and their uses must rank with religion, politics and sex. Tests, at least in part, are designed to do a dirty job: they help us make discriminating judgements about ourselves, about others, about levels of accomplishment and achievement, about degrees of effectiveness. They are no less controversial when they perform their tasks well than when they perform them poorly. Indeed, it can be argued that the better the test, the more controversial its use becomes. (p. 1113)

THE BELL CURVE CONTROVERSY

Throughout most of the mid-1980s to mid-1990s, emotional debates over hereditary versus environmental factors in intelligence had largely subsided, or at least had been less emphasized in the popular press. Heated discussion in the press and on television talk-shows was rekindled when a book, by Herrnstein and Murray (1994) entitled *The Bell Curve: Intelligence and Class Structure in American Life* appeared and asserted that black-white differences in intelligence were primarily due to genetic influences. Group intelligence tests, along with individual tests, supposedly supported the gene and IQ linkage.

Debate over the relative contributions of genes and of the environment to IQ has existed for many years, as has the contention that certain racial and ethnic groups differ from others with regard to abilities. Approximately 2,000 years ago Cicero acknowledged that Britons were too stupid to make good slaves. The latest such discourse takes place in Hernstein and Murray's 845-page book, which marshals a vast array of data, tables, and statistics to support a number of specific points. These points include the stance (a) that IQ is due primarily to genes and therefore cannot be easily altered; that blacks score 15 points lower than whites on IQ tests, but that Asians outscore whites, and these score differences are due primarily to genetic differences; (b) that social programs such as welfare and similar efforts designed to assist those at the bottom of the socio-economic barrel are wasteful because low socio-economic functioning is probably due to low IQ which the social programs will not be able to raise; that the high-IQ readers of their book have been unfairly burdened by having to support these relatively wasteful social welfare programs; and (c) that the future belongs to those with high IQs, or what Herrnstein and Murray cavalierly refer to as the "cognitive elite." Response to Herrnstein and Murray's work, in both the media and within the community of scholars, has been far more critical than supportive (e.g., Gardner, 1995; Gould, 1995; Kamin, 1995; Miller, 1995; Nisbett, 1995).

Those opposed to the conclusions drawn in *The Bell Curve* state their opposition on many levels. Some of the objections to the book include (a) that Herrnstein and Murray's work presents no new data or novel analyses but simply restates old eugenics arguments; (b) that the authors erroneously present the relative contributions of both genes and the environment; trained geneticists (which *they* are not) would be unable to delineate the relative contributions of these variables; and (c) that the expression of IQ represents the contributions of both genes the and environment, and that these two factors cannot be discussed in terms of their individual inputs to IQ. It is also argued that there is an overwhelming number of studies (including studies cited in Herrnstein and Murray's own work) showing that IQ can be raised substantially by environmental interventions; and that the observed 15-point gap between black and white IQs which they attribute to race is also seen between racially homogeneous groups, such as Catholics and Protestants in Ireland (Herbert 1995), and between Ashkenazic and Sephardic Jews. It has also been argued, as stated at the outset of this chapter, that IQ is not a comprehensive measure of intelligence and that it is inaccurate to use IQ measures constructed by a dominant social group (whites) to assess those (African Americans, Latinos, and others) who have been excluded from full participation and integration into the larger society .

The level of scientific and public ire that arose in response to *The Bell Curve* is reflected in the following vitriolic quotations: "The Bell Curve...contains no new arguments and presents no compelling data to support its anachronistic social Darwinism (Gould, 1995, p4); "In short, the Bell Curve is not only sleazy; it is, intellectually, a mess" (Ryan, 1995, p. 28); "I gradually realized that (when reading the Bell Curve) I was encountering a style of thought previously unknown to me: scholarly brinkmanship" (Gardner, 1995, p. 63); "The Bell Curve, a scabrous

piece of racial pornography masquerading as serious scholarship" (Herbert, 1995, p. 249); "The book has nothing to do with science" (Kamin, 1995, p. 99); "The book lays out its evidence in very convincing and well thought out ways, but it is just scholarly window dressing for the same old prejudice that has plagued this country since its very conception. The most detrimental aspect of this book is that it attempts to absolve individuals and society from the necessity of positive interventions for a diverse majority of the American people. Historically, there is significant evidence that this would be a grave error" (Richardson, 1995, pp.43–44). Many in the Asian community, whom Hernstein and Murray claim to be at the top of the IQ pyramid, have also objected to the genetic and racial positions espoused in *The Bell Curve*. "Asian Americans must not allow themselves to be misused in the service of Murray and Hernstein's political agenda." (Chon, 1995, p. 239). The political agenda Chon refers to is the elimination of social programs to aide those disenfranchised members of our society.

One of the most basic errors *The Bell Curve* is accused of making is its attempt to equate human intelligence with IQ scores. The single number representing IQ is only distantly related to one's ability to perceive and understand the environment, to draw reasonable conclusions, to understand social context, to demonstrate creative cognitive processes, and to show altruism and empathy. All of these capabilities, and many more, have been tied to intelligence and cannot be meaningfully reflected in a single score.

Those supporting the position of genetically based black-white IQ differences find fault with explanations that emphasize environmental influences. Jensen (1995), for example argues that "Individual differences in adult IQ are largely genetic, with a heritability of about 70 percent" (p. 335). The three most common environmental explanations for black-white differences in tested IQs are those that point to disadvantages or oppression cultural differences, and psychological maladjustments (Frisby, 1995). Disadvantages or Oppression explanations assert that black children are unable to achieve as a group on a level commensurate with whites because they have been historically denied commensurate opportunities to develop educationally (e.g., Myrdal, Sterner, & Rose, 1962). The cultural-differences explanation proposes that blacks are hindered in academic and testing situations because they are forced to accept and learn about a culture that is alien to their natural culture (e.g., Allen & Boykin, 1992). Finally, the psychological-maladjustment explanation asserts that a combination of racism, poverty, and cultural incongruence causes psychological damage, such as impaired self-esteem, and that such maladjustment impairs academic functioning and performance on IQ tests (e.g. Boykin, 1986).

Frisby (1995) refutes these environmental explanations as being the primary reason for black-white differences and supports Hernstein and Murray's work on the "facts" of genetic influence, by stating: "When facts and orthodoxy collide, the bearer of the facts is reflexively accused of being 'elitist,' 'racist,' 'incompassionate [sic] toward the plight of minorities', 'culturally insensitive,' and 'ideologically reactionary'." Thus, Frisby sees the arguments put forth in *The Bell Curve* as convincingly supporting a genetic basis of intelligence and environmental explanations as efforts to deny harsh reality.

The debate over the environment or genes and IQ scores has in no way been resolved nor is it likely to be settled anytime in the near future. It is a debate that evokes strong emotions, and seemingly compelling data can be presented to support both sides. Intelligence tests themselves, whether in group or individual format, are not constructed to shed light on the heredity-environment issue, but rather, have been used as a tool in the debate. What intelligence tests do is to reliably measure how an individual functions in response to specific items at a given point in time. The types of items seen on intelligence tests are intended to correlate with various areas of "real world" functioning, such as performance in school or in employment settings. Whether this generalization from test to real world is, in fact, achieved is another matter for debate. However, items within group intelligence tests assess a variety of areas such as reasoning, general knowledge, vocabulary, problem-solving ability, and nonverbal skills.

TEST CONTENT

The most common items on group intelligence tests are those that require some form of reasoning ability. For example, analogies can be presented in a verbal format or in a nonverbal manner where one is required to understand the relationship

between various forms. A typical verbal analogy that involves reasoning skills would be of the form below:

> WOOD is to TREE as PAPER is to: LAKE IRON PEN MILL PULP

Other types of verbal reasoning items include similarities, such as the example below:

> SMART means the same as: LIVELY HAPPY AGREEABLE CLEVER

Another type of item called oddities, is of the type that follows:

> Underline the word that does not belong with the others: DESK TABLE CHAIR BOOK BOOKCASE

Other types of reasoning items involve logical reasoning, as in the example below:

> Bob is shorter than John
>
> Ralph is taller than Bob
>
> Who is the shortest?

Items that depend upon reasonable inferences and judgments, based on the information given, are called inferential conclusions. These items are similar in form to items of reading comprehension, except that when used in intelligence tests, the level of vocabulary and reading difficulty are kept simple so that items are not dependent upon vocabulary or reading per se. An example of this kind of item, reported by Jensen (1980), is as follows:

> In a particular meadow there are a great many rabbits that eat the grass. There are also many hawks that eat the rabbits. Last year a disease broke out among the rabbits and most of them died. Which one of the following things most probably occurred?
>
> a) The grass died and the hawk population decreased.
> b) The grass died and the hawk population increased.
> c) The grass grew taller and the hawk population decreased.
> d) The grass grew taller and the hawk population increased.
> e) Neither the grass not the hawks were affected by the death of the rabbits. (p. 151)

In a random sample of the adult population in the United States, 52 percent chose the correct answer which was "c".

Other reasoning items are of a numerical nature, as shown below:

> John is twice as old as Jim, who is four years old. How old will John be when Jim is 15?

An example of a number series requiring reasoning is as follows:

> Write the number that will complete the following series: 3 8 13 18 23 _____

The majority of items found on group intelligence tests are of the multiple-choice variety. Concern has been raised about multiple-choice tests in that some assert that such items measure only superficial knowledge. This argument may certainly be true in some instances. For example little reasoning is needed for the following question:

> Which measure is equivalent to an average?
>
> a) Mean
>
> b) Median
>
> c) Mode
>
> d) Quartile

However, the following question requires fairly sophisticated reasoning and knowledge of statistics:

> The correlation of SAT-verbal and SAT-Math among all test takers is about .5. For a group of applicants admitted to Harvard University, the correlation is probably
>
> a) Greater that .5
>
> b) About .5
>
> c) Less than .5
>
> d) No way to determine

For this question one must reason that since Harvard University is highly selective, the group of applicants who are admitted must have fairly homogeneous test scores, so the correlation will be lower than that of the national group.

In addition to items that require reasoning, there are questions that are based on knowledge of vocabulary, as shown in the example below:

> Enmity means a) opponent b) hatred c) love d) vacant

Vocabulary items have been criticized because they rely heavily on educational background. Acquisition of vocabulary is not just a matter of learning and memory, but also requires discrimination, generalization, and education. Throughout one's life everyone hears many more words than become part of their vocabularies. Some people,

however, acquire much larger vocabularies than others, and this is true even among siblings of the same family. The effective use of vocabulary requires the ability to make fine discriminations and to reason abstractly. Therefore, it is of little surprise that knowledge of vocabulary is a critical component in "g." In addition to written presentations, vocabulary items can be presented in pictorial form to be used with children and nonreaders; they sometimes appear on group intelligence tests in the lower grades.

Items that tap an individual's general fund of information may also be used on group intelligence tests and are open to the same criticism that is applied to vocabulary items. They correlate highly with other noninformation measures of intelligence because an individual's range of knowledge is a good indication of ability. These items provide the most problems with respect to cultural differences because of the difficulty in determining the range of information an individual from a different culture might be expected to know. For this reason, vocabulary items and general information items do not appear as frequently on many group intelligence tests today as they once did. There are many other types of verbal and nonverbal test items, all of which have been shown to make a contribution to "g." When tests must be administered to large groups, as most group intelligence tests are, issues such as ease of administration and ease of scoring become important factors, and these influence the selection of items.

TEST SCORES

Group intelligence tests used in academic settings can be subdivided into three categories, which are based on the types of score they yield (Nitk,1983). These are *the single-score omnibus tests, the three-score tests,* and *the multiple-aptitude tests.* A single-score omnibus test reports one score that encompasses several different aspects of general scholastic ability combined into a single number. An example of an omnibus group-administered intelligence test is the Otis-Lennon School Ability Test (Otis & Lennon, 1977). Here a single score incorporates various areas of cognitive functioning. Three-score tests are divided into levels, and yield three different scores. For example, the Cognitive Abilities Test (CogAT) will yield scores measuring verbal, quantitative, and nonverbal abilities. Finally, multi-scored tests provide a wider profile about the testee, with the popular Differential Aptitude Tests (DAT) yielding nine different subtest scores.

GROUP ABILITY TESTS

What follows is an overview of some of the group intelligence and abilities tests that are currently in use and which fall into Nitk's (1983) categories. The tests covered are by no means an exhaustive listing, as a complete summary of group intelligence tests would become a large text. Rather, what is provided is a representative sampling of measures so that the reader will develop a perspective of the kinds of group tests that are in common use. Included in this review will be standard group intelligence measures along with tests that are less sensitive to cultural influences and specialized tests for college entry and employment.

Otis-Lennon Mental Ability Test (OLMAT):

One of the more popular group intelligence tests that is intended to provide a measure of "g," the general intellectual factor, is the OLMAT (Otis & Lennon, 1969). The OLMAT evolved from the Army Alpha Examination of World War I, as Arthur Otis was a contributor to both instruments. The current test reflects many characteristics of the Army Alpha, but it is more refined with regard to psychometric properties. The stated purpose of the test is to generate a comprehensive, carefully articulated assessment of general mental ability, or scholastic aptitude, of American students, through a battery of tests that provides scores from kindergarten through Grade 12 (ages 5–18 years). Items are hand-scored, and the scale has a mean of 100 (called a deviation IQ) and a standard deviation of 16. Despite the fact that an IQ is derived, the authors of the test appear to take a tentative stand on whether the test is, in fact, a measure of intelligence. At one point the test is called a measure of general mental ability, yet at a later point, the reader is informed that the tests do not measure native endowment. Regardless of whether group intelligence tests do measure intelligence, virtually all of the group measures report statistically significant reliability and validity data. The manual for the OLMAT provides extensive data on reliability and is based on samples in excess of 120,000; the

split-half and KR-20 coefficients range from 0.93 to 0.96. Reliability coefficients vary as a function of the age and grade level assessed. The test manual does not report validity data.

A further outgrowth of the Otis series of tests is the Otis-Lennon School Ability Test (OLSAT) (Otis & Lennon, 1977). Like its predecessor, the OLMAT, the OLSAT is a pencil-and-paper, multiple-choice test that is group administered and objectively scored. The test is a multilevel battery that is suitable for school settings (grades 1 through 12) and is designed to measure abstract thinking and reasoning ability. The purposes of the OLSAT are to assess the examinee's ability to cope successfully with school-learning tasks and use the results for placing students in classes. The focus on school learning dispenses with the potential interpretational problems that arise when terms such as *mental ability, intelligence,* or *mental maturity* are used. In fact, there is a change from Deviation IQ (DIQ) as used in the OLMAT to the School Ability Index (SAI) on the OLSAT. Nevertheless, the OLSAT, like the OLMAT, is designed to assess a verbal educational factor, and the SAI has the same psychometric properties as the DIQ. The OLSAT, as its predecessor, the OLMAT, is based on a defensible standardization procedure involving 200,000 students in 200 school districts (Swerdlik, 1992). Reliability estimates range from .84 to .95, depending on the level within the test that is assessed and the method of computing reliability. Validity coefficients range from .40 to .60, and these values are typical of well-constructed psychological tests. The Primary I level of the test consists of objects familiar to the child—ice cream cones, animals, stars, etc., and is thereby helpful in holding the child's attention.

Lorge Thorndike Intelligence Test

The Lorge Thorndike Intelligence Test (Lorge & Thorndike, 1966) is another popular group-administrated scale that is applicable to grades 3 through 13. This test measures abstract intelligence, and like the OLMAT, contains both verbal and nonverbal items. The verbal battery is made up of five subtests that include vocabulary, verbal classification, sentence completion, arithmetic reasoning, and verbal analogies. The nonverbal battery contains subtests of pictorial classifications, pictorial analogies, and numerical relationships.

The current edition, called the Multi-Level Edition, has more representative norms than the earlier Separate Level Edition. Validity estimates are readily established as the Lorge-Thorndike was normed on the same samples used for the Iowa Test of Basic Skills, a group administered test for grades 3 through 8 and the Tests of Academic Progress for grades 9 through 12. Correlations with school performance are typical of the various group tests, and in one case are reported to be .87 with reading and .76 with math. Moderate but significant correlations are found between the Lorge-Thorndike and the WISC and the Stanford Binet, and range from .54 to .77. The conglomerate of different types of verbal and nonverbal items appears to represent an attempt to assess "g" by utilizing some arrangement of tests that correlate with each other and which therefore are assumed to share a common global or general intellectual process.

Multidimensional Aptitude Battery (MAB)

The MAB (Jackson, 1984) has been administered to various normal and special populations, such as business people, high school and college students, prison inmates, and psychiatric patients. The proposed purpose of the test is the assessment of intellectual abilities. The MAB attempts to transfer the structure of the WAIS-R into a format suitable for group administration. The MAB can be administered either individually or in a group setting, and it consists of five verbal scales and five nonverbal performance scales. Each subtest has a time limit of seven minutes, which means that the entire battery can be administered in 90 minutes. Verbal, Performance, and Full Scale IQs have been calibrated to match those of the WAIS-R. Test-retest reliabilities for the MAB are .95 for Verbal IQ, .96 for Performance IQ, and .97 for Full-Scale IQ. The MAB has the advantage of standardized group administration without sacrificing reliability and validity, and it is generally seen as having strong psychometric properties.

Cognitive Abilities Test (CogAT)

The CogAT, Form 5 is based on the CogAT, Forms 1–4 (Thorndike & Hagen, 1986). and the Lorge-Thorndike Intelligence Test (Lorge & Thorndike, 1966), It can be used for Kindergarten through 12th grade. There are 10 different levels of

the test (Levels 1–2 and Levels A–H.) Each level of the test contains three separate batteries that yield separate scores for verbal-, quantitative- and nonverbal-reasoning abilities. A composite score is also computable. The standardization sample consisted of over 160,000 students from public, Catholic, and private non-Catholic schools. The CogAT is a popular and well-established test of educational aptitude that has undergone complete restandardization. It has strong psychometric properties with reliability and validity estimated from the .70s to .90s.

Henmon-Nelson Tests of Mental Ability

The Henmon-Nelson Tests of Mental Ability (Lamke, Nelson, & French, 1973) are group measures of mental ability that have four levels. There is a Primary Battery (Grades K–2), a battery for Grades 3–6, a battery for Grades 6–9, and a battery for grades 9–12. The Primary Battery requires approximately 30 minutes to administer, while the batteries for Grades 3–12 take exactly 30 minutes to administer.

Test items are centered around subjects related to academic functioning, such as vocabulary, sentence completion, opposites, general information, verbal analogies, verbal classification, verbal inference, number series, arithmetic reasoning, and figure analogies. No reading is required for the Primary Battery. The norms for the Henmon-Nelson Tests of Mental Ability were obtained by stratified random sampling of over 40,000 students in the years 1972–1973. Raw scores, Deviation IQ scores by age, age percentile ranks and stanines, and grade percentile ranks and stanines can be calculated.

The Culture-Fair Intelligence Test (CFIT)

The CFIT (R. B. Cattell, 1973) is a nonverbal measure of an individual's intelligence. This assessment instrument is designed to overcome the influences of verbal fluency, cultural background, and educational level. The CFIT is said to be unique in that it was designed to measure fluid abilities, whereas traditional tests stress the measurement of crystallized abilities. Thus, in theory, the CFIT allows an evaluation of the future potential of an individual, rather than assessing past achievements or lack of achievements.

The tests are of paper-and-pencil format and have time limits for each subtest. Scale 1 is intended for children aged 4–8 years and for retarded adults. This particular scale is not considered by the test authors to be group-administered or fully culture-free. Scale 2 is for ages 8 to 13 years, and for average adults, and Scale 3 is for high school students and superior adults. The participant's total working time is only 12 1/2 minutes, but the total administration time is closer to 30 minutes. The CFIT has been criticized for its lengthy instructions that cause children to lose attention and become bored. Another criticism is that bright adults with learning disabilities, particularly those with left-right reversal difficulties, are said to obtain low scores on this test (Vane & Motta, 1990) .

Internal consistency coefficients averaged across samples are: Scale 1, .91; Scale 2, .82; and Scale 3, .85. Test-retest reliabilities are: Scale 1, .80; Scale 2, .84; and Scale 3, .82. The CFIT correlates with other intelligence measures in the mid-.70 range. Despite this, several studies of the CFIT have produced mixed results. For example, it has been shown that there are only moderate correlations from .20 to .50 with scholastic achievement, although predictive validities have been fairly impressive for certain groups and criteria. Moreover, correlations with other intelligence tests are mostly between the .50 to .70 range, suggesting that the test is measuring the "g" factor. The CFIT has been administered to many culturally diverse groups outside of the United States and produces scores that are comparable between groups. Although the tests show somewhat lower correlations with socioeconomic status than culture-loaded or other primarily verbal tests, and some bilingual immigrant groups score higher on these tests than on conventional IQ tests, the CFIT does not greatly reduce the magnitude of score differences when administered to culturally disadvantaged groups.

Raven's Progressive Matrices

Another test that might be considered to be culture-fair or culturally reduced is the Raven's Progressive Matrices (Raven, 1941, 1981; Raven, Court, & Raven, 1983, 1985). There are two widely used versions of the test: the Standard and the Colored versions (Naglieri & Prewett, 1990). The test was introduced in 1938 and has gone

through many revisions. Because it is nonverbal, and in most situations requires little more than having the examinee point to the correct item, it is often used in situations where examiners want a measure of ability that is not biased by educational background or by cultural or linguistic deficiencies. All of the test items are composed of geometric figures that require the test taker to select among a series of designs the one that most accurately represents or resembles the one shown in the stimulus material. The test items are presented in graded levels of difficulty and there are test booklets for different age levels. Validity measures involving the correlation of the Raven Matrices with the Stanford-Binet and the Wechsler Scales range from .54 to .86 (Raven, Court, & Raven, 1983, pp. 8–9). The authors indicate that "the scales can be described as 'tests of observation and clear thinking,...By themselves they are not tests of 'general intelligence'....They should be used in conjunction with a vocabulary test" (p. 3). Despite this caution, the Progressive Matrices have been viewed as measures of intelligence and have been widely used in many countries to test military groups because they are considered to be independent of prior learning.

Test of Nonverbal Intelligence-2 (TONI-2):

The TONI-2 by Brown, Sherbenou, and Johnsen (1990) is a language-free intelligence test which does not require the examinee to read, write, speak, or listen. It can be used in small groups and is often given individually (Murphy, Conoley, & Impara, 1994). The test is intended to be useful for those who are bilingual, non-English speaking, or who have difficulty reading, writing, speaking, or hearing. The items require the subject to decide how several figures are related by choosing from four to six alternatives, and to indicate which one of these goes best with three or more of the presented stimuli. The figures are black-and-white line drawings; some are simple geometric figures; others are more abstract. The examinee is required to point to the correct response. The TONI-2 is appropriate for individuals aged 5–86 years, and requires 15 minutes to administer in either group or individual format. According to Naglieri and Prewett (1990), "The primary ability assessed by the TONI is problem solving" (p. 359). The national standardization sample consisted of 2,500 persons. There are two equivalent forms of the TONI-2; each contains a variety of problem-solving tasks presented in ascending order. A "TONI quotient" is yielded with a mean of 100 and a standard deviation of 15.

Expressive One-Word Picture Vocabulary Test: Upper Extension (EOWPVT-U)

This vocabulary test was designed to be an upward-age extension of the Expressive One-Word Picture Vocabulary Test-Revised (EOWPVT-R) (Gardner, M.F., 1983), which is individually administered to individuals aged 2–12 years. The EOWPVT-U (Gardner, 1990) is used for individuals aged 12 to 16 years and can be group or individually administered. It was developed by psychologists, counselors, physicians, learning specialists, speech therapists, social workers, diagnosticians, and other professionals. It is said to provide a valid and readily obtainable assessment of a student's verbal intelligence. The testee is shown a stimulus picture and is required to demonstrate his or her ability to understand and use words by naming simple objects or providing vocabulary words for abstract concepts that are illustrated in pictures. The EOWPVT-U can be administered individually in an oral-testing situation, or in a group format by having students write their responses. The EOWPVT-U provides mental ages, percentiles, and stanines, and deviation-IQ scores are calculated.

Black Intelligence Test of Cultural Homogeneity (BITCH)

At the other end of the continuum, away from tests which claim they are culture-free or -fair, are those tests that are designed to reflect a unique knowledge of a given culture. An example of this type of scale is the BITCH Culture Specific Test by Williams (1972). The author of this test set out to demonstrate that one's performance on a test can be affected by cultural experience. The test contains a vocabulary that reflects African-American slang, and as might be expected, African Americans score significantly better on it than do whites. The test's value probably is that it illustrates the extent to which one's performance can vary as a result of prior knowledge; its major drawback is that it does not correlate with known measures of intelligence (Matarazzo & Wiens, 1977). The latter finding can, of course, be countered by the notion

that one would not expect to find a correlation between popular standardized intelligence tests that are considered by some to be culturally biased in terms of the dominant culture and one that is biased in favor of a minority group.

The Draw-A-Person (DAP) Test (A Quantitative Scoring System)

The DAP Test (A Quantitative Scoring System) (Naglieri, 1988) is a recently published system of scoring human-figure drawings "to obtain an estimate of ability" (Naglieri & Prewett, 1990, pg. 363). Although administered individually, it can also be administered in groups. There are 64 items comprising a rating scale for the drawings of a man, a woman, and one's self. The drawings are rated according to the number of body parts that are drawn and the extent to which these parts are elaborated and detailed. Also of importance in the scoring system is the proportion of the parts of the body to one another, and the manner in which these parts are connected. The DAP Test is intended to be used as a part of the larger group of tests or for screening purposes. Since the test is nonverbal and easily administered, the influences of verbal skills, primary language, fine motor coordination, cultural diversity, and language disabilities are said to be reduced. (Naglieri & Prewett, 1990, pg. 364).

The DAP Test was normed on 2,622 individuals aged 5–17 years. According to Naglieri (1988) the normative sample is representative of the 1980 U.S. census data on the stratification variables of age, sex, race, geographic region, ethnic group, socioeconomic status, and community size. The test yields possible scores for each of the three drawings (man, woman, self). These three raw scores are then combined to form a total test score, which is then converted to a standard score with a mean of 100 and a standard deviation of 15. Percentile ranks, age-equivalent scores, and confidence intervals are available for the individual drawings and for the DAP Total Test Score.

The use of the DAP Test as a measure of nonverbal ability or intelligence has not gone without criticism. Motta, Little, and Tobin (1993), for example, point out that the DAP Test has doubtful validity because it correlates weakly with more established measures of intellectual functioning. These authors suggest that the DAP Test has little use in psychodiagnostics because of its psychometric weaknesses and that it should be used only for rapport-building purposes. Similar doubts regarding the utility of drawings in psychological assessment have been voiced by other researchers (e.g., Oakland & Dowling, 1983; Phil & Nimrod, 1976; Weerdenburg & Janzen, 1985). Despite these concerns, many continue to use figure drawings to assess intellectual functioning.

The Scholastic Aptitude Test (SAT)

In testing for college entrance, one test dominates the field: the College Entrance Examination Board's Scholastic Aptitude Test (SAT). This test is given by the College Board to all high school students throughout the nation who wish to take it. Many selective colleges require SAT scores, but in many colleges scores are only one of the factors used in admitting students. Most colleges, however, do have minimum SAT-cutoff scores.

The SAT is a paper-and-pencil test containing 150 multiple-choice items, with five choices each. There is a verbal section involving reading comprehension, antonyms, verbal analogies, and sentence completion. The mathematics section consists of numerical and quantitative-reasoning items, but does not tap formal mathematical knowledge per se. The SAT would undoubtedly load heavily on the "g" factor in any factor-analytic study that included other mental-ability tests (Vane & Motta, 1984, 1990). The verbal section has been found to correlate higher than the quantitative score with college grade-point average.

The validity of the SAT in predicting scores of minority-group students has frequently been challenged. For example, Stanley and Porter (1967), conducted a study that involved students in three African-American, coeducational, four-year state colleges and compared them with students in 15 predominantly white state colleges in Georgia. Correlations of the combined scores with freshman grade-point averages was .72 for white females, .63 for African-American females, and .60 for both white and African-American males, suggesting that the prediction for white females is better than for the other three groups. Several studies have shown that high school grade-point averages predict college grade-point averages better than the SAT for whites, but not for African -Americans (Cleary, 1968; McKelpin, 1965; Munday, 1965; Peterson, 1968).

The Wonderlic Personnel Test (WPT)

In the area of employment, tests have been used in making employment decisions in the United States for over 70 years. Although there are many content-validated job-knowledge tests and job-sample tests such as typing tests, the most commonly used measures have been measures of cognitive skill, called either aptitude or ability tests. According to Schmidt and Hunter (1981), who performed a meta-analysis of a large number of studies in the field of employment testing, the results show that:

> professionally developed cognitive ability tests are valid predictors of performance on the job and in training for all jobs in all types of settings...[and that] cognitive ability tests are equally valid for minority and majority applicants and are fair to minority applicants in that they do not underestimate the expected job performance of minority groups. (p.1128)

Schmidt and Hunter (1981) reported results of a study of 370,000 clerical workers, which showed that validity of seven cognitive abilities was essentially constant across five different clerical-job families. All seven abilities were highly valid in all five job families.

The WPT (Wonderlic, 1977) is one of the group intelligence tests designed for use in employment selection. It can be administered individually or in a group setting. The author intentionally uses the term *personnel* rather than *intelligence* to reduce the anxiety of those who must take the test. Despite this, the test manual clearly indicates that the intended use of the instrument is to assess mental ability so that a suitable match can be made between the applicant's ability and the ability demanded for a particular job area. Dodrill and Warner (1988) found a high degree of correspondence between the WAIS and the WPT in an evaluation of psychiatric, neurological, and normal participants. They conclude that their results "point to the Wonderlic as a measure of general intelligence" (p.146)

The WPT is administered in only 10–12 minutes. There are 50 questions, which examinees usually do not finish; test items require the examinee to reason in terms of words, numbers, and symbols, and to use ideas when thinking. The test items include vocabulary, sentence rearrangement, logic, arithmetic, and interpretation of proverbs (e.g., a rolling stone gathers no moss). Reliabilities range from .82 to .94. The test has 14 different forms and has been standardized in business situations, using large numbers of people and test sites. Minimum scores are reported for professions ranging from custodian to administrator and executive. The score reported is the number correct instead of an IQ, thus reducing some of the controversy evoked by the latter term. Norms are available based on sex, age, range, and educational level. The test reportedly correlates .91–.93 with the WAIS Full-Scale IQ.

Drawbacks of the WPT are that reading skill is required to take the test and speed is a factor. As a result, it would penalize those with psychomotor deficits or reading deficiencies. Because the test provides only a single score, it may not be as diagnostically useful as longer tests. These disadvantages are offset by the obvious benefits of a reliable and valid group measure of intelligence that is easily administered and scored. If used in the right context, it is a valuable test.

SUMMARY

Group and individual assessment of human abilities have a long, colorful, and often emotional history. Many theoretical perspectives regarding the nature of intelligence have been put forward, and numerous assessment devices have been developed, and continue to be developed, for the purpose of assessing abilities. Most of the group intelligence measures being used today correlate significantly with individually administered intelligence tests, and both have proven their utility in the areas of education and employment. Questions, such as the nature of intelligence, whether greater emphsis should be placed on practical measures of ability, or whether existing measures can fairly assess minority groups or members of other cultures, remain embroiled in controversy and will continue to be debated as they have been in the past. Nevertheless, group intelligence tests provide an economical way to readily assess large numbers of individuals, and for this reason, will continue to be useful tools in helping to make personnel decisions.

REFERENCES

Aiken, L. R. (1976). *Psychological testing and assessment* (2nd ed.). Boston: Allyn & Bacon, Inc.

Allen, B. & Boykin, W. (1992). African-American children and the educational process: Alleviating cultural discontinuity through prescriptive pedagogy. *School Psychology Review, 21*(4), 586–596.

Barrett, G. V. & Depinet, R. L. (1991). A reconsideration for testing for competence rather than intelligence. *American Psychologist, 46,* 1012–1024.

Binet, A., & Simon, T. (1916). *The development of intelligence in children.* (E.S. Kit, Trans.). Baltimore: Williams & Wilkins. (Original work published 1905)

Boykin, A. W. (1986). The triple quandary and the schooling of Afro-American children. In U. Neisser (Ed.), *The school achievement of minority children: New perspectives* (pp. 57–92). Hillsdale, NJ: Lawrence Earlbaum.

Brigham, C. C. (1923). *A study of American intelligence.* Princeton NJ: Princeton. University Press.

Brown, L., Sherbenou, R. J., Johnsen, S. K. (1990). *Test of non-verbal intellingence (Toni-2)* (2nd ed.) Austin, TX: Pro-Ed, Inc.

Campione, J. C., & Brown, A. L. (1978). Toward a theory of intelligence: Contributions from research with retarded children. *Intelligence, 2,* 279–304.

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology, 54,* 1–22.

Cattell, R. B. (1973). *Cattell Culture Fair Intelligence Tests.* Champaign, IL: Institute for Personality and Ability Testing.

Chon, M. (1995). The truth about Asian Americans. In R. Jaccoby & N. Glauberman (Eds.), *The Bell Curve Debate* (pp. 238–240). New York: Random House.

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and white students in integrated colleges. *Journal of Educational Measurement, 5,* 115–124.

Cronbach, L. (1975). Five decades of public controversy over mental testing. *American Psychologist, 3,* 1–14.

Das, J. P (1973). Structure of cognitive abilities: Evidence for simultaneous and successive processing in children. *Journal of Educational Psychology, 65,* 103–108.

Dodrill, C. B., & Warner, M. H. (1988). Further studies of the Wonderlic Personnel Test as a brief measure of intelligence. *Journal of Consulting and Clinical Psychology, 56,* 145–147.

Eysenck, H. J. (1971). *Race, intelligence and education.* London: Temple Smith.

Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin, 101,* 171–191.

Freeman, F. N. (1922). A referendum of psychologists. *Century Illustrated Magazine, 107,* 237–245.

Frisby, C. L. (1995). When facts and orthodoxy collide: The Bell Curve and the robustness criterion. *School Psychology Review, 24,*12–19.

Galton, F. (1870). *Hereditary genius: An inquiry into its laws and consequences.* London: D. Appleton.

Garber, H., & Austin G. R. (1982). Learning, schooling, scores: A continuing controversy. In H. Garber and G. R. Austin (Eds.), *The rise and fall of national test scores* (pp.1–8). New York: Academic Press.

Gardner, H. (1995). Scholarly brinksmanship. In R. Jaccoby & N. Glauberman (Eds.), *The bell curve debate* (pp. 61–72). New York: Random House.

Gardner, H. (1983). *Frames of mind: The theory of multiple intelligences.* New York: Basic Books.

Gardner, M. F. (1983). *Expressive One-Word Picture Vocabulary Test-Revised.* Novato, CA: Academic Therapy Publications.

Gardner, M. F. (1990) *Expressive One-Word Picture Vocabulary Test: Upper Extension* (1983). Novato, CA: Academic Therapy Publications.

Gordon, E. W., & Terrell, M. D. (1981). The changed social context of testing. *American Psychologist, 36,* 1167–1171.

Gordon, H. (1923). Mental and scholastic tests among retarded children. *Board of Education Pamphlet,* 1023, No.44. London: HMSO.

Gould, S. J. (1995). Mismeasure by any measure. In R. Jaccoby & N. Blauberman (Eds.), *The Bell Curve Debate* (pp. 3–13). New York: Random House.

Guilford, J. P. (1967). *The nature of human intelligence.* New York: McGraw-Hill.

Haney, W. (1981). Validity, vaudeville and values. *American Psychologist, 36,* 1021–1034.

Hargadon, F. (1981). Tests and college admissions. *American Psychologist, 14,* 469–479.

Hebb, D. O. (1949). *The organization of behavior.* New York: John Wiley.

Herbert, R. (1995). Throwing a curve. In R. Jaccoby & Naomi Blauberman (Eds.), *The Bell Curve Debate* (pp. 249–251). New York: Random House.

Hernstein, R. J. (1973). *IQ in the meritocracy.* Boston: Little, Brown.

Hernstein, R. J., & Murray, C. (1994). *The bell curve: Intelligence and class structure in American life.* New York: The Free Press.

Hirsch, N. D. M. (1928). An experimental study of the East Kentucky mountaineers: A study in heredity and environment. *Genetic Psychology Monographs, 3,* 183–244.

Horn, J. L. (1985). Remodeling old models of intelligence. In B. Wolman (Ed.), *Handbook of intelligence* (pp.267–300). New York: Wiley.

Jackson, D. (1984). *Multidimensional Aptitude Battery.* Port Huron, MI: Research Psychologists Press, Inc.

Jensen, A. R. (1967). The culturally disadvantaged: Psychological and educational aspects. *Educational Research, 10,* 4–20.

Jensen, A. R. (1969). How much can we boost IQ and scholastic achievement? *Harvard Educational Review, 39,* 1–123.

Jensen, A. R. (1973). *Educability and group differences.* New York: Harper & Row.

Jensen, A. R. (1980). *Bias in mental testing.*New York: Free Press.

Jensen, A. R. (1995). Paroxysms of denial. In R. Jaccoby & Naomi Glauberman (Eds.), *The Bell Curve Debate* (pp. 335–337). New York: Random House.

Kamin, L. J. (1995). Lies, damned lies, and statistics. In R. Jaccoby & N. Glauberman (Eds.), *The Bell Curve Debate* (pp. 81–105). New York: Random House.

Lamke, T. A., Nelson, M. J., & French, J. L. (1973). *Henmon-Nelson Tests of Mental Ability.* Boston: Houghton Mifflin Co.

Lippman, W. (1922a). Tests of hereditary intelligence. *New Republic, 32,* 213–215.

Lippman, W. (1922b). Tests of hereditary intelligence. *New Republic, 32,* 328–330.

Lippman, W. (1923). Rich and poor, girls and boys. *New Republic, 34,* 295–296.

Lorge, I., & Thorndike, R. L. (1966). *Lorge-Thorndike Intelligence Tests.* Chicago, IL. Riverside Publishing Co.

Matarazzo. J. D., & Wiens, A. N. (1977). Black Intelligence Test of Cultural Homogeneity and Wechsler Adult Intelligence Scale scores of Black and White Police Applicants. *Journal of Applied Psychology, 62,*. 57–63.

McClelland, D. C. (1973). Testing for competence rather than for "intelligence." *American Psychologist, 28,* 1–14.

McKelpin, J. C. (1965). Some implications of intellectual characteristics of freshmen entering a liberal arts college. *Journal of Educational Measurement, 2,* 161–166.

Miller, A. (1995). Professors of hate. In R. Jaccoby & N. Glauberman (Eds.), *The Bell Curve Debate* (pp. 162–178). New York: Random House.

Motta, R. W., Little, S. G., & Tobin, M. I. (1993). The use and abuse of human figure drawings. *School Psychology Quarterly, 8,* 162–176.

Munday, L. (1965). Predicting college grades in predominantly Negro colleges. *Journal of Educational Measurement 2,* 157–160.

Murphy, L. L., Conoley, J. C., & Impara, J. C. (1994). Test of Nonverbal Intelligence (2nd ed.). In L. L. Murphy, J. C. Conoley, & J. C. Impara (Eds.), *Tests in Print IV, Vol.2* (p. 890). Lincoln, NE: University of Nebraska Press.

Myrdal, G., Sterner, R., & Rose, A. (1962). *An American dilemma: The negro problem and modern democracy.* (2nd ed.). New York: John Wiley & Sons.

Naglieri, J. A. (1988). *Draw A Person: A quantitative scoring system.* New York: Psychological Corporation.

Naglieri, J. A., & Prewett, P. N. (1990). *Handbook of psychological and educational assessment of children's intelligence and achievement.* (R. W. Kamphaus & C. R. Reynolds, Eds.) (pp. 348–370. New York: Guilford.

Neisser, U. (1976). General, academic, and artificial intellignece, In L. Resnick (Ed.), *Human intelligence: Perspectives on its theory and measurement* (pp. 179–189). Norwood, NJ: Ablex.

Newman, H. H., Freeman, F. N., and Holzinger, K. J. (1937). Twins: A study of heredity and environment. Chicago: University of Chicago Press.

Nisbett, R. (1995). Dangerous, but important. In R. Jaccoby & N. Glauberman (Eds.), *The Bell Curve Debate* (pp. 110–114). New York: Random House.

Nitk, A. J. (1983). *Educational tests and measurement: an introduction.* New York: Harcourt Brace Jovanovich, Inc.

Oakland, R., & Dowling, L. (1983). The Draw-A-Person Test: Validity properties for nonbiased assessment. *Learning Disabilities Quarterly,* 534.

Otis, A. S., & Lennon, R. T. (1969). *Otis-Lennon Mental Ability Test.* New York: Harcourt, Brace, & World.

Otis, A. S., & Lennon, R. T. (1977). *Otis-Lennon School Ability Test.* New York: Harcourt Brace Jovanovich.

Peterson, R. E. (1968). Predictive validity of a brief test of academic aptitude. *Educational and Psychological Measurement, 28,* 441–444.

Phil, R. O., & Nimrod, G. (1976). The reliability and validity of the Draw-A-Person Test in IQ and personality assessment. *Journal of Clinical Psychology, 32,* 470–472.

Pintner, R. (1923). *Intelligence testing.* New York: Holt, Rinehart & Winston.

Raven, J. C. (1941). Standardization of progressive matrices. *British Journal of Medical Psychology.* 19, 137–150.

Raven, J. C. (1981). *Manual for Raven's Progressive Matrices and Mill Hill Vocabulary Scales* (Research Suppl. 1). London: H. K. Lewis.

Raven, J. C., Court, J. H., & Raven, J. (1983). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Standard Progressive Matrices,* (Sect 1.). London: H. K. Lewis.

Raven, J. C., Court, J.H., & Raven, J. (1985). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Standard Progressive Matrices* (Sect 3.). London: H.K. Lewis.

Reshley, D. J., Kicklighter, R., & McKee, P. (1988). Recent placement litigation: Part III. *School Psychology Review, 15,* 39–50..

Richardson, R. Q. (1995). The window dressing behind *The Bell Curve. School Psychology Review, 24,* 42–44.

Ryan, A. (1995). Apocalypse now? In R. Jaccoby & N. Glauberman (Eds.), *The bell curve debate* (pp. 14–29). New York: Random House.

Sattler, J. M. (1988). *Assessment of Children* (3rd. ed). San Diego: J. M. Sattler.

Schmidt, F. L., & Hunter, J. E. (1981). Employment testing: Old theories and new research findings. *American Psychologist, 36,* 1128–1137.

Shockley, W. (1971). Negro IQ deficit: Failure of a "malicious coincidence" model warrants new research proposals. *Review of Educational Research, 41,* 227–248.

Shuey, A. M. (1958). *The testing of Negro intelligence.* New York: Social Science Press.

Shuey, A. M. (1966). *The testing of Negro intelligence.*(Rev. ed.) New York: Social Science Press.

Spearman, C. E. (1923). *The nature of intelligence and the principles of cognition.* London: Macmillan.

Stanley, J. C., & Porter, A. C. (1967). Correlation of scholastic aptitude test scores with college grades for Negroes versus whites. *Journal of Educational Measurement, 4,* 199–218.

Sternberg, R. J. (1986). *Intelligence applied: Understanding and increasing your intellectual skills.* San Diego: Harcourt Brace.

Sternberg, R. J., Wagner, R. K., Williams, W. M., Horvath, J. A. (1995). Testing common sense. *American Psychologist, 50,* 912–926.

Swerdlik, M. E. (1992). Review of Otis Lennon school ability test, sixth edition. In J. J. Krammer & Jane C. Conoley (Eds.). *The eleventh mental measurements yearbook.* (pp. 635–639).

Terman, L. M. (1922a). The great conspiracy. *New Republic, 33,*116–120.

Terman, L. M. (1922b). The psychological determinist: Or democracy and the IQ. *Journal of Educational Research, 6,* 57–62.

Thorndike, E. L. (1927). *The measurement of intelligence.* New York: Bureau of Publications, Teachers College, Columbia University.

Thorndike, R. L., & Hagen, E. (1986). *Cognitive abilities test, forms 1–4.* Chicago, IL: The Riverside Publishing Co.

Thurstone, L. L. (1938). Primary mental abilities. *Psychometric Monographs,* Vol 1.

Vane, J. R., & Motta, R. W. (1984). Group intelligence tests. In G. Goldstein & M. Hersen (Eds.), *Handbook of psychological assessment* (pp. 100–116). Elmsford, New York: Pergamon Press.

Vane, J. R., & Motta, R. W. (1990). Group intelligence tests. In G. Goldstein & M. Hersen (Eds.), *Handbook of psychologcial assessment* (2nd ed., pp. 102–119). Elmsford, New York: Pergamon Press.

Vernon, P. E. (1950). *The structure of human abilities.* New York: Wiley.

Vernon, P. E. (1978). Intelligence: Hereditary and environment. San Francisco: W. H. Freeman and Co.

Wechsler, D. (1958). *The measurement and appraisal of adult intelligence* (4th ed.). Baltimore: Williams and Wilkins.

Weerdenburg, G., & Janzen, H. L. (1985). Predicting grade success with a selected kindergarten screen battery. *School Psychology International,* 13–23

Williams, R. L. (1972). *The BITCH Test (Black Intelligence Test of Cultural Homogeneity).* St. Louis, MO: Black Studies Program, Washington University.

Wonderlic, E. F. (1977). *Wonderlic Personnel Test.* Northfield, IL: E. F. Wonderlic and Associates.

Woodworth, R. S. (1941). *Hereditary and environment.* New York: Social Science Research Council.

Yerkes, R. M. (1923). Testing the human mind. *Atlantic Monthly, 131,* 358–370.

PART IV

ACHIEVEMENT, APTITUDE, AND INTEREST

This Page Intentionally Left Blank

CHAPTER 7

ACHIEVEMENT TESTING

Lynda J. Katz
Gregory T. Slomka

INTRODUCTION

The second edition of the *Handbook of Psychological Assessment* appeared in 1990 and included a chapter on "Achievement Testing" by these authors. Developments and trends in the field are updated in this chapter.

During the mid-1970s the use of standardized tests among a variety of elementary, secondary, and post-secondary educational programs came under severe criticism. The use of standardized achievement tests in particular involved over 80 percent of American school children, with some of these children taking 26 achievement tests during a school career (National School Boards Association, 1977). And yet, as recent as 1992, 80 percent of the state system-wide tests given to some 14.5 million students were achievement tests. It has been postulated that the nation-wide concern with the use of standardized tests resulted from competition among the "baby boom" generation children of the late 1940s and early 1950s for "scarce slots in the choicest schools and businesses," so that their stakes of doing well or poorly on tests went up. Second, those same baby boomers were looking back on their experiences with years of taking standardized tests and were very sensitive to the perceived abuses of such testing (Strenio, 1981, p. xviii). The main criticisms of these tests, however, have centered around the equality of the tests themselves; the use to which they are put; the behavior of the testing industry, with some 40 to 50 test publishers responsible for 90 percent of the tests used in the country today (Haney, Madaus, & Lyons, 1993); and the consequences for society of the misuse of these tests. In addition, major court cases and federal legislation for exceptional children have addressed specifically the use of tests and testing as part of the overall assessment process (*Larry P. v. Riles* and P.L. 94-142), again in response to these same criticisms.

In November 1975, at a conference on testing sponsored by the National Association of Elementary School Principals and the North Dakota Study Group on Evaluation, 25 national organizations, including the U.S. Office of Education, drafted the following statement:

> We believe that the public, and especially educators, parents, and children, need fair and effective assessment processes that can be used for diagnosing and prescribing for the needs of individual children....
>
> In regard to standardized achievement tests, we have agreed on the following recommendations:
>
> 1. The profession needs to place a high priority on developing and putting into wide use new processes of assessment that are more fair and effective than those currently in use and that more adequately consider the diverse talents, abilities, and cultural backgrounds of children.
> 2. Parents and educators need to be much more actively involved in the planning and processes of assessment.
> 3. Any assessment results reported to the public must include explanatory material that details the limitations inherent in the assessment instruments used.

4. Educational achievement must be reported in terms broader than single-score national norms, which can be misleading.
5. Information about assessment processes should be shared among the relevant professions, policy makers, and the public so that appropriate improvements and reforms can be discussed by all parties.
6. Every standardized test administered to a child should be returned to the school for analysis by the teachers, parents, and child.
7. Further, the standardized tests used in any given community should be made publicly available to that community to give citizens an opportunity to understand and review the tests in use.
8. The professions, the public, and the media need to give far greater consideration to the impact of standardized testing on children and young people, particularly on those below the age of ten.
9. A comprehensive study should be conducted on the actual administration and use of standardized tests and the use of test scores in the schools today. (National School Boards Association, 1977, p. 18).

In 1983 *A Nation at Risk* was published, one of the most widely publicized education reforms reports of the 1980s. In that report the authors warned "the educational foundations of our society are presently being eroded by a rising tide of mediocrity that threatens our very future as a nation and a people" (p. 5).

The National Commission on Excellence in Education went on to recommend that "standardized tests of achievement (not to be confused with aptitude tests)...be administered at major transition points from one level of schooling to another and particularly from high school to college or work" (p. 28). The purposes of testing would be to certify a student's credentials, identify needs for remedial instruction and identify opportunities for accelerated work.

In 1990, U.S. President Bush and the National Governors Association announced the "America 2000" strategy for educational reform (National Education Goals Panel, 1990). That reform called for new achievement tests in the core subjects of English, mathematics, science, history, and geography. These tests were to differ from traditional norm-referenced assessments and focus instead on problem solving or task performance. The author of one recent review article has suggested that over-reliance on multiple-choice tests in the 1980s "led teachers to emphasize tasks that would reinforce rote learning and sharpen test-taking skills, and discouraged curricula that promote complex thinking and active learning (Wells, 1991, p. 55).

In addition, the Individuals with Disabilities Act (IDEA) of 1990 (Amendments to the Education for all Handicapped Children Act of 1975) called specifically for nondiscriminatory testing and multidisciplinary assessment (Hardman, Drew, Egan, & Wolf, 1993) for children with disabilities, explicitly supporting a major role for testing in the Individual Educational Plan (IEP). It has been estimated that between 8 and 20 million tests were used for special-education testing alone in the late 1980s (Haney, Madaus & Lyons, 1993). These estimates are based on 4.4 million students aged 3 to 21 years who served in special education programs in elementary and secondary schools between 1984 and 1985 (Snyder, 1987), and for whom an average of five to ten tests were used for initial assessment and one to two tests were used at least every three years thereafter. The majority of these tests were tests of achievement.

Thus, it remains both relevant and timely more than a decade later to

- review the historical development, classification, and psychometric properties of traditional achievement tests;
- update their status and use in terms of contemporary educational and clinical research and practice;
- consider the relationship of achievement testing to ecological and sociocultural variables and their use with special population groups; and
- take a futuristic look at the impact of modern computer technology on test construction and utilization.

Such a discussion may determine whether recommendations made 20 years ago regarding the use of achievement tests have been or will continue to need to be addressed.

Historical Development of Achievement Tests

The standardized objective achievement test based on a normative sample was first developed by Rice in 1895. His spelling test of 50 words (with alternate forms) was administered to 16,000 students in grades 4 through 8 across the country. Rice went on to develop tests in arithmetic and language, but his major contribution was his objective

and scientific approach to the assessment of student knowledge (DuBois, 1970). Numerous other single-subject-matter achievement tests were developed in the first decade of the twentieth century, but it was not until the early 1920s that the publication of test batteries emerged; in 1923, the Stanford Achievement Test at the elementary level, and in 1925, the Iowa High School Content Examination (Mehrens & Lehmann, 1975). Since the 1940s, there has been a movement toward testing in broad areas as well, such as the humanities and natural sciences rather than in specialized, single-subject-matter tests. Moreover, attention has been directed toward the evaluation of work-study skills, comprehension, and understanding, rather than factual recall per se. In the 1970s, standardized tests were developed that were keyed to particular test books, the use of "criterion-referenced" tests (CRTs) emerged (their dissimilarity from norm-referenced tests will be addressed in the next section), and the development of "tailored-to-user specifications" tests (Mehrens & Lehmann, 1975, p. 165) was initiated.

Early in the 1990s, the literature on achievement testing was concerned with latent-trait theory, item-response curves, and an assessment of learning achievement that is built into the instructional process. With the later 1990s, concerns have tended to focus on the intrinsic nature of the achievement test itself. Computer-adaptive testing is not the computerization of standardized norm-referenced paper-and-pencil tests but a radically different approach. The approach is based on a concept of a continuum of learning and where a particular child fits on that continuum so that his or her experience with testing is one of success rather than failure.

In addition to computer-adapted testing, the use of alternative assessment tools has taken a front-row seat (Improving America's Schools, Spring, 1996). This performance based assessment approach involves testing methods that require students to create an answer or product that demonstrates knowledge or skill (open-ended or constructed-response items, presentations, projects or experiments, portfolios). As Haney & Madaus (1989) have pointed out, these alternatives to multiple-choice tests are not new; and in fact, multiple-choice testing replaced these alternative forms of assessment in the late 19th and early 20th centuries because of the expense involved, the difficulties with standardization, and their use with large numbers of people. To appreciate fully this dramatic shift in the conceptualization of the assessment of achievement, it is first necessary to understand (a) the nature of tests which fall under the domain of achievement; (b) the psychometric underpinnings of achievement tests; (c) the basis for criterion-referenced as opposed to norm-referenced measurement; and (d) special issues which arise when achievement tests are used for particular purposes.

Classification of Achievement Tests

Achievement tests have generally been categorized as single-subject tests, survey batteries, or diagnostic tests and further dichotomized as group- or individually administered tests. Reference to the *Ninth Mental Measurement Yearbook* (Mitchell, 1985) reveals the prevalence of multitudinous published objective tests, and elsewhere it has been reported that some 2,585 standardized tests are in use (Buros, 1974). Table 7.1 is a listing of the most commonly used achievement tests. They have been categorized as (a) group administered, (b) individually administered, and (c) modality-specific tests of achievements, which can be either group or individually administered.

Typically one administers achievement tests in order to obtain an indication of general academic skill competencies or a greater understanding of an individual's performance in a particular area of academic performance. In this regard achievement tests are specifically designed to measure "degree of learning" in specific content areas. There are several distinct applications of achievement tests which vary as a function of the setting in which they are applied. Tests such as the Metropolitan Achievement Tests, Stanford Achievement Tests, California Achievement Tests, and Iowa Tests of Basic Skills represent instruments that typically consist of test-category content in six or more skill areas. The benefit of the battery approach is that it permits comparison of individual performances across diverse subjects. Because all of the content areas are standardized on the same population, differences in level of performance among skill areas can reflect areas of particular strength or deficit. Many of these instruments provide a profile as well as a composite score that allows ready comparison of levels of performance between tests. The representative content of these batteries typically includes core assessment of language, reading, and mathematics abilities. The extensiveness of the coverage of allied curricula, that is, science,

Table 7.1. Commonly Used Achievement Tests

Group Administered Achievement Tests	
California Achievement Tests	CTB/McGraw Hill (1984). California Achievement Tests. Monterey, CA: Author.
Iowa Test of Basic Skills	Hieronymus, E. F., Lindquist, H. D., & Hoover, D., et al. (1978). Iowa Test of Basic Skills. Chicago: Riverside Printing.
Metropolitan Achievement Test	Balow, I. H., Farr, R., Hogan, T. P., & Prescott, G. A. (1978). Metropolitan Achievement Tests (5th ed.). Cleveland, OH: Psychological Corporation.
Stanford Achievement Test	Gardner, E. G., Rudman, H. C., Karlson, B., & Merwin, J. C. (1982). Stanford Achievement Test. Cleveland, OH: Psychological Corporation.
SRA Achievement Services (SRA)	Naslond, R. A., Thorpe, L. P. & Lefever, D. W. (1978). SRA Achievement Series, Chicago: Science Research Associates.
Individually Administered Achievement Tests	
Basic Achievement Skills Individual Screener (BASIS)	Psychological Corporation (1983). Basic Achievement Skills Individual Screener. San Antonio: Author.
Kaufman Test of Educational Achievement	Kaufman, A. S., & Kaufman, N. G. (1985). Kaufman Test of Individual Achievement, Circle Pines, MN: American Guidance Service.
Peabody Individual Achievement Test-Revised	Markwardt, F. C. (1989). Peabody Individual Achievement Test. Circle Pines, MN: American Guidance Services.
Wide Range Achievement Test 3	Wilkinson, G. S. (1993). Wide Range Achievement Test 3. Wilmington, DE: Jastak Associates.
Woodcock Johnson Psychoeducational Battery-Revised	Woodcock, R. W. (1989). Woodcock Johnson Psychoeducational Battery-Revised: Technical Report. Allen, TX: DLM Teaching Resources.
Modality Specific Achievement Tests	
Reading	
Classroom Reading Inventory	Silvaroli, N. J. (1986). Classroom Reading Inventory (5th ed.). Dubuque, IA: Wm. C. Brown.
Diagnostic Reading Scales	Spache, G. D. (1981). Diagnostic Reading Scales. Monterey, CA: CTB/McGraw-Hill.
Durrell Analysis of Reading Difficulty	Durrell, D. D., & Catterson, J. H. (1980). Durrell Analysis of Reading Difficulty (3rd ed.). Cleveland, OH: Psychological Corporation.
New Sucher-Allred Reading Placement Survey	Sucher, F., & Allred, R. A. (1981). New Sucher-Allred Reading Placement Inventory. Oklahoma City: Economy Company.
Gates-MacGinitie Reading Tests	MacGinitie, W.H., et al. (1978). Gates-MacGinitie Reading Tests. Chicago: Riverside Publishing.
Gray Oral Reading Tests	Wiederholt, J. L., Bryant, B. R. (1992). Gray Oral Reading Tests, Third Edition. Austin, TX: Pro-Ed.
Nelson-Denny Reading Test	Brown, J.I., Fishco, V.V., & Hanna, G. (1993). Nelson-Denny Reading Test. Chicago: Riverside Publishing Co.
Stanford Diagnostic Reading Test	Karlson, B., Madden, R., & Gardner, E. F. (1976). Stanford Diagnostic Reading Test (1976 ed.). Cleveland, OH: Psychological Corporation.
Woodcock Reading Mastery Tests-Revised	Woodcock, R. W. (1987). Woodcock Reading Mastery Tests-Revised. Circle Pines, MN: American Guidance Service.

(*continued*)

Table 7.1. (Continued)

Mathematics	
Enright Diagnostic Inventory of Basic Arithmetic Skills	Enright, F. E. (1983). Enright Diagnostic Inventory of Basic Arithmetic Skills; North Billerica, MA: Curriculum Associates.
Keymath Revised	Connolly, A. J. (1988). Keymath Revised. A Diagnostic Inventory of Essential Mathematics. Circle Pines, MN: American Guidance Service.
Sequential Assessment of Mathematics Inventories	Reisman, F. K. (1985). Sequential Assessment of Mathematics Inventories, San Antonio, TX: Psychological Corporation.
Stanford Diagnostic Mathematics Test	Beatty, L. S., Madden, R., Gardner, E. G., & Karlsen, B. (1976). Stanford Diagnostic Mathematics Test. Cleveland, OH: Psychological Corporation.
Test of Mathematical Abilities	Brown, V. L., Cronin, M. E., & McEntire, E. (1994). Test of Mathematical Abilities, Second Edition. Austin, TX: PRO-ED.
Language	
Spellmaster	Greenbaum, C. R. (1987). Spellmaster. Austin, TX: Pro-Ed.
Test of Written Language-3	Hammill, D. D., Larsen, S.C. (1996). Test of Written Language, Third Edition. Austin, TX: Pro-Ed.
Woodcock Language Proficiency Battery - Revised	Woodcock, R.W. (1991). Woodcock Language Proficiency Battery-Revised. English and Spanish Forms. Chicago: The Riverside Publishing Company.
Written Language Assessment Test	Grill, J. J., & Kerwin, M.M. (1989). Written Language Assessment Test. Novato, CA: Academic Therapy Publications.

humanities, and social studies, varies significantly. Sax (1974) provides a description of the major differentiating characteristics of 10 of the most commonly used achievement test batteries.

In contrast to the "survey" type tests or screening batteries described above are the more content-focused diagnostic achievement tests. Although any of the survey instruments is available to identify areas of academic strength or weakness (Radencich, 1985), they are not in themselves sufficient for diagnostic or remediation-planning purposes. Their use in screening large groups helps to identify those individuals in need of more specific individualized diagnostic evaluation. Through the use of a diagnostic battery, an area of identified deficit is examined in a more extensive fashion to determine what factors contribute to the academic dysfunction. Typically, these tests include a broad enough sampling of material so that areas of need are specified in order to develop remedial instructional objectives. For example, the Woodcock Reading Mastery Tests-Revised (Woodcock, 1987) provides five subtests which examine component processes associated with overall reading ability. These include Letter Recognition, Word Attack, Word Recognition, Word Comprehension, and Passage Comprehension. More in-depth examination at this level permits hypothesis generation regarding the nature of the specific academic deficit to be further tested. Similar tests are available to assess other aspects of academic performance: mathematics, spelling, writing, language skills, etc. Refined assessment at this level is necessary for differential diagnosis and remedial intervention. Screening batteries simply do not permit sufficient evaluation of an area for this kind of decision making to take place.

Although most achievement tests have the potential to be used as screening instruments to identify individuals in need of remedial instruction, fewer instruments actually appear to have been used for diagnostic purposes. In a national survey conducted in the early 1980s, Goh, Teslow, and Fuller (1981) reported that the Wide Range Achievement Test and the Peabody Individual Achievement served as the general achievement batteries most commonly utilized by school psychologists. At that point in time, in the area of specific achievement tests, the Key Math Diagnostic Achievement Test, the Illinois Test of Psycholinguistic Abilities (ITPA), and the Woodcock Reading Mastery Tests ranked as the instruments used most frequently for the assessment of specific academic content areas. However, in the late 1990s,

one rarely, if ever, encounters reference to the ITPA either in reported research studies or in diagnostic test reports used as part of an Individualized Education Plan.

Criterion-Referenced versus Norm-Referenced Achievement Tests

One other highly significant dichotomy must be addressed when discussing the classification of achievement tests and certain of their psychometric properties, namely, the distinction between criterion-referenced tests (CRTs) and norm-referenced tests (NRTs). While it is not possible to differentiate one from the other in terms of visual inspection (a criterion-referenced test can also be used as a norm-referenced test: for example, Basic Achievement Skills Individual Screener), there are intrinsic differences between the two approaches to achievement testing. Traub and Rowley (1980) described the decade of the 1970s as a time when "the notion of criterion-referenced measurement captured and held the attention of the measurement profession unlike any other idea" (p. 517). Mehrens and Lehmann (1975) asserted that the issues of accountability, performance contracting, formative evaluation, computer-assisted instruction, individually prescribed instruction, and mastery learning created a need for a new kind of test, the criterion-referenced test.

The concept of criterion-referenced achievement measurement was first detailed in the 1963 paper by Robert Glaser entitled "Instructional Technology and the Measurement of Learning Outcomes: Some Questions." In that landmark publication Glaser wrote:

> Underlying the concept of achievement is the notion of a continuum of knowledge acquisition ranging from no proficiency at all to perfect performance. An individual's achievement level falls at some point on this continuum as indicated by behaviors he displays during testing. The degree to which his achievement resembles desired performance at any specified level is assessed by criterion-referenced measures of achievement or proficiency... Criterion levels can be established at any point in instruction....
>
> Criterion-referenced measures indicate the content of the behavioral repertory.... Measures which assess student achievement in terms of a criterion standard...provide information as to the degree of competence attained by a particular student which is independent of reference to the performance of others. (p.519)

Glaser further stated that achievement measures are appropriately used to provide information regarding a student's capability in relation to the capabilities of his or her fellow students as well. Where an individual's relative standing along the continuum of attainment is the primary concern, the appropriate achievement measure is one that is norm referenced. Whereas both CRTs and NRTs are used to make decisions about individuals, NRTs are usually employed where a degree of selectivity is required by a situation, as opposed to situations in which concern is only with whether an individual possesses a particular competence and there are no constraints regarding how many individuals possess that skill. Thus, at the core of the difference between the two kinds of tests is the issue of variability. "Since the meaningfulness of a norm-referenced score is basically dependent on the relative position of the score in comparison with other scores, the more variability in the scores the better" (Popham, 1971). This obviously is not a requirement of the criterion-referenced measure.

Because of basic differences in the theories underlying test construction, there have been several hundred publications on CRTs dealing with such issues as test reliability, determination of test length (Millman, 1973), score variability (Hambleton & Cignor, 1978; Hambleton, 1980), and test validity (Linn, 1982). The psychometric properties of CRTs have undergone close scrutiny, and one of the most critical dimensions reviewed has been the issue of validity. In the words of Linn (1980):

> Possibly the greatest short-coming of criterion-referenced measurement is the relative lack of attention that is given to questions of validity of the measures. The clear definitions of content domains and well-specified procedures for item generation of some of the better criterion-referenced measures place the content validity of the tests on much firmer ground than has been typical of other types of achievement tests. Content validity provides an excellent foundation for a criterion-referenced test; but...more is needed to support the validity of inferences and uses of criterion-referenced tests. (p. 559)

In their review of 12 commercially prepared criterion-referenced tests, Hambleton and Cignor (1978) did not find a single one that had a test manual that included satisfactory evidence of validity (Hambleton, 1980). Validity has too often been assumed by both developers and users of criterion-referenced tests. This is no more acceptable for a criterion-referenced test than it is for any other test. It is time that questions of validity of the uses and

interpretations of criterion-referenced tests be given the attention they deserve.

Despite these criticisms from the point of view of traditional test-construction theory, criterion-referenced measurement has been found to have major utility with respect to the development of computer-assisted, computer-managed, and self-paced instructional systems. In all of these instructional systems, testing is closely allied with the instructional process, being introduced before, during, and after the completion of particular learning units as a monitoring, diagnostic, and prescriptive mechanism (Anastasi, 1982). Moreover, it has had practical applications with respect to concerns with minimum competency testing (Hunter & Burke, 1987; Lazarus, 1981) and mastery testing (Harnisch, 1985; Kingsbury & Weiss, 1979).

Curriculum-Based Measurement

In addition to criterion-referenced and norm-referenced tests of achievement, one additional "hybrid"—which appears to be surfacing, particularly in the area of special education—*curriculum-based measurement* (CBM), merits a brief note in this review. From the Institute for Research on Learning Disabilities at the University of Minnesota, Deno (1985) and his colleagues have proposed a method of measurement which lies somewhere between the use of commercialized tests and informal teacher observations. Their initial research with the procedure in the areas of reading, spelling, and written expression, and concerns with reliability, validity, and limitations are reviewed by Deno. Among the limitations are its utility only with the domain of reading at present, its lack of stability estimates as indicative of reliability, and its lack of generality that enables aggregation across curricula.

However, one aspect of CBM that appears to mark a distinct embarkation from traditional achievement testing is the concept of frequent measurement. In addition to the work of Mirkin, Deno, Tindal, and Kuehnle (1982) on the measurement of spelling achievement with learning disabled students, LeMahieu (1984) reported on the extensive use of a program of frequent assessment known as the Monitoring Achievement in Pittsburgh (MAP) which began in 1980 and involved 81 schools with a total enrollment of 40,000 students. Students were tested every six weeks with curriculum-based measures developed by committees of teachers. Serious risks in this kind of achievement testing involve the potential for teachers to narrow the curriculum and to teach to the assessment instrument as well as for students themselves to develop and refine test-wise behaviors as opposed to attaining specific academic skills.

USE OF ACHIEVEMENT TESTS

Achievement Tests in Education

Within the context of educational programs there is a continual process of evaluation that also includes teacher-made tests and letter-grade performance standards. The continuous monitoring of student performance within a particular academic content area provides means not only to assess student progress but also to link instructional strategies and learning objectives with identified student learning needs or skill deficits. Out of a concern for the performance of public schools, statewide minimum competency testing programs proliferated in the 1990s. "Policymakers reasoned that if schools and students were held accountable for student achievement, with real consequences for those that didn't measure up, teachers and students would be motivated to improve performance" (Improving America's Schools, 1996, p.1). Traditional achievement tests were judged to be "low-end" tests (p.1), and the advent of standards-based reform was seen as impetus to revamp methods of student assessment, a revamping which is ongoing at the time of this writing.

In a similar vein, a study by Herman, Abedi, and Golan (1994) assessed the effects of standardized testing on schools. They surveyed 341 elementary teachers in 48 schools, although the location of the schools was not identified. In their study, classes in which disadvantaged students were the majority were more affected by mandated testing than those serving their more advantaged peers. Results suggested that teachers serving disadvantaged students were under greater pressure to improve test scores and more driven to focus on test content and to emphasize test preparation in their instructional programs.

Despite such criticisms with respect to the misuse or inappropriate use of these tests, the periodic administration of achievement tests has traditionally been viewed as an educationally

Table 7.2. Achievement Tests: Purpose and Outcome

PURPOSE OF TESTING	OUTCOME CRITERION
Screening:	Identification of students potentially eligible for remedial programming.
Classification/Placement:	Specific academic deficiencies have been ascertained. Question now arises regarding whether student meets eligibility criteria.
Prescriptive Intervention:	A specific developmental arithmetic disorder is manifest in a child identified with visuo-perceptual processing problems. What curriculum adjustments appear warranted?
Program Evaluation:	Administrators seek to evaluate benefits of an accelerated reading program for gifted students.

sound procedure by professionals in the field. From a positive perspective, Anastasi (1988) provided a summary of their usefulness in educational settings. First, their inherent objectivity and uniformity provide an important tool in assessing the significance of grades. While individual classroom-performance measures can be susceptible to fluctuation because of a number of variables, their correlation with achievement-test scores provides a useful comparative validity criterion for grades. They are especially useful in the identification of students whose limited progress in a content area will require remedial intervention. Within this context, individualization of specific needs can be identified so that individual and group curricula can be modified. In this regard, the use of achievement tests prior to the initiation of training can become particularly efficacious. When these measures are utilized at the end of an instructional period they have the potential to serve as a means for assessing the quality of instructional programming and aiding in programmatic evaluation.

In general, then, achievement tests are used to make decisions, decisions which may involve instructional, guidance, or administrative issues. For example, what is the efficacy of a particular method of instruction? What are the specific outcomes of learning? Is there a need for remediation? Are grading practices accurate? Is the curriculum responsive to the acquisition of basic and specific academic skills? Is counseling appropriate for any given student? Is appropriate placement a concern? Thus, the breadth of the assessment will be predicated upon the rationale for the use of particular achievement measures. Table 7.2 illustrates the types of questions or problems that may be addressed and the expected benefit(s) to be derived from the testing process.

Achievement versus Aptitude

One further point, which any review of achievement tests must certainly address with respect to their classification and use, is the notion of aptitude versus achievement. This contrast dates back to the preoccupation of educational psychologists in the 1920s and 1930s with the role of heredity versus environment in the learning arena. This early simplistic notion that innate capacity or potential could be measured by aptitude tests independent of an individual's learning history or "reactional biography" (Anastasi, 1984, p. 363) has been disavowed. Replacing the traditional concepts of aptitude and achievement in psychometrics is the concept of "developed abilities," the level of development attained by an individual in one or more abilities (Anastasi, 1982, p. 395). In line with this conceptualization of the measurement of abilities, Anastasi provides a continuum of testing in terms of the "specificity of experimental background" that particular tests presuppose. The continuum ranges from course-oriented achievement tests to broadly oriented achievement tests to verbal-type intelligence to "culture-fair" tests. This continuum more accurately reflects the overlapping of aptitude and achievement tests. This analysis has been demonstrated empirically over and over in terms of the high correlations between achievement and intelligence tests. "In some instances, in fact, the correlation between achievement and intelligence tests is as high as the reliability coefficients of each test" (Anastasi, 1982, p. 395).

Finally, Anastasi notes that the continued labeling of some tests as aptitude or achievement measures has led to misuses of test results—in particular, the identification of certain children as underachievers when their respective achievement-test scores are lower than their scholastic aptitude- or intelligence-test scores. In the words of Anastasi (1982):

> Actually, such intraindividual differences in test scores reflect the universal fact that no two tests...correlate perfectly with each other....Among the reasons for the prediction errors in individual cases are the unreliability of the measuring instruments, differences in content coverage, the varied effects of attitudinal and motivational factors on the two measures, and the impact of such intervening experiences as remedial instruction or a long illness. (p. 396)

Scoring Systems Associated with Tests of Academic Achievement

Before further discussion of the application of traditional achievement-test data, it is necessary to consider how the results of these tests are conveyed. Raw scores derived from achievement tests are typically converted to age- or grade-equivalent scores, standard scores, or percentile scores. Hoover (1984) makes a useful distinction between two scoring dichotomies. *Developmental scores* compare individual performance to that of a series of reference groups that differ systematically and developmentally in average achievement, with developmental scores being expressed as age- or grade-equivalent scores. *Status scores* compare test performance with a single normative reference group and are expressed as standard scores and percentiles. It is important to distinguish between the two types of measurement as each has unique strengths and limitations.

Developmental Scores

Age-Equivalent Scores. Educational Age (EA) represents a scoring criterion which has come under significant criticism and is used very infrequently in reporting educational test data. The scaling of items on some achievement tests is presented in a developmental sequence such that a particular score represents mean level of performance for a specific-age reference group. An individual who attains a specific score on the test is reported to function at a particular age level. This system of score reporting is useful for descriptive purposes, especially for "measuring growth." As in grade-equivalent scores, which will be discussed next, serious flaws are encountered when one attempts to utilize such scores for comparative purposes.

Grade-Equivalent Scores. A grade-equivalent score (GE) reflects the presumed level of performance of an average student at a particular grade level. For example, if the mean score of a group of sixth graders on an achievement test is reported as M = 6.2, children who attain the same score are imputed to function at a level of performance commensurate with sixth graders in general. Although it is quite important to have available a continuous scale describing developmental level as a means to demonstrate progress in attainment and growth, the GE represents one of the most frequently misinterpreted sources of educational data. First, it should be noted that GE scores are reported in a format that reflects both grade level and month. The typical school year is approximately 10 months. Hence, scores of 6.2 and 6.9 contrast levels of performance commensurate with the beginning and end of the school year. There are, however, limitations on direct interpretation of GE scores. The scaling of achievement-test data is rarely a continuous process. Scores for many grade equivalents are frequently extrapolated or interpolated and consequently do not reflect actual derived scores. They are, in fact, estimations based on a hypothetical grade-equivalent curve. The use of such a scale also presumes that the teaching of such skills is a continuous process reflected across grades. This is not, however, reflected in the reality of the educational experience. Gains made by students are more realistically seen as a combination of spurts and plateaus, and not as a continuous process as is mathematically interpolated in scale construction.

The most significant limitation in the use of GE scores appears to arise because they are ordinal measures. The difference between a one-year gain in proficiency at a lower grade level in comparison to that same gain at a higher grade level may be significant. Further, because most of the basic core academic competencies are taught within the first through eighth grades, one cannot presume that grade-equivalent scores associated with the terminal stages of the educational career are equivalent. Finally, it must be noted that relatively small differences in performance can result in exaggerated differences in grade-level equivalency owing to the nature of scale construction.

The most frequently cited problem with using GE scores is the potential for misinterpretation of significant differences in level of performance. For example, a fourth grader obtains a score of

6.7 in reading. One cannot directly compare this youngster to other sixth graders. It is an erroneous assumption to state that this child's reading ability is commensurate with that of a sixth grader. His reference group remains fourth graders. He clearly demonstrates well-above-average performance in comparison to this reference group. One cannot, however, compare him to sixth graders, who by the nature of their development and experience with reading, are different from our fourth grader. Because of the inherent potential for parents to set inappropriate standards of performance for their children based on such scores, the use of scores has been abandoned in many quarters.

Status Scores

A wide variety of standard score methodologies are available for reporting test results. These represent scores scaled along a continuum which permits one to ascertain where a particular score may fall in comparison to other scores in a distribution. There are two distinct advantages to the utilization of this scoring system. Standard scores permit the opportunity to compare individual performance to a normative standard, and they make possible the comparison of individual performance across two or more different tests. The latter represents an important criterion for the application of achievement tests within the context of a larger test battery.

Percentiles

Percentile rank represents a point in a distribution at or below which the scores of a given percentage of subjects fall. If a student scored at the 95th percentile, this would mean his or her score was better than 95 percent of the other students who took the same test. When clearly conveyed in the context of a psychological report, this scoring methodology represents one of the most readily understandable forms of test description. The potential for inappropriate comparisons of level of performance, as reflected in the GE score example, is significantly reduced.

Standard Scores

Standard scores represent raw scores that have been scaled relevant to a constant mean and standard deviation. As a function of the magnitude of the standard deviation, one can, through linear transformation, readily ascertain how far from the mean performance lies. Most tests standardize scores within defined age groups. Therefore, regardless of the age of the subjects under evaluation, a specific standard score will have the same meaning. For example if two students, ages 8 and 10 years, obtain the same standard score on a reading test, relative to the normal curve, one can readily distinguish that in comparison to their age mates, they are functioning at equal distance from the mean. Standard score conversions also include *z* scores, *t* scores, and occasionally stanine scores which can be interpreted in like manner. In general, standard scores are considered the more accurate and precise means of reporting test results. Finally, it is not uncommon for test developers to provide multiple methods for performance description. For example, the Wide Range Achievement Test-3 provides grade equivalent scores, age-based standard scores, percentiles, normal-curve equivalents, and absolute scores.

Achievement Test Scores and the Diagnosis of Learning Disabilities

The relevance of understanding the scoring systems utilized in the interpretation of achievement-test results can be dramatically illustrated when one considers the educational diagnosis of a specific learning disability. Learning disabilities have become the dominant handicap of school-age children in the country, with some 42 percent of all students ages 3 to 21 years in special education programs diagnosed as learning disabled (Databank, 1985).

A basic assumption underlying learning disabilities is the failure of the student to acquire primary academic skills at levels expected for age, grade placement, and level of intellectual functioning. The identification of individuals with learning disabilities has traditionally been based on the notion of a "significant discrepancy" between ability level and demonstrated academic skill attainment. Regardless of which of the many formulas is used to diagnose a learning disability, all require data from standardized achievement tests. Thus, the use of achievement testing has become an integral component in the differential diagnosis of learning disabilities. In this regard, the concept of "significant discrepancy" has been an important one, for it forms the basis for distinguishing specific learn-

ing-disability diagnoses from conditions such as underachievement or mental retardation.

Under Public Law 94-142, the Education for All Handicapped Children Act of 1975, it was specified that a team could render a determination of specific learning disability if a child did not achieve at his or her ability level when provided with appropriate educational instruction and if a severe discrepancy existed between intellectual ability and achievement in one or more of seven areas of achievement, including oral expression, listening comprehension, written expression, basic reading skills, reading comprehension, mathematics calculation, or reasoning. Specifically excluded along with mental retardation were other factors which could impinge on limited academic proficiency, such as peripheral sensory or motor handicaps, emotional disturbance, or socioeconomic or cultural disadvantage. The actual specification of the means of ascertaining discrepant performance is left vague in this definition. Algozzine, Ysseldyke, and Shinn (1982) emphasize that the field of learning disabilities has always suffered a definitional dilemma. Federal guidelines have not appreciably corrected this situation. No clear consensus across school districts exists nationally for arriving at workable definitions of learning disability diagnoses (Shaw, Cullen, McGuire, & Brinkerhoff, 1995).

In spite of the lack of consensus regarding definition, the notion of severe discrepancy has been defined most frequently by the use of an ability-achievement discrepancy. Inherent in this conceptualization of learning disability is the potential for at least average intellectual functioning with academic performance well below expectations. A number of strategies have been applied in an attempt to operationalize criteria representative of a severe discrepancy.

Deviation from Grade Level

A commonly encountered criterion used to define a potential learning disability might be "grade level performance in academic achievement two grade levels below expectation for age." This criterion has been criticized as inadequate for a number of reasons. First, as previously discussed, grade-level equivalents represent the weakest psychometric criterion upon which to base comparisons of academic performance. Second, utilization of such a constant criterion fails to take into consideration the significance of discrepant performance at various points in the continuum of educational programming. For example, performance two grade levels below expectation in a third grader can be far more significant than the same magnitude of score deficit in an eighth grader. Further, in the assessment of adult populations, the efficacy of grade-equivalent scores loses predictive validity. It is extremely difficult to ascertain whether eighth-grade academic skills in a 40 year old are indicative of any significant disparity in level of performance.

Finally, problems have been identified with potential identification of students who may be learning-disabled. Use of grade-level discrepancy criteria tends to overidentify children whose intellectual functioning is below average and to underidentify those students who may be above average. A student with an IQ of 82 might in fact be functioning at a grade level which is not discrepant for his or her overall level of intellectual functioning. On the other hand, a fourth grader who is reading at or just below grade level, but who has an IQ in the superior range and who should clearly be reading at well above grade level expectations, would be excluded.

Standard Score Discrepancy Models

The process of comparing standard scores derived from academic and intelligence tests holds apparent benefits over grade-discrepancy scores on purely psychometric grounds. Typically, a criterion level is arbitrarily selected, a 1 or 2 standard-deviation-point discrepancy between general ability and achievement test score. This methodology can, however, also impose bias into the discrimination process. Many such models do not take into consideration the regression of IQ on achievement. One cannot assume direct correspondence between IQ and standard score equivalents. It can be demonstrated that academic achievement-test scores fall somewhat short of IQ for individuals, manifesting above-average performance, and in lower-functioning individuals, academic achievement-test scores are actually higher. The use of a simple discrepancy-score formula implicitly assumes a perfect correlation between general ability and achievement tests which in fact does not exist. It would also require that each test be based on the same standard-score distribution.

Regression Equations

The most sophisticated methodologies available for determining significant score discrepancies are based upon complex computations or tables designed from formulas based on regression equations. A number of strategies have been developed, each with unique distinguishing properties. A number of reviews are available (Forness, Sinclair, & Guthrie, 1983; Reynolds, 1984; Wilson & Reynolds, 1984) that describe the characteristics of these methodologies. There remains, however, no one mathematical model that is commonly accepted or in fact utilized.

Reynolds (1984) reported on the findings of the Work Group on Measurement Issues in the Assessment of Learning Disabilities, a study section formed in 1983. This group was delegated the responsibility of addressing questions directed toward identification of "best practice" solutions to the learning disabilities definitional dilemma. In their findings, models of discrepancy analysis based upon grade-equivalent scores were rejected outright. Factors related to their imprecision and their ready misinterpretation were noted. Most critical, however, was the inherent lack of the mathematical properties necessary for conducting comparative analyses that are associated with this scoring system. The group concluded that age-based standard-score discrepancy models represent potentially the best methodology available. However, while developmental standard scores are to be preferred over grade-level or status-standard scores, their value has been challenged also because they require greater growth for below-average children than for average or above-average children (Clarizio & Phillips, 1986).

One cannot, however, focus exclusively on the concept of discrepancy as the sole basis for the diagnosis of a learning disability. To quote Reynolds (1984), "The establishment of a severe discrepancy is a necessary but insufficient condition for the diagnosis of a learning disability" (p. 468). A host of factors other than a specific learning disability (LD) can contribute to significant academic underachievement. Among these are limited sociocultural opportunity, dysmotivation, sensory-perceptual dysfunction, or functional psychiatric impairment. It is Reynolds' bias, however, that only when a severe discrepancy can be demonstrated is a child considered eligible for a diagnosis of LD. This bias has come under severe criticism of late because identification, and therefore remediation, must wait until the student fails (Shaw, Cullen, McGuire, & Brinkerhoff, 1995). In contrast, research studies consistently support the efficacy of early intervention. Studies supporting the identification of reading problems as early as the pre-school years, with programs in kindergarten that include a focus on phonological and orthographic awareness, are compelling in this regard (Foorman, Francis, Beeler, Winikates, & Fletcher, 1997; Wasik & Slavin, 1993; Lundberg, Frost, & Petersen, 1988).

Some Thoughts on the Validity Issue

There are, among educators and researchers, those who question the focus on the *reliability* of the IQ and achievement discrepancy versus its *validity* (Shepard, 1983). In a study by Shepard and Smith (1983), which evaluated the identification practices of psychologists and teachers within the state of Colorado involving 1,000 student files and 2,000 specialists, 50 percent of those professionals surveyed were unaware that an IQ of 90 falls at the 25th percentile. For children with IQs of 90, the expectation was that achievement would be at grade level (the 50th percentile) because the IQ was "in the normal range" (Shepard, 1983). The authors continued that these specialists were unaware also that after the first or second grade, it is not uncommon for large numbers of children to have grade-equivalent scores below their grade placement.

Other technical problems identified in this study further complicate the identification of LD. Most of the tests used in the diagnosis of LD were technically inadequate with the exception of the WISC-R and one or two achievement batteries. Many clinicians were unaware of the differences between technically adequate and inadequate tests. Specialists often selected technically inadequate measures even when more valid instruments were available; their choices tended to follow traditional preferences associated with each professional group. Many clinicians continued to apply inaccurate conventional wisdom regarding the symptoms of the disorder (relying on interpretations of subtest scatter, underestimating normal patterns of difference, etc.) (Shephard & Smith, 1983).

Reynolds (1984) and the Task Force advanced a number of recommendations which attempted to bridge this validity-reliability gap with respect to the diagnosis of LD:

1. Instruments applied should meet criteria defined in PL 94-142.
2. Well-standardized national norms should form the basis for statistical comparison of individual levels of performance.
3. Normative comparisons should be based upon co-normed samples. The ideal scenario is one in which the two tests compared are normed on the same sample. Where this is not possible, the two normative groups should be clearly comparable.
4. Only individually administered tests of achievement and intellectual ability should be utilized.
5. Age-based standard scores based upon a common scale represent the most statistically robust means for score comparison.
6. Measures employed should conform to acceptable criteria for validity and reliability.
7. Special technical considerations should be addressed when using performanc-based measures of achievement (e.g., writing skill).
8. Bias studies should have been conducted and reported.

In summary, while the psychometrics involved in scoring and interpreting the results of achievement tests can be fraught with complexity and controversy, as illustrated in the case of the diagnosis of learning disabilities, the consequences of the resolution of the issues involved are even further reaching. Consider the effects of labeling, the contraction of teacher competence to deal with a variety of learning styles in the classroom, the allocation of resources available to those students with the most severe disability, and the costs of providing for special education resources themselves (Shepard, 1983). All of these can be viewed as negatives. It is not difficult nor unrealistic to extrapolate these same issues to include diverse groups of students in educational programs today. Thus, we are left with ethical responsibilities to insure the appropriate utilization of achievement tests based on the most current thinking and research available, which is macrocosmic rather than microcosmic in nature.

Messick (1980) has argued this point in his "Test Validity and the Ethics of Assessment." He had written earlier, with specific reference to the measurement of personality, that tests should be evaluated not only in terms of their measurement properties but also in terms of their potential social consequences (Messick, 1965). Messick emphasized the importance of construct validity, arguing "that even for purposes of applied decision making reliance upon criterion validity or content coverage is not enough" (Messick, 1975, p. 956), and that "the meaning of the measure must also be comprehended in order to appraise potential social consequences sensibly" (Messick, 1980, p. 1013). He defined test validity as an overall evaluative judgment of the adequacy and appropriateness of inferences drawn from test scores, opining that values questions arise with any approach to psychological testing, whether it be norm-referenced or criterion- referenced, a construct-based ability test, or a content-sample achievement test. This evaluative judgment of test validity is based on (a) convergent and discriminate research evidence as to the test scores interpretability in terms of the particular construct under review; (b) an appraisal of the value implications of that interpretation; (c) justification of the relevance of the construct and its utility of the particular application proposed; and (d) dealing with the potential social consequences of the proposed use as well as the actual consequences upon implementation of the testing procedure.

> Intervening in the model between test use and the evaluation of consequences is a decision matrix to emphasize the point that tests are rarely used in isolation but rather in combination with other information in broader decision systems. The decision process is profoundly influenced by social values and deserves, in its own right, massive research attention beyond the good beginning provided by utility models. (Messick, 1980, p. 1025)

Messick concluded his remarks by paraphrasing Guion (1976): "The formulation of hypotheses is or should be applied science, the validation of hypotheses is applied methodology, but the act of making...(a) decision is...still an art" (p. 1025).

The Use of Achievement Tests in Clinical Practice

Achievement testing conducted with clinical populations is generally regarded as an extension of intelligence and aptitude testing. It provides one further means to ascertain "general ability level." Results are typically utilized for drawing inferences regarding the capacity of the individual under evaluation to apply knowledge or native intelligence in practical problem-solving situa-

tions. One equates intelligence and exposure to educational opportunity with the ability to conform with the demands of achievement testing at commensurate levels of success. Typically, one is not engaging in achievement testing with this population in anticipation of identification of potential performance discrepancies, but to gauge overall adaptive competency. The identification of any significant discrepancies would of course result in further clinical investigation. Cognitive as well as noncognitive variables would then be explored.

Achievement Test Results Applied in Neuropsychological Evaluation

Achievement tests play a definitive role in the administration of standard neuropsychological test batteries. For example, a number of extended versions of the Halstead-Reitan Neuropsychological Test Battery include an administration of the Wide Range Achievement Test or another age-appropriate screening battery within the test protocol. Data derived from such tests offer clinical utility beyond discrepancy analysis. They can be used as a method to infer an estimated level of premorbid intellectual functioning (Lezak, 1983). As basic academic skill competencies are generally not susceptible to significant deterioration in mild-to-moderate generalized cerebral dysfunction, standard scores derived from general achievement-test measures offer one means to interpolate a coarse estimation of premorbid functioning when other means of documentation are not available.

Achievement-test results can be incorporated into the pattern analysis of other neuropsychological test variables to aid in the specification of the effects of focal-lesion processes. For example, problems exclusively with the spatial components of arithmetic processes in an individual manifesting no evidence of linguistic defects would help suggest a post-Rolandic lesion of the right cerebral hemisphere, when other markers of right hemisphere dysfunction are present. It is not uncommon to consider achievement-test performance within the context of a formal aphasia examination as a means to extend the assessment to the integrity of lexical-skill functions and writing ability.

Beyond their application in the documentation of the effects associated with focal-lesion processes, such test results hold even greater potential utility in aiding in the development of hypotheses regarding functional limitations associated with cerebral dysfunction. As primary academic-skill competencies are intimately related to aspects of autonomous functioning in a number of instrumental activities associated with daily living, the degree of preservation of such primary skills as reading and arithmetic abilities can be important prognostic indicators associated with long-term recovery and adaptation.

Achievement Test Results Applied to Rehabilitation Assessment Methodologies

In the areas of both psychiatric and vocational rehabilitation, the specification of the degree to which core academic competencies are developed holds a number of prognostic implications. With low-level functioning individuals, the specifications of primary literacy skills is an important determinant of the level of complexity of programming in which they might participate. The degree to which a learning curriculum might emphasize effective reading comprehension might be potentially exclusionary, for example.

An important component of the rehabilitation assessment is determination of the degree to which any remedial intervention might be required prior to implementing programming. Inadequate educational opportunity or underachievement related to psychosocial factors must be distinguished from developmental academic disorders and conditions which cause a loss of previously attained ability. Intervention strategies to remediate or supplant deficient academic skills are determined by the thorough analysis of their cause. Prognostically, it is important to identify those individuals functioning at their plateau versus those who have the potential to develop these skills further.

In summary, with the use of achievement testing in clinical settings the focus is typically divested towards two lines of inquiry: (a) obtaining knowledge of the degree to which basic academic skill competencies are developed in a particular individual, and (b) examining individual performance within a particular area of academic performance. The basic referral question in large measure determines what armamentarium of techniques will be brought to bear in the assessment. It will also influence how test scores will be compared and interpreted.

COGNITION, METACOGNITION, AND ACHIEVEMENT TESTING

The application of cognitive-theory research to educational psychology can be traced back as early as 1960 with the publication of David Ausubel's paper "The Use of Advance Organizers in the Learning and Retention of Meaningful Verbal Material," the later work of Rothkopf (1965) on mathemagenic behaviors, Ausubel's (1968) text, *Educational Psychology: A Cognitive View,* Anderson's (1972) work on how to construct achievement tests to assess comprehension, and the work of Marton and Saljo (1976a, 1976b) who argued that a description of what is learned is more important than a summary of how much is learned (Clarke, 1982). Glaser (1981) reviewed current research in cognitive and developmental psychology addressing its potential influence on the development of new psychometric methodology. He cited Bartholomae's (1980) work on error analysis with college students in remedial-writing programs and Siegler's (1976) work on rule assessment in the acquisition of scientific concepts as illustrative of the "necessary interrelationships between the analytical assessment of performance and effective instruction" (Glaser, 1981, p. 929). Interest in the assessment of mastery or competence can be traced also to developments in cognitive psychology, artificial intelligence, and language understanding. Herein the works of Chase and Simon (1973) on the chess master and the work of Larkin, McDermott, Simon, and Simon (1980) on problem solving in the area of elementary physics were cited by Glaser.

Finally, research in the realm of metacognition—the knowledge, regulation, and management of one's own cognitive processes and products—(Flavell, 1976) has led to a concern with the measurement of these self-regulatory skills in terms of predicting successful problem solving which then leads to learning. Metacognitive abilities develop with maturity, and current research in learning instruction has demonstrated that these skills may be less well developed in those individuals who have learning disabilities.

Thus, it becomes quite clear that an understanding of the learning process and its assessment can yield more fruitful data than those traditionally obtained by achievement tests. This is particularly important in light of the social-educational demands outlined by Glaser (1981) which will shape and mold the future of educational assessment:

- the shift from a selective educational system to one designed to help individuals succeed in educational programs (zero-reject system);
- the requirement for improved levels of literacy and problem-solving ability in a variety of knowledge and skill domains (minimum competency and mastery certification);
- the need to understand individual differences in the process of measurement so that abilities can be improved to facilitate further learning (cognitive, sociocultural, gender specific.

The application of cognitive and metacognitive principles with respect to the measurement of learning have been detailed in the areas of reading (Curtis, 1980; Curtis & Glaser, 1983), spelling (Henderson & Beers, 1980; Nolen & McCartin, 1984), and foreign language (Fischer, 1981; Stevenson, 1983; Terry, 1986). Curtis and Glaser (1983) describe the current level of understanding regarding the theoretical framework utilized to study the process of learning to read, a process which involves a complex of interrelated skills (word decoding, accessing semantic word-information, sentence processing, and discourse analysis), proficiency in one affecting success in the others. The results of traditional reading-achievement tests have made it impractical to diagnose reading problems in terms of remediation or instructional strategies thus far. However, current theory on efficiency in word identification, the qualitative features of semantic knowledge, and research on schemata can be utilized as a form of construct validity and thus allow measurement of achievement that reflects both the development of competence and the process of instruction. "With developing knowledge of reading it should be possible to establish standards of performance...[and]...combined enterprise representing test design based on knowledge of human learning and performance, psychometric requirements, and studies of test use should improve our ability to link testing and instruction" (Curtis & Glaser, 1983, p.144).

Diagnostic Application of Achievement Test Results

As an illustration of the application of cognitive and metacognitive strategies in the process of achievement testing, the remainder of this discussion focuses on an expanded level of analysis that can be undertaken in the clinical setting for purposes of both diagnosis and remediation interventions.

Reading

Converging lines of research (Fletcher et al., 1994; Shaywitz, Fletcher, & Shaywitz, 1996) have emphasized the primacy of core phonological-processing deficits in disabled readers. Unlike language competencies which unfold naturally in a fairly predictable fashion, reading represents an acquired skill. As such, not only constitutional but environmental determinants may contribute to failures in reading. Without an explicit model of normal reading development, patterns of impairment cannot be described.

Reading development has been traditionally dichotomized across two component skills which must be mastered. These have been described within a dual-route model of reading as involving (a) a phonological, and (b) a direct lexical route in which whole-word or orthographic-recognition skills facilitate active word-recognition (Morton, 1969). In the earliest stages of the development of pre-reading competencies, the logographic stage, mastery of the visual-orthographic properties of letters, memorization of the visual gestalts of a limited repertory of words, and utilization of visual associative skills to foster word recognition from pictures that accompany text, are accomplished. The early reader is simultaneously developing sound-symbol associative skills. Such abilities are predicated upon a child's ability to first decompose speech into component structures (phonological awareness). Via these processes, the early reader is learning to associate visual symbols (graphemes) with their corresponding sound equivalents (phonemes). With increasing experience, direct orthographic-associative competency is established and familiar words are decoded on "sight". Confrontation with low frequency, novel words is presumed to require some combination of orthographic- and phonological-coding abilities (Coltheart, 1978). Skilled readers are presumed to have developed automatized orthographic skills in reading by the fourth grade. As such, phonological skills are relegated primarily to the processing of less familiar words. A competent reader is presumed to have developed equivalent proficiency in both aspects of word analysis. In addition, with advancing age, the capacity to incorporate morphological cues as well as lexical-semantic, and other contextual cues, further contributes to the act of reading. As such, words rich in meaning tend to be decoded with greater faculty than more ambiguous function words.

The complexity of reading-skill acquisition expressed over time requires recognition of the reciprocal contributions of higher-level processing systems, beyond the dual-route model. Cognitive models of reading development (see Chase & Tallal, 1991, for review) take into consideration increased capacity to bring on-line higher cognitive faculties (the simultaneous development of not only "bottom-up" but "top-down" processing) to the development of reading fluency and comprehension (McClelland & Rumelhart, 1981). Progression in reading proficiency requires the act of word recognition becoming automatized (Liberman, Liberman, Mattingly, & Shankweller, 1980). Thus, on-line cognitive activities are directed less at the act of decoding and permit the use of metalinguistic awareness, selective attention and working memory to support semantic encoding and comprehension monitoring. Although phonological-coding skills (word- analysis abilities) represent the most widely studied aspect of reading development, more recent research has focused upon the contribution of other types of linguistic-rule knowledge (semantic, morphological, and syntactic conventions) to the development of higher-level reading skills, ie., comprehension skills.

Associated with research on phonetic processing in deficient readers has been the identification of specific linguistic deficiencies associated with reading disabilities. In a review by Mann (1994), deficiencies associated with naming and verbal productivity, expansion of semantic knowledge, auditory sequential memory, sentence recall, and grammatical and syntactical analysis skills (particularly comprehension demands requiring the processing of more complex grammatic structures) have been identified. Catts (1989) futher identified higher rates of oral-speech deficits, including early history of articulation inefficiencies among the reading disabled. A commonality linked to these

deficits involves the processing of sound patterns of language. These deficits have been broadly conceived as evidence that specific maturational lags in the systems supporting language development represent a secondary source of the cognitive deficit expressed by disabled readers. The specificity of language dysfunction related to reading inefficiency is conveyed by Tallal's finding (1987) that 85 percent of children exhibiting language disorders in the preschool years develop language-related learning disabilities (i.e., reading problems).

In addition to phonological and language-processing abilities, visual-feature analysis is also required in the act of grapheme-phoneme correspondence. Although low-level visual deficits have been identified among reading-disabled populations (Lovegrove, Martin, & Slaghuis, 1986), their impact as factors significantly impinging upon reading-skill development appears minimal (Hulme, 1988; Vellutino & Scanlon, 1987). In addition, select memory inefficiencies have been identified in some impaired readers. A sparse network of associations in working memory and retrieval deficits have been hypothesized in these instances. Employing hierarchical-regression analysis, Vellutino, Scanlon and Tanzman (1994) utilized measures of phonological coding and analysis abilities, verbal-memory measures, semantic- and syntactical-analysis tests, and visuo-perceptual task demands as dependent variables in predicting word-analysis proficiency. Phonological-processing skills accounted for the majority of the variance associated with word-identification proficiency. Semantic and syntactic measures were identified as intermediate predictors of reading proficiency and visual abilities.

Findings such as these have reshaped conventional wisdom applied to the assessment of reading disabilities. Dissociations between normal versus impaired readers have been traditionally specified by decision rules, i.e., aptitude-achievement test discrepancies. Children with a discrepancy between IQ and an objective reading measure are classified as disabled and deemed eligible for special-education supports. Low-achieving readers are conceptualized as reading below normative standards for age. But because of associated lower level IQ scores and imputed, more generalized cognitive inefficiencies, these low-achieving readers are not presumed eligible or appropriate for special-education services. These long-standing practices have led to a bimodal conceptualization of reading deficiencies, with reading disabilities representing a hump on the lower tail of this distribution (Shaywitz, Escobar, Shaywitz, Fletcher, & Makuch, 1992). Data derived from the Connecticut Longitudinal study (Shaywitz et al., 1992) support the contention of Stanovich (1991) that there actually may be no qualitative differences between disabled and low-achieving readers, and that phonological-processing deficits represent a core deficit indentifiable in both groups. These investigators have argued that reading abilities exist on a continuum which includes superior, average, and impaired readers. This model argues against the use of any arbitrary cut point indicating normalcy versus disability (i.e., discrepancy models) and instead suggests that intervention for any individual with reading inefficiencies be driven by identification of his or her unique processing deficiencies (phonological and associated cognitive limitations). As such, virtually all poor readers will exhibit a primary phonological-processing deficit. Variance in the expression of deficiencies in impaired readers, that is, heterogeneity in the expression of reading deficiencies, is a function of the severity of the core phonological processing deficits *and* the nature and extent of any underlying or associated cognitive dysfunction.

These findings have aided in validating Stanovich, Nathan, and Zolman's (1988) initial hypothesis regarding the variability expressed among impaired readers. This model presumes all disabled readers manifest a phonological-processing deficit. The most severe forms of reading disability are characterized by a fundamental or "core" deficit in the ability to establish grapheme-phoneme correspondence. As the nature of manifest impairment extends beyond core phonological-processing deficiencies, the term "variable" is attributed to the idiosyncratic manifestation of other language, attention, memory, or perceptuo-integrative skill deficits that may be additionally expressed. The model takes into account the remarkable heterogeneity expressed in reading deficiencies and why categorical models of reading, e.g, subtyping schemas, may not satisfactorily characterize the unique attributes expressed in individual cases.

Conceiving of reading problems in this fashion emphasizes the importance of defining the individual array of strengths and weaknesses expressed by any reader. This represents an alternative to models which posit more discrete subtypes of disabled readers and permits a means to conceive of reading on a continuum from normal variability in reading

proficiency to the heterogeneous expression of impaired reading development.

Assessment

A major portion of diagnostic reading assessment focuses on the sophistication and accuracy of decoding skills. This assessment is accomplished through the presentation of reading material as isolated phonemes, nonsense words, familiar and unfamiliar words, as well as words presented "in context," that is, in the form of sentences or complex paragraphs. At a first level of analysis the rule-out of basic visuo-perceptual dysfunction is necessary. The reader must be able to appreciate fully the visuo-symbolic configuration of letters and words. Here one is concerned with the rule-out of visual-sequential and modality-specific attentional deficits which could prevent the accurate assimilation of the written material. Perceptual errors such as reversals (reading "b" for "d" or "p" for "q") would also be excluded.

With the rule-out of primary perceptual dysfunction, analysis of grapheme-phoneme correspondence is undertaken. Basic decoding ability is ascertained for vowels, consonants, and consonant blends of letter combinations. Increasing the level of complexity of syllabic blends permits analysis of any sequential information-processing deficits that may be present. One is interested in the capacity not only to analyze and decode written material sequentially but aural material as well.

There are tasks which tap auditorization or syllabication, that is, the ability to decode the component phonetic properties of a word. On the Auditory Analysis Test, for example, one is asked to say "Germany" without the "ma" sound, thus transforming the remaining syllables to "journey." Some individuals, who on the Word Attack subtest of the Woodcock Reading Mastery Tests are reasonably successful in reading isolated phonemes, have great difficulty blending these same sounds into their appropriate phonological expression when confronting them in complex words. For example, when asked to read "phonological" the student struggles to isolate—"pho"..."no"..."loge"..."ee"..."cal"—only to pronounce the word then as "phonograph," a word more embedded in auditory memory. Frequently the effort required to analyze words laboriously in this fashion is exacted at great expense in terms of comprehension and memory for material read.

Assessment techniques that require rapid identification of words serve as a means to assess sight-recognition vocabulary. Speed of recognition is not factor-controlled in many types of reading tests. "Automatic recognition" represents the most sophisticated and efficient means of reading. Reading performed at this level taxes working memory minimally and frees the reader to focus on the semantic organization of the material for greater understanding and for committing textual information to memory. There are, however, individuals who have not attained adequate levels of sight-recognition skills. They maintain a more labored phonologically based reading style. These individuals may present a variety of deficits that impede their ability to process complex visuo-symbolic material. This might involve visual inattention, visuo-perceptual processing problems, spatial- or gestalt-recognition deficits, or weak visual memory. An analysis of the approach taken during "word attack" can be helpful in isolating the contributing deficit or deficits.

Within this context, the overall complexity of the word presented can be important. Errors encountered with relatively simple reading material can suggest problems in processing the basic visual morphology of written material. In terms of the simultaneous processing of visual input, there may be a finite limit on how complex a word can be for it to be realized. In attempts to compensate, some children "guess" at the whole word by processing only the prefix or first few syllables. Poor visual-gestalt functions or whole-word recognition skills are usually typified by gross lexical "word substitution" errors. Here words that share a similar visual gestalt to the word at hand are substituted, often resulting in flagrant misreading. In this regard it is necessary to rule out impulsivity as a contributing factor. The absence of other evidence of attention-deficit-disorder symptoms in ancillary testing or observation is particularly helpful.

Finally, comparisons of the relative efficiency of oral and silent reading under timed conditions can be potentially useful. A sample of oral reading of both word-recognition material and passage material can be extremely beneficial. Dramatic improvement in passage versus isolated-word reading immediately suggests the potential for the reader to compensate via the use of semantic cues. There are students whose oral-reading efficiency can be significantly compromised by anxiety or

inhibition. Far greater efficiency can be expressed by them in silent reading.

Reading Comprehension

Examination of reading comprehension is generally undertaken via the reading of a paragraph and the answering of questions about the content. A quantitative score is applied based on the number of correct responses and an estimation of reading-comprehension level is ascertained. This procedure does not in itself reflect the myriad factors which can contribute to comprehension difficulties. Level of investment can be a significant factor. Motivation can be influenced by interest in the factual material presented as well as general investment in reading as a preferred learning modality. Basic reading proficiency in terms of adequate word-recognition skills will also influence comprehension. Without strategies for the decoding of unfamiliar or complex reading material, adequacy of understanding will suffer. There are also a number of higher cognitive skills that influence performance, including linguistic proficiency, memory, cognitive flexibility, and semantic-organization skills. In order to ascertain where on a continuum of contributing factors comprehension problems lie, a number of informal strategies have been recommended to augment the reading-comprehension examination (Aaron & Poostay, 1982; Levine, 1987).

These strategies focus on the reading of restricted passages of known grade-level difficulty with the examiner focusing on a number of direct questions that permit an informal task analysis of potential contributions to comprehension failure. For example, Levine (1987) recommends beginning with the oral reading of simple sentences as the starting point. Limiting the amount of information to be assimilated restricts the degree to which active memory and semantic organizational skills are required, thus permitting direct access to potential problems based on decoding lexical information. At this level, basic questions regarding word-recognition errors, limited functional vocabulary, and problems with understanding morphology and syntax can be ascertained. Increasingly more complex lexical material is then presented. With each passage a number of profiles are presented in which the reader is asked to recall details, sequence events, and identify main ideas. More sophisticated demands can be made, such as summarizing the overall content of the passage. Responses can be evaluated on a continuum of literal to inferential depending on their level of complexity. It is also of value to compare general level of performance on oral comprehension and memory tests to determine whether reading comprehension is related to more generalized cognitive impairment.

The comprehensive evaluation of reading competencies requires utilization of diverse methodologies that typically involve an amalgam of standardized tests. In addition to academic achievement tests, language measures which tap lexical retrieval, semantic knowledge, linguistic short-term memory, as well as auditory comprehension are required. The Test of Language Development-Primary (Newcomer & Hammill, 1988), Test of Language Development-II (Intermediate) (Hammill & Newcomer, 1988), the Clinical Evaluation of Language Fundamentals-R (Semel, Wiig, & Selord, 1995), and Test of Adolescent and Adult Language-3 (Hammill, Brown, Larson, & Wiederholt, 1994) represent important adjunctive measures.

Thus, the primary role of the diagnostician should be geared less towards the documentation of any aptitude-achievement disparity and more upon the multivariate description of underlying cognitive processes contributing to reading impairment. While categorical diagnosis remains a requirement for eligibility determination within most classificatory systems, the potential explanatory power of assessment lies not with the discrepancy analysis, but with detailed multi-variate description of the constituent cognitive processes subserving reading. In this fashion, the cognitive basis of reading impairment can be linked empirically to remediation strategies.

Mathematical Abilities

In the past, an at-minimum fourth-grade math competency was considered adequate for adult functioning. Adaption to an increasingly technological society requires greater fluency in mathematics (Semrud-Clikeman & Hynd, 1992). The complexity of the subject matter, predominance of the use of a spiral curriculum, and other factors related to instructional technology have been related to trends noted in the preceding two decades that reflect lower overall math achieve-

ment in American children. Math-skill development is the subject of renewed interest.

Success in skill acquisition varies as a function of developmental stage, mastery of acquired-skill competencies, as well as a variety of intrinsic and extrinsic factors. Among intrinsic factors associated with underachievement, anxiety, negative self-attributions towards mathematics, and other motivational factors have been identified. Additional intrinsic or constitutional factors, such as the potential influence of heritability, remains essentially unknown. Multiple cognitive deficits have been imputed as potentially adversely impacting math-skill development. Early studies (Larsen & Hammill, 1975; McLeod & Crump, 1978) found general intellectual ability, visual perceptual and visuo-motor competencies, memory for visual sequences, verbal abilities, sequential-information processing, and comprehension and reasoning skills to be correlated with success in math performance.

Among those identified with math disabilities, Baker and Cantwell (1995) note comorbidity for reading disorders, disorders of written expression, expressive and receptive language disorders, and developmental coordination disorders. Attention-Deficit Hyperactivity Disorder (ADHD) represents the most commonly occurring Axis I diagnosis. Greater overall risk for social immaturity, school or personal adjustment problems, social skill deficits, anxiety, and depression have also been identified as risk factors expressed in this population.

Compared to the investigation of reading disabilities, the specification of integrated cognitive and neuropsychological models of math disability are lagging. Spreen and Haaf (1986) as well as Rourke and colleagues (see Rourke, 1993, for review), utilizing empirically derived clustering methodologies, have identified subtypes of mathematics impairment. There are, however, multiple sources of variability that may impinge on acquisition of mathematics competencies over the course of development. These include problems with decoding symbols; writing and copying numbers, appropriate sequencing and alignment of numbers; fact mastery; acquisition of the semantics of mathematics; memorization; capacity to convey multistep, sequenced cognitive operations; monitoring the quality of on-going performance; higher level linguistic competencies related to both reading and segmental language processing; development of spatial processing; as well as reasoning and abstract conceptual abilities (Semrud-Clikeman & Hynd, 1992). Heterogeneity in the expression of mathematics disabilities tends to be the rule rather than the exception.

Assessment

An excellent overview of the history of mathematics assessment as well as suggestions for assessment strategies can be found in the paper by Bryant and Rivera (1997). The authors suggest that in the field of learning disabilities, norm-referenced mathematics instruments play a vital role in developing a profile of strengths and weaknesses but that they are not intended as tools for instructional planning. "Rather, other math assessment practices, such as criterion-referenced testing, curriculum-based measurement, error analysis, clinical interviews, and so forth, can be used...to develop appropriate mathematics progress and to document student progress" (p. 66). Consistent with the remarks by Bryant and Rivera, the major objective of conducting diagnostic standardized testing in mathematics abilities is to ascertain areas of strength and deficits relative to developmental level. That is, assessment of math abilities should be developmentally driven. It requires knowledge of the developmental sequences of math concept development and an understanding of the curriculum demands faced by a youngster relative to age or grade-based expectations. Informal mathematical concepts and skills are acquired via spontaneous interaction with the environment during the preschool period, for example, acquisition of concepts of "more" and "less" conservation, additive and subtractive qualities within events, rudimentary counting and enumeration (Ginsburg, 1987). Levine, Jordan, and Huttenlocher (1992) demonstrated that as early as age 4 years, math strategy utilization is driven via well-established nonverbal conceptual abilities. It is not until age 5 or 6 years that conventional number fact or story problems can be assimilated. Thus, potential dissociations between nonverbal versus verbal conceptual abilities may contribute to divergence in developmental pathways associated with mathematics-skill development by the time a youngster reaches the primary grades. With the attainment of school age, mastery of basic conventions of number facts (counting and grouping), the alphanumeric symbol code of integers, as well as number alignment and place value are established. These skills permit mastery of written calculation. With advancing

age, mastery of more complex algorithms is achieved and the curriculum includes greater emphasis on concept development, mathematical reasoning and problem solving.

As previously noted, at different age levels, differing cognitive styles may be brought to bear in problem resolution. More than one means to go about solving a particular problem may be chosen. The level of sophistication of the processes brought to bear in task resolution can in itself be diagnostic. Even though a correct answer is ultimately obtained, the strategies utilized in reasoning may be developmentally deficient, hence affecting overall efficiency in performance. The capacity for greater "automatization" and use of "formal operations" with maturity is anticipated. Lack of expression of efficient problem-solving strategies can be diagnostically important. Standardized achievement tests are helpful, therefore, in identifying both the failure to develop appropriate numerical reasoning or problem-solving strategies as well as in identifying their types. As multiple pathways can lead to the expression of developmental arithmetic problems, it is important for the assessment of mathematical abilities to be tied to the larger domain of higher cognitive functioning.

According to Fleischner (1994), a core battery of achievement tests appropriate for the assessment of math disability should provide: (a) coverage in areas of conceptual understanding, conceptual proficiency and skill applications as well as (b) a means for the integration of qualitative error analysis and clinical interview procedures. The former measures permit normative comparisons requisite for the establishment of any severe discrepancy criteria as well as pattern analysis of errors. These later strategies offer insights into errors in thinking and strategy utilization. Criticisms have been leveled against a number of standardized assessment methodologies because the preponderance of coverage is limited only to computational measures, i.e., the Wide Range Achievement Test-3 (WRAT-3. Such screening measures are insufficient for the comprehensive diagnosis and description of mathematics disabilities (Romberg, 1992).

Examples of more comprehensive assessment methodologies include the Key Math-Revised: A Diagnostic Inventory of Essential Mathematics (Connolly, 1988), the Test of Mathematical Abilities (Brown, Cronin, & McEntire, 1994), and the Test of Early Mathematics Ability (2nd ed.) (Ginsburg & Baroody, 1990). These tests offer the opportunity for multiple observations of performance across developmentally age-appropriate measures of conceptual understanding, computational skills, and applied problem solving. Utilization of a diagnostic battery further permits the profiling of a pattern of performance which can then be correlated with cognitive and neuropsychological data.

In addition, a number of informal strategies should be objectified to aid in furthering the analysis of procedural deficits or problem analysis. Clinical interview as well as "testing of limits" procedures (Ginsburg, 1987) provides additional insights into the nature of identified problems with analysis and conceptualization. Levine (1994) provides one such structured interview that can be utilized to further pinpoint a subject's appreciation of the nature of his or her underlying math inefficiencies.

The augmentation of achievement-test data with the utilization of neuropsychological measures provides a means to correlate the nature of the manifest skills with any cognitive deficits that may contribute to learning problems. Further, such augumentation helps to define the procedural math deficits from an information-processing perspective as well as to establish a data base from which remediation or accommodations can be developed (Fletcher, Levin, & Satz, 1989).

Disorders of Written Expression

Disorders of written expression are identified by a demonstrable lag in the development of one or more components of writing. These could include deficits in spelling, punctuation, grammatical form and structure, or composition and organization. (Manifestations of only spelling deficits or legibility problems do not currently constitute a differential diagnosis of this disorder based on DSM-IV diagnostic criteria.) Objectification of criteria defining this form of learning disability has been challenging owing to the multi-factorial nature of writing. Intra-individual expression of a writing disorder may vary as a function of impairment of expression in any of the following areas: handwriting, spelling, language, attention and memory, written-narrative organizational skills, or metacognitive abilities.

No data are available that formally characterize the prevalence of this disorder are available. Gender influences remain ill-defined. Hooper et al.

(1994) suggest that prevalence may likely parallel the expression of a developmental language disorder. Determination of whether the deficits are primary or secondary to language or reading disabilities is integral for differential diagnosis. The occurrence of isolated disorders of written language is comparatively rare. The multi-dimensional nature of the neuropsychological underpinnings of writing suggest that deficits in any of the areas sufficient to impact writing would also impact other domains of academic performance. Thus, the evaluator must be prepared to identify a variety of potential co-morbid conditions. These include not only other learning disabilities but ADHD, depression, low academic self-esteem, anxiety, or thought disorder.

Developmentally, writing has been conceived of as the final pathway in the ontogeny of language (Johnson & Myklebust, 1967). Levine (1987) provides a heuristic model describing the progression of writing skills. Stage one begins with the establishment of basic graphomotor control (including drawing, tracing, and coloring abilities of preschoolers), attempts at proto-writing via pretend activities, and the initiation of writing training in first grade. Stage two concerns the honing of the basic orthographic skills related to letter and word formation as well as the establishment of greater graphomotor control. Stage three, associated with late second grade, is characterized by the progressive incorporation of skills such as capitalization, punctuation, syntax, and grammar (cursive writing is subsequently introduced). In the automatization stage, progressive mastery of primary competencies permits greater capacity for self-monitoring of the written product, expansion in the length of written expression, and utilization of more complex grammatical forms. In addition, planning and organizational skills begin to be incorporated into writing. In the elaboration stage (grades 7 through 9), the act of writing is sufficiently automatized in order to permit its use as a means to support the development of ideation. Greater capacity for ideational integration is expressed, and summarization skills are subsequently displayed. Capacity to form and express a viewpoint develops. In the final stage (9th grade and beyond), diversification and early development of a writing style are achieved. Writing progressively increases in versatility within such a model in order to not only augment communicative effectiveness (oral as well as written) but to support reasoning skills and creativity.

Although substantial intraindividual specification of deficits underlying writing abilities can be identified utilizing a comprehensive assessment approach, the identification of interindividual variability, or how patterns or subtypes of writing disorder manifest themselves remains more incompletely understood. The investigation of writing disability subtypes is in its infancy. One of the few comprehensive attempts is reflected in the work of Sandler et al. (1992). This factor notwithstanding, attempts have been made to formalize assessment methodologies based upon emerging empirical studies of writing disability (Berninger, 1994). Six components of writing assessment have been identified: (1) handwriting quality (legibility); (2) writing fluency (number of words copied within time constraints); (3) spelling from dictation; (4) spelling accuracy within composition; (5) compositional fluency (number of words produced within time constraints); and (6) compositional quality (content, cohesiveness, and organization of written narrative material).

The presence of identified hand-writing deficiencies requires assessment of motor, perceptual, and visuo-motor integrative competencies, as well as rule-out of any other pervasive developmental output failures. (Refer to specific recommendations for assessment of handwriting disorders by Bain, (1991). Skills essential in spelling include the mastery of grapheme-phoneme correspondence, overlearning the orthographic representation of word structures, and development of morphological knowledge. Strategies for assessment in this area have been covered in the reading section of this chapter (refer to phonological awareness and linguistic measures). Moats (1994) in her "Assessment of Spelling in Learning Disability Research" suggests among other things that a well-designed measure of spelling would sample "the broad domain of orthographic patterns, sound-symbol relationships, and morphophonemic patterns that must be learned by the writer of English" (p 335), while containing a wide enough range of items to accurately measure incremental development.

In addition to accrual of graphomotor samples and analysis of spelling errors, an evaluation of basic language competencies and reading skills is required. Proficiency in oral language has traditionally marked the starting point for the investigation of written language. As such, the language measures previously cited for use in assessment of linguistic underpinnings of reading disabilities

offer a means to assess semantic and general linguistic competencies (syntax) which are prerequisites for the conveyance of ideas within written form.

Developmental output failure in writing may also be a reflection of deficits in other aspects of higher cognitive functioning (Berninger, 1994). Attentional inefficiencies may be expressed as a function of input or output faculties. Deficits in the capacity to monitor quality of on-going cognitive performance impact writing demands which place a premium on simultaneous information-processing abilities. So too, executive function deficits would be expected to impact the planning and organizational skills essential in orchestrating complex ideation.

Larsen (1987) provides a review of 13 individual and group administered achievement tests based on methods of administration, test format, and coverage which are applicable to the assessment of writing abilities. The majority of these instruments provide only a cursory evaluation of writing abilities. Among traditional broad-focused academic achievement tests, the Woodcock Johnson Tests of Achievement-Revised (Woodcock & Johnson, 1989) offer the broadest coverage of conventional skills underlying writing. The Wechsler Individual Achievement Test (WIAT), while less comprehensive in assessment of core competencies related to the mechanics of writing, also offers a means to evaluate written expression within a standardized format.

Most psychoeducational screening batteries applied to the assessment of learning disabilities are limited in their sensitivity to the identification of developmental writing disorder features. When a writing disorder is suspected, a focused screening battery would consist of at minimum a spelling test (e.g., WRAT-3), the Proofing and Writing Samples subtest of the Woodcock Johnson Battery, and a sample of thematic writing, the WIAT Narrative Writing subtest. Vulnerabilities expressed on any of these measures would identify areas for more comprehensive testing. Unfortunately, the range of standardized tests available to assess written language remains restricted. No single standardized assessment tool comprehensively evaluates the heterogeneous language and cognitive deficits that characterize this disorder (Gregg, 1992). Of those systems available, many are time consuming and challenges associated with scoring can be daunting.

Two of the most frequently utilized standard written language tests are briefly summarized herein. The Test of Written Language-3 (TOWL-3) (Hammill & Larsen, 1995) was standardized for use with children from ages 7-6 through 17-11. It utilizes samplings of both spontaneous and contrived writing abilities. Eight subtests tap spelling, punctuation and capitalization, applications of semantic knowledge, syntax and grammatic cohesiveness. Factors relevant to story construction and thematic maturity are tapped. It provides an overall index of written language competency which can be contrasted against other standard scores related to intellectual ability or language mastery. *The Test of Early Written Language* (Hresko, Herron, & Peak, 1995) represents a downward extension of the TOWL-3. It was designed to assess emergent writing abilities. In addition to sampling linguistic skills, it taps discrimination of verbal and nonverbal representational forms as well as handwriting abilities. Like the TOWL-3, it provides a ready means for profile analysis.

Given the labor intensity of these direct assessment methodologies, an alternative to these fixed battery approaches is Berninger's Core Battery for Writing Assessment (1994). It utilizes the Writing Samples and Dictation subtests of the Woodcock Johnson to assess handwriting, the WRAT-3 Spelling Test to assess handwriting fluency and spontaneous spelling, and a variety of hybrid measures, for which norms are available for grades one through nine, to assess compositional fluency and quality.

Acknowledgement of writing disorders as conditions worthy of neuropsychological investigation has been slow in development. Although literature is established relating acquired agraphia to aphasia or apraxia in adulthood, this literature is not generalizable to developmental disorders. Within this context, writing problems continue to be conceptualized as conditions secondary to language, motor, attention, reading, or other deficits and not as multi-dimensional disorders worthy of investigation in their own right. As yet, a comprehensive model delineating the ontogeny of writing abilities remains to be developed.

ACHIEVEMENT TESTING WITH SPECIAL POPULATIONS

Exceptional Children

Under the educational opportunity safeguards included within Section 504 of the Rehabilitation Act, P.L. 94-142 and its amendments are specific components dealing with the process of evaluation. What is mandated by law is that all students who potentially have an educational disability receive a comprehensive evaluation that fairly assesses their abilities and does not discriminate against them because of cultural or racial factors or a disabling condition. Moreover, in all areas of exceptionality, federal and state legislation require the development of individualized education plans (IEPs) for handicapped students. Educational assessment data from standardized tests provide one necessary source of information used in the development of strategies for diagnostic prescriptive teaching. Here diagnostic achievement testing plays a particularly important role not only in identifying areas in need of remediation but also in placement and classification decisions. With the importance attached to assessment in the identification, diagnosis, placement, and instruction of children with disabling conditions, it is no surprise that the use of achievement tests, particularly the use of norm-referenced measures, has come under increasing criticism (Fuchs, Fuchs, Benowitz, & Barringer, 1987; Fuchs, Fuchs, Power & Darley, 1985; LaGrow & Prochnow-LaGrow, 1982; Ysseldyke, Algozzine, Regan, & Potter, 1980; Ysseldyke & Shinn, 1981).

Fuchs et al. (1987) conducted an extensive study of the 27 most well-known and commonly used tests in special education in order to determine the degree of participation of children with handicaps in the creation of test norms, and item selection, and in the establishment of their reliability and validity. Fourteen of these tests were measures of achievement classified as either screening (battery) or diagnostic (content specific). The user manual and/or technical supplement of each test was then analyzed in terms of (a) norms, (b) item development, (c) internal and test-retest reliability, and (d) concurrent and predictive validity. In only two of the achievement measures were children with handicaps included in the norming process and on only one measure were they included in item development. Otherwise, no other information was available. Such findings led the authors to state: "[I]f, in fact, test constructors have not validated their instruments for use with handicapped people, they 'should issue cautionary statements in manuals and elsewhere regarding confidence in the interpretation' based on these tests" (p. 269. Note: The quotation in Fuchs is taken from Standard 14.2, p. 79, the Standards for Educational and Psychological Testing, 1985).

Numerous studies have analyzed the performance on standardized tests of academic achievement of students with learning disabilities (Caskey, 1986; Estes, Hallock, & Bray, 1985; McGue, Shinn, & Ysseldyke, 1982; Shinn, Algozzine, Marston, & Ysseldyke, 1982; Webster, 1985), behavioral disturbances (Altrows, Maunula, & LaLonde, 1986; Eaves & Simpson, 1984), and hearing impairments (Allen, White, & Karchmer, 1983; Karchmer, Milone, & Wolk, 1979; Trybus & Karchmer, 1977), as well as students who are gifted (Karnes, Edwards, & McCallum, 1986). The findings from these studies and others demonstrate empirically (a) the variability in test results across achievement measures; (b) particular item biases where low socio-economic status (SES) is a factor; (c) the influence of the examiner on the testing process; (d) the differential effect of diagnosis and (e) the roles of time pressure, anxiety, and sex (Doolittle, 1986; Plass & Hill, 1986). It is critical that the professionals who utilize these tests be aware of the significant validity issues involved when assessing persons with disabilities or other areas of exceptionality.

Minority Children

Cautionary comments have been made also by those persons concerned with the standardized testing of minority students. Critics of the testing movement assert that tests which purport to measure achievement, among other things, are biased against certain ethnic/racial groups. Those in favor of testing regard test misuse as the real problem. Underlying the debate is the belief by the critics that the model used to assess performance and competence in society is monocultural. "A main criticism is that the model ignores the relevance of culturally different experiences that foster other equally important competencies essential to the survival of the group or individual" (Williams, 1983, p. 192). Similarly, Green and Griffore (1980) report that in one study 46 percent of the errors made on the Gray Oral Reading Test by

minority children were due to dialect differences. Others have suggested that lack of "test-wiseness" (Millman, Bishop, & Ebel, 1965) may serve to lower the scores of minority students on tests of aptitude and achievement. Johnson (1979), commenting about the variables that may invalidate test scores for African-Americans and other minorities, wrote:

> Many factors operate to attenuate or lower test scores, and these factors tend to have their greatest effects on Blacks and other minority applicants. These include factors which affect the actual performance of individuals on the test, such as socioeconomic status, differences in educational opportunity, motivation, narrowness of content of the tests, atmosphere of the testing situation, and the perceived relevance of the test to success. They also include factors that affect the test score more directly such as the composition of the group used for item tryouts and item selection and analysis which precede the actual standardization, composition of the standardization or normative group, and the techniques and procedures employed in item construction. Also, the validity or appropriateness of tests often differ for Black and white applicants, in relation to the same future performance of criterion. (p.3)

In addition, it has been substantiated that minority and white children are exposed to different curricula through the practice of ability tracking (Coleman, 1966; Findley, 1974; Green & Griffore, 1980; McPartland, 1969). Reviewers of the hundreds of ability grouping studies conducted since the 1920s have concluded that while superior students may benefit from this method of curricular offering, students with lower class ranking may not. The primary areas of concern are exposure to undemanding curricula and the social stigma attached to students in low-ability groups.

In a study by Abadzi (1985), the effects on both academic achievement and self-esteem of students placed in ability grouping classrooms were investigated with a population of 767 students from grades 4 to 8 in a large Texas school district. Contrary to earlier studies, her findings were that high-ability students did not maintain in the long run the performance gains made in the first year of grouping. Only the lower-level high-ability students in grouped classes were to benefit from the educational and social opportunity provided the highest-ability students. Students near the cutoff score in all groups were the ones most influenced by grouping in terms of both achievement and self-concept. Support for these findings was provided in spite of a general downward trend in performance at the end of elementary school that was characteristic of the school district's test scores and those of other districts as well. The author hypothesized that the steady drop in scores with the high-ability students may have been the result of reduced achievement motivation brought on by a "sense of invincibility, which the high status of the program combined with nonexistent exit criteria helped reinforce" (Abadzi, 1985, p. 39).

The concept of achievement motivation raised in Abadzi's conclusions has been systematically studied since the publication of David McClelland's *The Achievement Motive* (1953). This concept has been defined as a learned motive, unconscious in nature, resulting from reward or punishment for specific behavior. While studies utilizing this definition of achievement motivation have been conducted across racial groups, they have been criticized because of their ethnocentric design, methodology, and instrumentation. Castenell (1984) suggests that future research incorporate the definition espoused by Katz (1969) and Maehr (1974) which posits that (a) achievement motivation is conscious, (b) the need to achieve is universal to all groups, but (c) "because different groups have different life experiences it is likely that situations or a set of tasks will evoke different group responses" (p. 442).

While concerns have been raised with respect to standardized testing with minority students in general, we have not addressed the issues involved in standardized achievement testing with language minority students. In this regard we defer to a thorough discussion of this topic by Lam (1993) in which he suggests guidelines to consider in exempting limited English proficiency (LEP) students from standardized achievement testing and for the development of special testing for these students.

This section on special populations concludes with guidelines set forth by Williams (1983) that are highly reminiscent of the recommendations put forth in 1975 and cited at the beginning of this chapter. They would appear to encompass concerns regarding the use of achievement tests regardless of students' race, color, national origin, or handicap.

- Test constructors should foster an awareness of the limitations of the tests and the meaning attributed to test scores.

- Test constructors should educate their consumers in selecting tests in terms of particular goals and objectives of educational evaluation.
- Test constructors should bear responsibility for including minorities in all aspects of test development and not limit this to the standardization sample.
- Test consumers must assume some responsibility for developing skills in administering tests and interpreting results in light of the culturally diverse experiences that pupils bring into the testing situation.
- The educational community should minimize or eliminate intelligence testing or substitute approximately modified assessment techniques and interpretive procedures that consider cultural differences.
- The educational community should focus on achievement rather than intelligence or aptitude testing to eliminate pernicious connotations and unfair placement practices that limit future educational attainment and opportunity (p. 205).

THE FUTURE OF ACHIEVEMENT TESTS

Computer Adaptive Testing

The final section of this chapter is a discussion of the growth and impact of computerized adaptive testing on the measurement of achievement and what this product of modern technology means to the field of measurement. This is a fitting topic to conclude the previous narrative because computer adaptive testing is the direct result of advances in the fields of psychometrics, mathematics, cognitive learning theory, educational measurement, human engineering, and science technology. It relies as heavily on Glaser's criterion-referenced measurement as it does on Ausubel's cognitive approach to learning, Deno's curriculum-based measurement, Messick's concern with test validity, and Anastasi's continuum of testing.

Overall, educational research and development is most currently preoccupied with enhancing the instructional value of tests, or as Haney (1985) describes it, "making testing more educational" (p.4). He states that one need not be a dyed-in-the-wool social Darwinist to recognize that the use of standardized testing is increasing because it serves some important social functions. However, certain deficits that currently exist tend to negate the value of these tests: (a) Most testing programs violate the one nearly universal desideratum in all learning theories—in order to learn, an individual needs to receive rapid and specific feedback. (b) Most standardized tests have a very uncertain relationship to the specific teaching and learning that occurs in particular schools and classrooms. (c) The frequent concern to keep standardized testing programs secure limits their educational utility. It is these deficits, both narrowly and broadly defined, that the process of adaptive testing can be seen to rectify.

Adaptive testing is based on the premise that a measurement continuum should parallel a learning/teaching continuum, and if this learning continuum could be adequately measured by an underlying scale extending through its entire range, a student could enter and exit the measurement continuum at points appropriate to his or her current development regardless of age or grade levels (Forbes, 1986). This test development system is based on a measurement model popularly named the Rasch model after its originator. This model is also referred to as a one-parameter model in contrast to three-parameter models of latent traits which are based not only on item difficulty (single parameter) but also on item discrimination (slope of the difficulty) and on the level of change performance (guessing).

All item-response theory models must have an item data bank from which test items are drawn in the process of test construction. These items are computer stored and are then retrieved following a logical format. Utilizing a computer, the test can be presented to the student on a video screen with the computer keyboard serving as the response mechanism. Under such a procedure, the computer represents one pre-constructed test selected from a group of such tests. The test is tailored so that the computer "jumps" the person to the appropriate item-difficulty range and then gives a preselected sequence of items based on the correctness or incorrectness of the previous response. Generally, fewer items are required to measure performance at a predetermined level of measurement error than is the case with traditional testing procedures. Computerized adaptive tests have been shown also to take less than half the testing time required by traditional achievement tests and to provide more precise ability estimates across the entire ability range. Because the ability estimates and the item parameters are calibrated on a common scale, these estimates are theoretically

independent of the particular sample of persons taking the test and the particular sample of items selected by each examiner.

Seminal work done by Weiss (1980) focused on applying computerized adaptive testing to the measurement of achievement, using a methodology to extend beyond the aptitude measurement to which this type of testing had been limited previously. In addition to extending the use of item-characteristic curve theory (ICC) methods from ability testing to the problems of achievement testing, the project was also concerned with developing solutions to unique problems raised in achievement testing, that is, assessment in multiple content areas, mastery testing, the issue of stability of measurement over time, and the effects of immediate feedback as to the correctness or incorrectness of test responses. The findings of this three-year research project supported the use of ICC theory and methods and computerized adaptive testing for the measurement of achievement. However, many new questions were raised in addition to those originally addressed by the research that were in need of further study.

One of the first studies to compare and equate achievement scores from three alternative methods of testing, paper-administration, computer-administration, and computerized adaptive testing, was conducted by Olson (1986) with all students in grades 3 to 6 within three California school districts. A total of 575 students were involved in the study. Results of the study indicated that (a) analysis of variance showed no significant differences among the three measures in terms of the comparability of measurement precision; (b) computerized adaptive testing (CAT) required only one fourth of the testing time required by the paper-administrated test; (c) the computerized adaptive test provided a more precise ability estimate with smaller variance than either of the other two measures; and (d) the ability estimates calculated from a 20-item CAT tended to show more precision than tests of 55 to 62 items used with the other two measures.

Since that time, work has been done to investigate an innovative application of item-response theory (IRT) in computerized testing known as self-adapted testing (SAT). With this model, the difficulty levels of items administered are chosen by the student rather than by a computer algorithm, with positive results (Rocklin & O'Donnell, 1987). Rocklin and O'Donnell (1991) later reported that anxiety influenced student performance less on the SAT than on the CAT. Their results were later substantiated by Wise, Plake, Johnson, and Roos (1992) and Roos, Plake, and Wise (1992). Wise, Kingsbury, and Houser (1993) then experimented with a restricted form of SAT to provide students with control over the testing situation while preventing large mismatches between item-difficulty choice and proficiency level, which had shown itself to be a factor with a limited number of students in previous studies. At this point, use of the RSAT procedure is yet to be empirically evaluated.

This section on adaptive testing concludes with the futuristic predictions raised by Hsu and Sadock (1985) in their review, *Computer-assisted Test Construction: The State of the Art*. The authors foresaw the following as commonplace in testing of the future:

1. The development of item construction theories that take advantage of artificial intelligence and the phrase recognizability of the computer.
2. The development of item banks in the area of criterion-referenced achievement tests and in conjunction with textbook publication.
3. Item calibration and test design available on microcomputer.
4. The regular use of computers in test administration.
5. The application of IRT in test design by non-measurement specialists.
6. The use of computerized adaptive and diagnostic testing in the classroom.

Writing about achievement tests in the 1984 edition of the *Handbook of Psychological Assessment*, Fox and Zerkin concluded: "[While] standardized tests are not perfect and can be misused and misunderstood...they are currently the best instruments educators have available for assessing the quality of curriculum and for individualizing and improving instructional programs for each child" (p. 130). These conclusions no longer hold.

It is no longer possible to call these standardized measures of achievement the "best" instruments available. With the 1970s, criterion-referenced tests were touted as useful alternatives to norm-referenced tests. In the 1980s the new fields of cognitive sciences and computer technology were cited as likely sources for new and better test development. And now, in the late 1990s, performance assessment is popularly viewed as a remedy to the past ills of standardized tests. While numerous social problems are associated currently with the

use of more traditional testing procedures, and in particular multiple-choice tests, Haney, Madaus, and Lyons (1993) suggest that the negatives associated with their use may have more to do with the myriad functions that standardized tests are expected to perform. "To the extent that we regain more balanced approaches to assessment, reflecting a wider range of the modes by which we ought to judge student learning,...to that extent will the distortions now associated with standardized tests be reduced" (p. 294).

It is hoped that the present discourse has led the reader to question practices of the present because of knowledge of the past and to look to the future with eager anticipation. Tests can be a flexible passport into that future or a rigid barrier bound to the past. It is our job as professional educators, in the broadest sense, to insure the former. When describing the failure of the testing profession to inform the public about the meaning of "objective" standardized tests, Strenio (1981) states: "At a minimum, testers have an obligation to avoid placing their particular jargon in any context that makes it even harder for the layman to interpret than it already is" (p. 65). The authors of this chapter hope that they have not been guilty of this same failing.

> "Then you should say what you mean," the March Hare went on.
>
> "I do," Alice hastily replied; "at least—at least I mean what I say—that's the same thing, you know."
>
> "Not the same thing a bit!" said the Hatter; "why, you might just as well say that 'I see what I eat' is the same thing as 'I eat what I see!'".
>
> —Lewis Carroll
> Alice's Adventures in Wonderland

REFERENCES

Aaron, I., & Poostay, E. (1982). Strategies for reading disorders. In C. Reynolds & T. Gutkin (Eds.), *Handbook of school psychology*. New York: Wiley.

Abadzi, H. (1985). Ability grouping effects on academic achievement and self-esteem: Who performs in the long run as expected. *Journal of Educational Research, 79* (1),36–40.

Algozzine, B., Ysseldyke, J., & Shinn, M. (1982). Identifying children with learning disabilities: When is a discrepancy severe? *Journal of School Psychology, 20*(4), 299–305.

Allen, T. E., White, C. E., & Karchmer, M. A. (1983). Issues in the development of a special edition for hearing-impaired students of the Seventh Edition of the Stanford Achievement Test. *American Annals of the Deaf, 128*, 34–39.

Altrows, I. F., Maunula, S., & LaLonde, B. D. (1986). Employing teachers' ratings in selection of achievement tests in reading and mathematics with a behaviorally disturbed population. *Psychology in the Schools, 23*, 316–319.

Anastasi, A. (1982). *Psychological testing* (5th ed.). New York: MacMillan.

Anastasi, A. (1984). The K-ABC in historical and contemporary perspective. *Journal of Special Education, 18*(3), 357–366.

Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: MacMillan.

Anderson, R. C. (1972). How to construct achievement tests to assess retention of meaningful verbal material. *Review of Educational Research, 42*(2), 145–170.

Ausubel, D. P. (1960). The use of advance organizers in the learning and retention of meaningful verbal material. *Journal of Educational Psychology, 51*, 145–170.

Ausubel, D. P. (1968). *Educational psychology: A cognitive view*. New York: Holt, Rinehart, and Winston.

Baker, L. & Cantwell, D. P. (1995). Learning disorders, motor skills disorder, and communication disorder. In H. I. Kaplan & B. J. Sadock (Eds.) *Comprehensive textbook of Psychiatry VI*. Baltimore: Williams and Wilkins.

Bain, A. M. (1991). Handwriting disorders. In A. M. Bain, L. Lyons-Bailet, & L. Cook-Moats (Eds.), *Written language disorders: Theory into practice*. Austin, TX: ProEd.

Bartholomae, D. (1980). The study of error. *College Composition and Communication, 31*, 253–269.

Berninger, V. W. (1994). Future directions for research on writing disabilities: Integrating endogenous and exogenous variables. In G. R. Lyon (Ed.), *Frames of reference for the assessment of learning disabilities: New views on measurement issues*. Baltimore: Paul H. Brooks Publishing Co.

Berninger, V., Yates, C., Cartiwright, A., Rutberg, J., Remey, E., & Abbott, R. (1992). Lower-level developmental skills in beginning writing. *Reading and Writing: An Interdisciplinary Journal, 4*, 257–280.

Boder, E. (1973). Developmental dyslexia: A diagnostic approach based on three atypical reading-spelling patterns. *Developmental Medicine and Child Neurology, 15*, 663–687.

Brown, V. L., Cronin, M. E., & McEntire, E. (1994). *Test of mathematical abilities*. Austin, TX: ProEd.

Bryant, B. R., & Rivera, D. P. (1997). Educational assessment of mathematics skills and abilities. *Journal of Learning Disabilities, 30*(1), 57–68.

Buros, O. (1974). *Tests in print II.* Highland Park, NJ: Gryphon Press.

Caskey, W. E. (1986). The use of the Peabody Individual Achievement Test and the Woodcock Reading Mastery Tests in the diagnosis of learning disability in reading: A caveat. *Journal of Learning Disabilities, 19*(6), 336–337.

Castenell, L. (1984). A cross-cultural look at achievement motivation research. *Journal of Negro Education, 53*(4), 435–443.

Catts, H. W. (1989). Speech production deficits in developmental dyslexia. *Journal of Speech and Hearing Disorder.* S4, 422–428.

Chase, C. H., & Tallal, P. (1991). Cognitive models of developmental reading disorders. In J. E. Obrzut & G. W. Hind (Eds.), *Neuropsychological foundations of learning disabilities.* San Diego, CA: Academdic Press.

Chase, W. G., & Simon, H. A. (1973). A perception in class. *Cognitive Psychology, 1*, 55–81.

Clarizio, H. F., & Phillips, S. E. (1986). The use of standard scores in diagnosing learning disabilities: A critique. *Psychology in the Schools, 23*, 380–387.

Clarke, A. M. (1982). Psychology and education. *British Journal of Educational Studies, 30*(1), 3–56.

Coleman, J. S. (1966).*Equality of educational opportunity.* Washington, DC: U.S. Government Printing Office.

Coltheart, M. (1978). Lexical access in simple reading tasks. In G. Underwood (Ed.), *Strategies in information processing.* London: Academic Press.

Connolly, A. J. (1988). *Key math revised: A diagnostic inventory of essential mathematics.* Circle Pines, MN: American Guidance Service.

Curtis, M. E. (1980). Development of components of reading skill. *Journal of Educational Psychology, 72*(5), 656–669.

Curtis, M. E., & Glaser, R. (1983). Reading theory and the assessment of reading achievement. *Journal of Educational Measurement, 20*(2), 133–147.

Databank (1985). *Education Week, 5*, 16.

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children, 52*(3), 219–232.

Doolittle, A. E. (1986, April). *Gender-based differential item erformance in mathematics achievement items.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

DuBois, P. H. (1970). *A history of psychological testing.* Boston: Allyn & Bacon.

Eaves, R. C., & Simpson, R. G. (1984). The concurrent validity of the Peabody Individual Achievement Test relative to the Key Math Diagnostic Arithmetic Test among adolescents. *Psychology in the Schools, 21*, 165–167.

Estes, R. E., Hallock, J. E., & Bray, N. M. (1985). Comparison of arithmetic measures with learning disabled students. *Perceptual and Motor Skills, 61*, 711–716.

Findley, W. (1974). Grouping for instruction. In L. P. Miller (Ed.), *The testing of black students: A symposium.* Englewood Cliffs, NJ: Prentice-Hall.

Fischer, R. S. (1981). Measuring linguistic competence in a foreign language. *International Review of Applied Linguistics, 19*(3), 207–217.

Flavell, J. H. (1976). Metacognitive aspects of problem-solving. n L. B. Resnick (Ed.), *The nature of intelligence.* Hillsdale, NJ: Erlbaum.

Fleischner, J. E. (1994). Diagnosis and assessment of mathematics learning disabilities. In G. R. Lyon (Ed.), *Frames of reference for the assessment of learning disabilities: New views on measurement issues.* Baltimore: Paul H. Brooks Publishing.

Fletcher, J. M., Levin, H. S., & Satz, P. (1989). Neuropsychological and intellectual assessment of children. In H. I. Kaplan & B. J. Sadock (Eds.). *Comprehensive textbook of psychiatry V.* Baltimore: Willilams & Wilkins.

Fletcher, J. M., Shaywitz, S. E., Shankweiler, D. P., Katz, L., Liberman, I. Y., Stuebing, K. K., Francis, D. J., Fowler, A. E., & Shaywitz, D. A. (1994). Cognitive profiles of reading disability: Comparisons of discrepency and low achievement definition. *Journal of Educational Psychology, 86*, 6–23.

Foorman, B. F., Francis, D. J., Beeler, T., Winikates, D., Fletcher, J. M. (1997, Winter). Early interventions for children with reading problems: Study designs and preliminary findings. *Learning Disabilities, 8*(1), 63–71.

Forbes, D. W. (1986, April). The Rasch Model as a practical and effective procedure for educational measurement. In *Taming the Rasch tiger: Using item response theory in practical educational measurement.* Symposium conducted at the meeting of the National Council on Measurement in Education, San Francisco.

Forness, S., Sinclair, E., & Guthrie, D. (1983). Learning disability discrepancy formulas: Their use in actual practice. *Learning Disability Quarterly, 6*, 107–114.

Fox, L. H., & Zerkin, B. (1984). Achievement tests. In G. Goldstein & M. Hersen (Eds.), *Handbook of psychological assessment* (pp. 119–131). New York: Pergamon Press.

Fuchs, D., Fuchs, L. S., Benowitz, S., & Barringer, K. (1987). Norm-referenced tests: Are they valid for use with handicapped students? *Exceptional Children, 54*(3), 263–271.

Fuchs, D., Fuchs, L. S., Power, M. H., & Darley, A. M. (1985). Bias in the assessment of handicapped children. *American Educational Research Journal, 22*, 185–197.

Ginsburg, H. P. (1987). The development of arithmetic thinking. In D. D. Hammill (Ed.), *Assessing the Abilities and Instructional Needs of Students*. Austin, TX: ProEd.

Ginsburg, H. P., & Baroody, A. J. (1990). *Test of early mathematics ability*, (2nd ed.). Austin, TX: ProEd.

Glaser, R. (1963). Instructional technology and the measuremen of learning outcomes: Some questions. *American Psychologist, 18*, 519–521.

Glaser, R. (1981). The future of testing. A research agenda for cognitive psychology and psychometrics. *American Psychologist, 36*(9), 923–936.

Goh, D. S., Teslow, C. J., & Fuller, G. B. (1981). The practice of psychological assessment among school psychologists. *Professional Psychology, 12*, 696–706.

Green, R. L., & Griffore, R. J. (1980). The impact of standardized testing on minority students. *Journal of Negro Education, 49*, 238–252.

Gregg, N. (1992). Expressive writing disorders. In S. R. Hooper, G. W. Hynd & R. E. Mattison (Eds.), *Developmental disorders: Diagnostic criteria and clinical assessment*. Hillsdale, NJ: Lawrence Eribaum Associates.

Guion, R. M. (1976). The practice of industrial and organizational psychology. In M. D. Dunnette (Ed.), *Handbook of industrial an organizational psychology*. Chicago: Rand McNally.

Hambleton, R. K. (1980). Test score validity and standard sellin methods. In R. A. Berk (Ed.), *Criterion-referenced measurement: The state of the art*. Baltimore: Johns Hopkins University Press.

Hambleton, R. K., & Cignor, D. R. (1978). Guidelines for evaluating criterion-referenced tests and test manuals. *Journal of Educational Measurement, 15*, 321–327.

Hammill, D. D., Brown, V. L., Larson, S. C., & Wiederholt, J.L. (1994). *Test of adolescent and adult language*. Austin, TX: ProEd.

Hammill, D., & Larsen, S. C. (1995). *Test of written language*, (3rd. ed.). Austin, TX: ProEd.

Hammill, D. D., & Newcomer, P. L. (1988). Test of languag development-Intermediate. (2nd ed.). Austin, TX: ProEd.

Haney, W. (1985). Making testing more educational. *Educational leadership, 43*(2), 4–13.

Haney, W. M., & Madaus, G. F. (1989). Searching for alternatives to standardized tests: The whats, whys and whithers. *Phi Delta Kappa, 70*(9), 683–687.

Haney, W. M., Madaus, G. F., & Lyons, R. (1993). *The fractured marketplace for standardized testing*. Norwell, MA: Kluwer Academic Publishers.

Hardman, M. L., Drew, C. J., Egan, M. W., & Wolf, B. (1993). *Human exceptionality: Society, school and family* (4th ed.). Boston: Allyn and Bacon.

Harnisch, D. L. (1985). *Computer application issues in certification and licensure testing* (ERIC Document Reproduction Service No. ED 261 079)

Henderson, E. H., & Beers, J. W. (Eds.) (1980). *Developmental and cognitive aspects of learning to spell*. (ERIC Document RIE Jan. 1986.)

Herman, J. L., Abedi, J., & Golan, S. (1994). Assessing the effects of standardized testing on schools. *Educational and Psychological Measurement, 54*(2), 471–482.

Hooper, S. R., Montgomery, J., Swartz, C., Reed, M. S., Sandler, A. D., Levine, M. D., Watson, T. E., & Wasileski, T. (1994). Measurement of written language expression. In G. R. Lyon (Ed.), *Frames of reference for the assessment of learning disabilities: New views on measurement issues*. Baltimore, MD: Paul H. Brooks Publishing.

Hoover, H. (1984). The most appropriate scores for measuring educational development in the elementary schools: GE's. *Educational measurement: Issues and practices, 3*, 8–14.

Hresko, W. P., Herron, S. R., & Peak, P. K. (1995). *Test of early written language*. (2nd ed.), Austin, TX: ProEd.

Hsu, T., & Sadock, S. F. (1985). *Computer-assisted test construction: The state of the art*. ERIC Clearinghouse on Tests, Measurement, and Evaluation, Princeton, NJ.

Hulme, C. (1988). The implausability of lowlevel visual defects as a cause of children's reading difficulties. *Cognitive Neuropsychology, 5*, 369–374.

Hunter, D. R., & Burke, E. F. (1987). Computer-based selection testing in the Royal Air Force. *Behavior Research Methods, Instruments, and Computers, 19*(2), 243–245.

Improving America's School: A newsletter on issues in school reform (1996, Spring).

Johnson, D. J., & Myklebust, H. (1967). Learning disabilities: Educationalprinciples and practices. New York: Grune and Stratton.

Johnson, S. T. (1979). *The measurement mystique*. Washington, D.C.: Institute for the Study of Educational Policy.

Karchmer, M., Milone, M., & Wolk, S. (1979). Educational significance of hearing loss at three levels of severity. *American Annals of the Deaf, 124*, 97–109.

Karnes, F. A., Edwards, R. P., & McCallum, R. D. (1986). Normative achievement assessment of gifted children: Comparing the K-ABC, WRAT, and CAT. *Psychology in the Schools, 23*, 346–352.

Katz, I. (1969). A critique of personality approaches to Negro performance, with research suggestions. *Journal of Social Issues, 25*, 13–27.

Kingsbury, G. G., & Weiss, D. J. (1979). *An adaptive strategy for mastery decisions-Research Report 79–5. Computerized adaptive performance evaluations: Final report, February 1976 through January 1980.* Minnesota University, Department of Psychology (Contact #N00014-76-C-0627). Arlington, VA: Office of Naval Research, Personnel and Training Research Programs Office.

LaGrow, S. J., & Prochnow-LaGrow, J. E. (1982). Technical adequacy of the most popular tests selected by responding school psychologists in Illinois. *Psychology in the Schools, 19*(2), 186–189.

Lam, T. (1993). Testability: A critical issue in testing language minority students with standardized achievement tests. *Measurement and Evaluation in Counseling and Development, 26*, 179–191.

Larkin, J., McDermott, J., Simon, D. P., & Simon, H. A., (1980).Expert and novice performance in solving physics problems. *Science, 208*, 1335–1342.

Larsen, S. C. (1987). Determining the presence and extent of writing problems. In D. D. Hammill (Ed.), *Assessing the instructional abilities and instructional needs of students*. Austin, TX: ProEd.

Larsen, S. C., & Hammill, D. D. (1975). The relationahip of selected visual-perceptual abilities to school learning. *Journal of Special Education, 9*, 281–291.

Larry P. v. Riles, 343 F. Supp. 1306 (N.D. Cal. 1972).

Lazarus, M. (1981). *Goodbye to excellence: A critical look atminimum competency testing*. Boulder, CO: Westview Press.

LeMahieu, P. G. (1984). The effects on achievement and instructional content of a program of student monitorin frequent testing. *Educational evaluation and policy analysis, 6*(2), 175–187.

Levine, M. (1987). Developmental variation and learning disorders. Cambridge, MA: Educators Publishing Service.

Levine, M. (1994). *Educational care: A system for understanding and helping students with learning problems at home and in school.* Cambridge, MA: Educators Publishing Service.

Levine, S. C., Jordan, N. C., & Huttenlocher, J. (1992). Development of calculation abilities in young children. *Journal of Experimental Child Psychology, 53*, 72–103.

Lezak, M. (1983). *Neuropsychological assessment* (2nd ed.). New York: Oxford University Press.

Liberman, I. Y., Liberman, A. M., Mattingly, I., & Shankweller, D. (1980). Orthography and the beginning reader. In J. F. Kavanagh & R. L. Venezky (Eds.), *Orthography, reading and dyslexia.* Baltimore: University Park Press.

Linn, R. L. (1980). Issues of validity for criterion-referenced measures. *Applied Psychological Measurement, 4*(4), 547–561.

Linn, R. L. (1982). Two weak spots in the practice of criterion-referenced measurement. *Educational Measurement*, Spring, *12–13*, 25.

Lovegrove, W., Martin, F., & Slaghuis, W. (1986). A theoretical and experimental case for a visual deficit in specific reading disability. *Cognitive Neuropsychology 3*, 225–267.

Lundberg, I., Frost, J., & Petersen, O. (1988). Effects of an extensive program for stimulating phonological awareness in preschool children. *Reading Research Quarterly, 23*, 263–284.

Maehr, M. (1974). *Sociocultural origins of achievement.* Monterey, CA: Brooks-Cole.

Mann, V. (1994). Phonological skills and the prediction of early reading problems. In N. C. Jordan & J. Goldsmith-Phillips (Eds). *Learning disabilities: New directions for assessment and intervention.* Boston: Allyn and Bacon.

Marton, F., & Saljo, R. (1976a). On qualitative differences in learning-I: Outcome and process. *British Journal of Educational Psychology, 46*(1), 4–11.

Marton, F., & Saljo, R. (1976b). On qualitative differences in learning-II: Outcome as a function of the learner's conception of the task. *British Journal of Educational Psychology, 46*(2), 115–127.

McClelland, D. C., Atkinson, J. S., Clark, R. A., & Lowell, E. L. (1953). *The achivement motive*. New York: Appleton-Century-Crofts.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception. Part I. An account of basic findings. *Psychological Review, 88*, 375–407.

McLeod, T. M., & Crump, W. D. (1978). The relationship of visuospatial skills and verbal ability to learning disabilities in mathematics. *Journal of Learning Disabilities, 11*, 237–241.

McGue, M., Shinn, M., & Ysseldyke, J. (1982). Use of cluster scores on the Woodcock-Johnson Psychoeducational Battery with learning disabled students. *Learning Disability Quarterly,5*, 274–287.

McPartland, J. M. (1969). The relative influence of school desegregation and of classroom desegregation on the academic achievement of ninth-grade Negro students. *Journal of Social Issues, 25*, 93–102.

Mehrens, W. A., & Lehmann, I. J. (1975). *Standardized tests in education.* New York: Holt, Rinehart, and Winston.

Messick, S. (1965). Personality measurement and the ethics of assessment. *American Psychologist, 20*, 136–142.

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist, 30*, 955–966.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35*(11), 1012–1027.

Millman, J. (1973). Passing scores and test lengths for domain-referenced measures. *Review of Educational Research, 43*, 205–216.

Millman, J., Bishop, C., & Ebel, R. (1965). An analysis of test-wiseness. *Educational and Psychological Measurement, 25*, 707–726.

Mirkin, P., Deno, S., Tindal, G., & Kuehnle, K. (1982). Frequency of measurement and data utilization as factors in standardized behavior assessment of academic skill. *Journal of Behavioral Assessment, 4*(4), 361–370.

Mitchell, J. V. (Ed.). (1985). *The ninth mental measurement yearbook.* Lincoln, NE: University of Nebraska Press.

Moats, L. C. (1994). Assessment of spelling in learning disabilities research. In R. G. Lyon (Ed.), *Frames of reference for the assessment of learning disabilities.* Baltimore: Paul H. Brookes Publishing Co.

Morton, J. (1969). The interaction of information in word recognition. *Psychological Review, 76* 165–178.

National Commission on Excellence in Education. (1983a). *A nation at risk.* Washington, DC: U.S. Government Printing Office.

National School Boards Association (1977). *Standardized achievement testing* (Report No. 1977–1). Washington, DC: Author.

Newcomer, P. L., & Hammill, D. D. (1988). *Test of language development-Primary* (2nd ed.). Austin, TX: ProEd.

Nolen, P., & McCartin, R. (1984, November). Spelling strategies on the Wide Range Achievement Test. *The Reading Teacher,* 148–158.

Olson, J. B. (1986, April). *Comparison and equating of paper-administered, computer-administered and computerized adaptive tests of achievement.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

P. L. 94-142. The Education for All Handicapped Children Act of 1975, 20 U.S.C. SS1401 *et seq.*, 45 C.F.R. 121(a).

Plass, J. A., & Hill, K. T. (1986). Children's achievement strategies and test performance: The role of time pressure, evaluation, anxiety, and sex. *Developmental Psychology, 22*(1), 31–36.

Popham, W. J. (1971). *Criterion-referenced measurement.* Englewood Cliffs, NJ: Educational Technology Publications.

Radencich, M. C. (1985). BASIS: Basic Achievement Skills Individual Screener. *Academic Therapy, 20*(3), 377–382.

Reisman, F. (1982). Strategies for mathematics disorders. In C. Reynolds & T. Gukin (Eds.), *Handbook of school psychology.* New York: Wiley.

Reynolds, C. R. (1984). Critical measurement issues in learning disabilities. *Journal of Special Education, 18*(4), 451–476.

Rocklin, T., & O'Donnell, A. M. (1987). Self-adapted testing: A performance-improving variant of computerized adaptive testing. *Journal of Educational Psychology, 79*, 315–319.

Rocklin, T., & O'Donnell, A. M. (1991, April). *An empirical comparison of self adapted and maximum information item selection.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Romberg, T. A. (1992). Evaluation: A coat of many colors. In T. A. Romberg (Ed.), *Mathematics assessment and evaluation: Imperatives for mathematics educators.* Albany, NY: State University of New York.

Roos, L. L., Plake, B. S., & Wise, S. L. (1992, April). *The effects of feedback in computerized adaptive and self-adapted tests.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Rothkopf, E. Z. (1965). Some theoretical and experimental approaches to problems in written instruction. In J. D. Krumbaltz (Ed.), *Learning and the educational process* (pp. 193–221). Chicago: Rand McNally.

Rourke, B. P. (1993). Arithmetic disabilities, specific and otherwise: A neuropsychological perspective. *Journal of Learning Disabilities, 26*, 214–226.

Sandler, A. D., Watson, T. E., Footo, M., Levine, M. D., Coleman, W. L., & Hooper, S. R. (1992). Neurodevelopmental study of writing, disorders in middle childhood. *Developmental and Behavioral Pediatrics,* 13, 17–25.

Sax, G. (1974). *Principles of educational measurement and evaluation.* Belmont, CA: Wadsworth.

Section 504 of the Rehabilitation Act of 1973, 29 U.S.C. 794, 45 C.F.R. 81, 84.

Semel, E., Wiig, E. H., & Selord, W. A. (1995). *Clinical evaluation of language fundamentals.* (3rd ed.). San Antonio, TX: Psychological Corporation.

Semrud-Clikeman, M., & Hynd, G. W. (1992). Developmental artihmetic disorder. In S. R. Hooper, G. W. Hynd, & R. E. Mattison (Eds.), *Developmental disorders: Diagnostic criteria and clinical assessment.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Shaw, S. F., Cullen, J. P., McGuire, J. M., & Brinkerhoff, L. C. (1995). Operationalizing a definition of learning disabilities. *Journal of Learning Disabilities, 28*(9), 586–597.

Shaywitz, S. E., Escobar, M. D., Shaywitz, D. A., Fletcher, J. M., & Makuch, R. (1992). Evidence

that dyslexia may represent the tower tall of a normal distribution of reading ability. *New England Journal of Medicine,* 326, 145–150.

Shaywitz, S. E., Fletcher, J. M, & Shaywitz, B. A. (1996). A Conceptual model and definition of dyslexia: Findings emerging from the Connecticut longitudinal study. In J. H. Beitchman, N. J. Cohen, M. M. Konstantaveas, & R. Tannock (Eds.), *Language, learning, and behavior disorders: Developmental, biological, and clinical perspectives.* Cambridge, England: Cambridge University Press.

Shepard, L. (1983). The role of measurement in educational policy: Lessons from the identification of learning disabilities. *Educational Measurement: Issues and Practice, 2*(3), 4–8.

Shepard, L. A., & Smith, M. L. (1983). An evaluation of the identification of learning disabled students in Colorado. *Learning Disability Quarterl, 6*(2), 115–127.

Shinn, M., Algozzine, B., Marston, M. A., & Ysseldyke, J. (1982). A theoretical analysis of the performance of learning disabled students on the Woodcock-Johnson Psycho-educational Battery. *Journal of Learning Disabilities, 15*(4), 221–226.

Siegler, R. S. (1976). Three aspects of cognitive development. *Cognitive Psychology, 5,* 481–520.

Snyder, T. D. (Ed.). (1987). *Digest of education statistics.* Washington, DC: U.S. Government Printing Office.

Spreen, O., & Haaf, R. G. (1986). Empirically derived learning disability subtypes. A replication attempt and longitudinal patterns over 15 years. *Journal of Learning Disabilities,* 19, 170–180.

Stanovich, K., Nathan, R., & Zolman, I. (1988). The developmental lag hypothesis in reading: Longitudinal and matched reading-l\evel comparisons. *Child Development, 59*(1), 71-87.

Stanovich, K. E. (1991). Discrepancy definitions of learning disability: Has intelligence led us astray? *Reading Research Quarterly,* 26, 7–29.

Stevenson, D. K. (1983). Foreign language testing: All of the above. In C. J. James, *Practical applications of research in foreign language teaching.* Lincolnwood, IL: National Textbook.

Strenio, A. J. (1981). *The testing trap.* New York: Rawson, Wade.

Tallal, P. (1987). *Developmental language disorders: Interagency Committee on learning disabilities- Report to the US Congress.*

Terry, R. M. (1986). Testing the productive skills: A creative focus for hybrid achievement tests. *Foreign Language Annals, 19*(6), 521–528.

Traub, R. E., & Rowley, G. L. (1980). Reliability of test scores and decisions. *Applied Psychological Measurement,4,* 517–545.

Trybus, R. J., & Karchmer, M. A. (1977). School achievement score of hearing-impaired children: National data on achievement status and growth patterns. *American Annals of the Deaf, 122,* 62–69.

Vellutino, F., & Scanlon, D. (1987). Phonological coding, phonological awareness, and reading ability: Evidence from a longitudinal and experimental study. *Merrill-Palmer Quarterly,* 33, 321–363.

Vellutino, F. R., Scanlon, D. M., & Tanzman, M. S. (1994). Components of reading ability: Issues and problems operationalizing word identification, phonological coding, and orthographic coding. In G. R. Lyon (Ed.), *Frames of reference for the assessment of learning disabilities: New views on measurement issues.* Baltimore: Paul H. Brooks Publishing Co.

Wasik, B. A., & Slavin, R. E. (1993). Preventing early reading failure with one-to-one tutoring: A review of five programs. *Reading Research Quarterly, 28,* 179–200.

Webster, R. E. (1985). The criterion-related validity of psychoeducaional tests for actual reading ability of learning disabled students. *Psychology in the Schools, 22,* 152–159.

Weiss, D. J. (1980). *Computerized adaptive performance evaluation: Final report, February 1976 through January 1980.* Minnesota University, Department of Psychology, Arlington, VA: Office of Naval Research, Personnel and Training Research Programs Office.

Wells, P. (1991). Putting America to the test. *Agenda, 1*(Spring), 52–57.

Williams, T. S. (1983). Some issues in the standardized testing of minority students. *Boston University Journal of Education, 165*(2), 192–208.

Wilson, V., & Reynolds, C. (1984). Another look at evaluating aptitude-achievement discrepancies in the diagnosis of learning disabilities. *Journal of Special Education, 18*(4), 477–494.

Wise, S. L., Kingsbury, G. G., & Houser, R. L. (1993). *An investigation of restricted self-adapted testing.* Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta, GA.

Wise, S. L., Plake, B. S., Johnson, P. L., & Roos, L. L. (1992). A comparison of self-adapted and computerized adaptive tests. *Journal of Educational Measurement, 29*(4), 329–339.

Woodcock, R. (1987). *Woodcock Reading Mastery Tests*. Circle Pines, MN: American Guidance Service.

Woodcock, R., & Johnson, M. (1989). *The Woodcock Johnson psychoeducational battery-revised.* Allen, TX: DLM.

Ysseldyke, J. E., Algozzine, B., Regan, R., & Potter, M. (1980). Technical adequacy of tests used by professionals in simulated decision making. *Psychology in the Schools, 17*(2), 202–209.

Ysseldyke, J. E., & Shinn, M. R. (1981). Psychoeducational evaluation. In J. M. Kauffman & D. P. Hallahan (Eds.), *Handbook of special education* (pp. 418–440). Englewood Cliffs, NJ: Prentice-Hall.

CHAPTER 8

EVALUATION OF APTITUDES

Daniel J. Reschly
Carol Robinson-Zañartu

Aptitude assessment and intervention have a long and distinguished role in the clinical evaluation of children and youth with learning and behavioral problems. This chapter will review the major approaches to assessment of aptitudes with emphasis on ways the assessment information is used in decisions about diagnosis, placement, and treatment. The chapter content does not include information on aptitude assessment with adults, nor the uses of aptitude information in career counseling or vocational guidance.

A critical theme in our review is treatment validity, that is, is treatment effective (or more effective) when based on a conception of aptitude and guided by the associated assessment procedures? This criterion requires consideration of models, conceptions, assessment procedures, treatment approaches, and intended outcomes. The latter is especially central in this review. Intended outcomes, or criteria for validity, must be an integral part of the evaluation of different aptitude models.

Several traditional models, although intellectually attractive and logically persuasive, fall short on the treatment-validity criterion. Their use in educational or clinical situations must be regarded as questionable *if* benefits to children and youth cannot be demonstrated. Several models will be described and critically reviewed. Major attention will be devoted to mediated learning, a model with enormous potential that has been examined by scholars in several western nations.

CONCEPTS OF APTITUDE, INTELLIGENCE, AND ACHIEVEMENT

Aptitude, intelligence, and achievement as psychological constructs or types of tests are not easily distinguished. The traditional distinction was that achievement tests reflected the effects of past learning, whereas aptitude and intelligence reflected the individual's potential for success. In this traditional view, both aptitude and intelligence were seen as relatively enduring traits of the individual, not easily modified by experience or special training. In some instances both aptitude and intelligence-tests results were regarded as indications of innate capacity.

These traditional meanings of aptitude, intelligence, and achievement tests were rejected in all the leading measurement texts published in the last decade (Anastasi, 1997; Brown, 1983; Cronbach, 1990). All now are viewed as tests of developed abilities, that is, they reflect the effects of experience, and, as maximum performance measures, it is assumed that the individual is encouraged to try as hard as possible to do well.

The most important differences among aptitude, intelligence, and achievement have to do with how they are used and with assumptions about *antecedent experiences* (Anastasi, 1997; Brown, 1983). Achievement tests are assumed to measure past learning that occurred in a specific teaching or instructional situation. In contrast, aptitude has a future reference. The aptitude concept involves inferences about performance in future learning or

training situations. Intelligence is usually seen as between achievement and aptitude on the continua of test use and antecedent experiences. Intelligence has a present reference as a reflection of the effects of general, broad learning experiences. When intelligence tests are used in diagnoses, a future reference is assumed because predictions typically are made about the continuing status of the individual.

As a construct, aptitude often is used quite broadly, especially in theory and research on aptitude by treatment interactions (Cronbach & Snow, 1977; Snow, 1980, 1992). Here, aptitude is virtually any psychological characteristic of the person that predicts differences among people in later learning or training situations. Included in this very broad conception of aptitude are cognitive abilities and processes and personality and emotional characteristics (Snow, 1992). Although this broad conception of aptitude is used in this chapter, more of the content will come from examination of cognitive abilities and processes than from emotional or personality characteristics.

COMMON FEATURES OF APTITUDE MODELS AND ASSESSMENT

The varying aptitude models used with children and youth have a number of common features. Perhaps the most basic commonality is the goal of improving diagnosis, placement, and treatment decisions with children and youth who exhibit varying degrees and kinds of learning problems.

Expanded Consideration of Cognitive Processes

Proponents of different aptitude models share a commitment to assessing cognitive processes that are not directly represented on conventional measures of intelligence and achievement. Numerous observers have noted the rather narrow range represented or the limited opportunity to observe distinct cognitive processes on traditional measures (e.g., Naglieri, 1989; Woodcock & Mather, 1989). Widely used measures of intellectual functioning such as the Wechsler Scales (Wechsler, 1974, 1991) reflect a rather narrow range of cognitive processes. Further, the Wechsler items involve a wide variety of complex tasks that require the simultaneous use of two or more cognitive processes rendering difficult the observation of distinct processes.

Some more recently developed measures, such as the Kaufman Assessment Battery for Children (K-ABC) (Kaufman & Kaufman, 1983), provide a slightly broadened array of cognitive processes and better opportunities to observe specific processes. The K-ABC and similar instruments, however, remain primarily as measures of general intelligence with limited specific information on processes that are drawn from a single model of aptitude. The recently published Cognitive Assessment System (Naglieri & Das, 1997) likewise reflects a single cognitive model and a limited array of processes.

Intra-individual versus Inter-individual Differences

The aptitude models and measures typically focus more heavily on intra than inter-individual differences. Intra-individual differences involve variations in the pattern of cognitive processes within the individual. For example, the individual's mean on several measures often is used as the point of comparison in assessing strengths and weaknesses in cognitive processes. Substantial variations from this mean then may be translated into goals for cognitive training or recommendations for appropriate instructional methodology. The inter-individual interpretation method typically used with conventional measures of intelligence and achievement examines differences between individuals. Here, the individual's variation from a population mean is the primary reference point for interpreting performance. Inter-individual approaches to interpretation are useful for determining level of performance in comparison to others while intra-individual approaches yield information on pattern of performance within the individual.

Improved School Achievement

Aptitude assessment and intervention models share the common goal of improving academic achievement. Indeed, aptitude assessment and intervention usually is initiated because an individual student is having difficulty with acquiring basic literacy skills in reading, writing, or mathematics. Although different aptitude models use

quite different assessment and treatment procedures (see next section), an overall goal is improved academic functioning.

Improved Overall Cognitive Functioning

The aptitude models have the common goal of improved overall cognitive functioning, although it should be noted that there are quite different assumptions about how best to produce improved general functioning (see next section). Treatment, based on assessment of aptitudes, is thought to lead to improved functioning either through direct changes in cognitive processing or through the contribution to cognitive functioning that occurs with improved acquisition of academic skills. Rapidly changing technology leading to the dawn of the "information age" throughout the world has made advanced thinking processes, cognitive flexibility, and rapid acquisition of new skills imperative goals for all children and youth in modern societies. Aptitude assessment and treatment attempts to address these needs as well, through direct or indirect procedures for improving overall cognitive functioning.

DIFFERENCES AMONG APTITUDE MODELS

Aptitude models vary significantly regarding assessment procedures, treatment techniques, and intended outcomes. These variations involve fundamental elements, including (a) assumptions about stability or modifiability; (b) product versus process of cognitive activities; (c) transfer of effects; (d) standardized versus dynamic assessment procedures; and (e) applications to decisions.

Stability versus Change in Aptitudes

The most basic difference is the assumption about whether aptitudes are relatively stable traits of the individual or whether aptitudes can be changed through treatment. In the former view, aptitudes are important individual characteristics that form the basis for relatively enduring recommendations for program placement or instructional methodology. Since the cognitive or neurological processes that underlie the aptitudes are believed not to change, better cognitive and academic performance is sought through matching the individual's permanent aptitude characteristics to curricula or methodology that capitalize on intact processes or realistic goals. If aptitudes are viewed as unchanging, then treatments to change aptitudes (remediate deficits) is regarded as futile (Reynolds, 1981, 1986, 1992).

In contrast, those who contend that aptitudes can be changed through treatment use initial assessment results as the starting point for the design of experiences that will modify the individual's basic ways of thinking. Here, the goal is nothing less than the modification of thinking and, thereby, the development of aptitudes previously not observed. Expanding the cognitive repertoire is seen both as possible and essential. It is difficult to think of a more basic difference among models than whether aptitudes are changeable. This fundamental point leads to numerous additional differences.

Product versus Process Orientation

The primary observational unit varies among aptitude models. Product oriented models focus on whether or not the individual can correctly perform certain tasks that are assumed to be reflections of underlying aptitudes. For example, can the individual correctly solve problems that make primary demands on visual-spatial processing or presumed right cerebral cortex functions? In contrast, aptitude models that are more process oriented attempt to examine underlying cognitive processes or thinking skills that the individual uses to achieve right or wrong answers to tasks. In process orientation the answer itself often is less important than the thinking skills that produced the answer. The process-product distinction is related both to the stability-modifiability assumption and the questions of transfer of training, kind of assessment, and intended outcomes.

Transfer Questions

The transfer question involves the issue of how the aptitude information is used and the effects of its use. In models that assume stability of aptitudes with a focus on products of cognitive activity, the transfer question is: do aptitude strengths translate to more successful academic learning through matching aptitude strengths to instructional methodology, or through designing curricula to match

aptitude levels and patterns? It is assumed that the aptitudes that are assessed do underlie instruction or control responses to different curricula. If these aptitudes do not transfer from assessment to school-learning tasks, the aptitude assessment as well as the instructional and curricular recommendations lack treatment validity.

The transfer question is different for the models that assume that aptitudes are modifiable. In these models cognitive processes observed in the individual's efforts to solve task-specific problems then become goals for cognitive modification treatments. Near transfer is demonstrated subsequent to treatment if the individual can perform more successfully new examples of the same or highly similar tasks. Such positive change suggests that cognitive processes have been modified, at least within a specific kind of task.

The cognitive modification is, however, limited unless far transfer can be demonstrated. Far transfer is the question of whether cognitive modifications in the individual lead to more successful performance in different problem-solving situations that involve new tasks or stimulus properties. Ultimately, most cognitive modifiability models seek better performance in generalized problem solving and learning, including the acquisition of academic skills. For obvious reasons, it is considerably easier to produce near than far transfer.

Standardized versus Dynamic Assessment

The rules that are established for valid and reliable assessment differ substantially among aptitude models. In modifiability models, the examiner engages in a complex interaction with the student wherein the purpose is to establish how the individual thinks and the strategies that individual uses to approach new learning, solve problems, and communicate ideas and findings. Accomplishing this purpose necessarily requires a format that can be at best partially structured. The assessment is "dynamic" in that what the examiner does is determined by the kinds of thinking processes used by the student, interacting with the student to test hypotheses regarding the reasons for correct and incorrect problem solutions, and using near-transfer tasks to further refine hypotheses and explore effective interventions. Although modifiability clinicians may use a limited array of problem solving tasks, the presentation of the tasks and the interaction with the student nearly always varies from case to case.

Models that assume aptitude stability typically use conventional standardized procedures; that is, they attempt to achieve uniformity in task presentation, evaluating responses, summing scores, and interpretation of score meaning. The relationship of clinician and student is assumed to be standardized as well, and similar to the relationship that is used in individual assessment of intelligence or achievement.

Application to Decisions

The major aptitude models are used primarily in cases where the child or adolescent is not acquiring academic skills to the degree expected. Assessment of aptitudes leads to decisions about these students, but the kinds of decisions vary significantly by model. Application of the stability models typically leads to decisions about the kind of instruction that should be used, most often in reading. Clinicians working with teachers attempt to match aptitude strengths to instructional methodology, assuming that the best match will lead to more efficient learning. In other instances, aptitude strengths and weaknesses may be used as part of classification and placement decisions, for example, to diagnose low achievement as stemming from a learning disability or dyslexia and, then, based on this diagnosis, placing the student in some kind of remedial or special education program.

Modifiability models are less likely to be used in classification and placement decisions, or in the prescription of specific curricula. Decisions subsequent to application of assessment are more likely to involve prescriptions for particular kinds of training in thinking processes and problem-solving strategies and their application to teaching methodology. That methodology embeds cognitive processes in the teaching of academic skills. The assessment, however, continues as an integral part of the cognitive training. Teaching new thinking processes or strategies occurs simultaneously with ongoing assessment of problem-solving competencies such that the teaching-assessment process is continuous and inseparable.

MODELS OF APTITUDE ASSESSMENT AND INTERVENTION

Three models of aptitude assessment and intervention are reviewed in this section. Each of the models is described using the common features and differences discussed in the previous section, after which the criterion of treatment validity is applied to each.

Psycholinguistic and Perceptual-Motor Models

Psycholinguistic (PL) and perceptual-motor (PM) assessment and intervention are the oldest, best researched, and most controversial of the aptitude models used currently with children and youth. The models make similar assumptions about cognitive processes, interventions, and anticipated effects on academic achievement. Due to these similarities, both are discussed in this section.

The primary PL model (Kirk, McCarthy, & Kirk, 1968), based on a communication theory (Osgood, 1957), had three major components: (a) channels of communication (auditory-vocal and visual-motor); (b) communication process (reception, association, and expression); and (c) levels of organization (representational and automatic-sequential). Kirk and colleagues developed the Illinois Test of Psycholinguistic Abilities (ITPA) to assess these components of language processes and several volumes were published to guide intervention efforts (Bush & Giles, 1977; Kirk & Kirk, 1971; Minskoff, Wiseman, & Minskoff, 1972). The ITPA subtests and the associated intervention procedures attempted to address the following cognitive processes: auditory reception, visual reception, auditory association, visual association, verbal expression, manual expression, grammatic closure, visual closure, auditory sequential memory, visual sequential memory, auditory closure, and sound blending. Adequate functioning on these processes was assumed to be required for acquisition of literacy skills in reading, writing and, to a lesser extent, mathematics.

The perceptual-motor assessment and intervention model emphasized processes such as visual and auditory discrimination and perception, visual-motor coordination and integration, visual-auditory integration, and motor skills. Some of the models claimed direct relationships between neurological functioning and motor-perceptual awareness, leading to interventions such as walking on balance beams, vestibular stimulation from movement of the entire body in space, and precise large-motor exercises. Other PM-model variations placed more emphasis on pencil-and-paper tasks designed to improve visual-motor skills or auditory exercises to improve discrimination and recognition of sounds. All variations of PM assumed relationships between these skills and academic achievement.

The PL and PM models are used to identify intra-individual differences in processes presumed to underlie overall cognitive functioning and school achievement. The models generally are used with younger children (aged 2 to 10 years) who have been identified as delayed in cognitive development, or as experiencing difficulty in acquiring beginning literacy skills, especially reading. Intervention is usually guided by careful, standardized assessment of perceptual-motor or psycholinguistic strengths and weaknesses, followed by specific teaching activities designed to overcome weaknesses.

Clearly, the PL and PM models assume that basic cognitive processes could be identified accurately and improved through systematic instruction. Transfer to improved school achievement is assumed in a chain of logic that proceeds through the following assumptions: (a) PL or PM processes underlie and are a prerequisite to successful school learning; and (b) untreated PL- or PM-process deficits would remain as barriers to, and improved PL or PM processes will lead to, more successful achievement. Many educational programs for young children with learning problems continue to make these assumptions and emphasize PL training.

In addition to remedial programming for young children, the PL and PM models are highly influential in the diagnosis of specific learning disabilities (SLD). The most widely used SLD *conceptual* definition contains the following language, "a disorder in one or more of the basic psychological processes involved in understanding or using language, written or spoken" (Mercer, King-Sears, & Mercer, 1990; Reschly & Gresham, 1989). This conceptual definition appears in U.S. law and is adopted in many states. Although, the *classification criteria* described in federal and state law typically do not require identification of PL- or PM-processing deficits as part of the diagnosis of SLD (Mercer et al., 1990), the PL and PM models con-

tinue to be highly influential regarding thought about the causes of SLD.

The PL and PM models dominated thought and practice in SLD until about 1980. Reviews of research on the assessment of PL processes and the outcomes of PL interventions began to appear in the mid-1970s, leading to diminished use. The Illinois Test of Psycholinguistic Abilities and various tests of PM processes were severely criticized on psychometric criteria, especially the low reliabilities on subtests that were used to diagnose weaknesses and prescribe interventions (Salvia & Ysseldyke, 1988). If the subtest reliabilities were low, then intra-individual strengths-and-weaknesses results were, by necessity, inaccurate and the prescriptions for interventions were based on faulty information. Even more devastating, though more controversial, were a number of separate reviews of the effects of PL and PM interventions. Hammill and Larsen (1974, 1978) and Newcomer, Larsen, and Hammill (1975), concluded that PL interventions had little positive effects on improving PL processes, and no documented positive effects on school achievement. The conclusions about the effects of PL interventions were disputed by Minskoff (1975), and later by Lund, Foster, and McCall-Perez (1978), in a series of increasingly heated debates with Hammill and his associates. Later examination of the same body of literature using the technique of meta-analysis (Glass, 1983) led to slightly more positive conclusions about interventions with some PL processes (verbal expressions, manual expression, visual closure, and auditory association) (Kavale, 1981,1990), but no solid evidence has been provided confirming that improved PL processes lead to improved academic achievement.

The treatment-validity evidence regarding the PM model is even more negative. Kavale and Mattson's (1983) meta-analysis of some 180 studies led to the disappointing conclusion that PM interventions had no effect on PM processes and no identifiable beneficial effects on school achievement. PM assessment and training appears to be a waste of teachers' and students' valuable time (Kavale, 1990).

The PM and PL models have strong intuitive appeal. Most of the cognitive processes identified in these models are logically related to school achievement and overall cognitive functioning. Several possible explanations exist for the failure to unequivocally establish gains in either the processes or school achievement from PL and PM interventions. First, the theory simply may be wrong; that is, the PL and PM processes may not be essential for overall cognitive functioning and school achievement. If so, it would not be the first time that an intuitively appealing idea was incorrect. Second, the essential PL and PM processes may not be assessed accurately by current measures. Several authors have noted the psychometric deficiencies of measures in these areas. Third, the interventions may not be sufficiently powerful to produce the effects that would be required to produce PL or PM gains and improved school achievement. Regardless of which explanation(s) is/are correct, continued use of PL and PM models for assessment and intervention in clinical or educational settings is questionable.

Aptitude by Treatment-Interaction Models

Three prominent treatment-by-aptitude-interaction (ATI) models are used by many clinicians and educators today. All specify relationships between cognitive processes and the methodology used to teach cognitive and academic skills.

The modality-matching model generally focuses on three kinds of information-processing strengths and weaknesses: (a) auditory, (b) visual, or (c) kinesthetic. Children experiencing difficulty in acquiring academic skills are assessed to determine strengths and weaknesses over these processes, and then an instructional method that utilizes the child's strengths is prescribed. The same procedures are used in the second and third models, cognitive style and neuropsychological. In the latter model additional inference(s) is/are made about underlying brain functioning. The actual prescriptions for instructional methodology are highly similar across the three ATI models (Reschly & Gresham, 1989).

The ATI models see aptitudes as relatively unchanging, even to the point in the neuropsychological variation of cautioning against "teaching dead tissue" (Hartlage & Reynolds, 1981). Focusing interventions on deficit areas, as is the case in the PL and PM models, is seen as inefficient and perhaps futile. Aptitudes typically are assessed with conventional standardized measures of cognitive functions including individually administered measures of general intellectual functioning. Two-well standardized general intelligence measures have been developed with primary attention to diagnos-

ing elements of cognitive style (Naglieri & Das, 1997; Kaufman & Kaufman, 1983). The ATI models are used primarily to design instruction with an underlying assumption that matching process strengths will generalize to higher learner performance in instructional settings. Neuropsychological concepts also have appeared in 1980s definitions of SLD. The phrases "presumed to be due to central nervous system dysfunction" and "presumed neurological origin" appeared recently in two separate definitions formulated by learning-disabilities advocacy groups (Reschly & Gresham, 1989). These phrases and neuropsychological diagnostic criteria have not appeared in federal or state SLD definitions or diagnostic criteria (Mercer et al., 1990).

Modality matching, cognitive style determination, and neuropsychological assessment and intervention depend heavily on the existence of aptitude by treatment interactions (ATI). The ATI notion has enormous intuitive appeal. Reynolds (1992) described this process as, "Instruction... formatted around the child's best developed processes, avoiding those that are poorly developed or inept" (p. 10). According to a survey by Arter and Jenkins (1977), 99 percent of teachers believe that there are differences in how children process information, and that instruction will be more effective if instructional materials and methods are matched to modality or neuropsychological strengths.

The case for ATI was stated by Cronbach (1957) in a highly influential and widely-cited article that appeared in the *American Psychologist*, the largest circulating psychology journal in the world. Cronbach asserted that aptitudes could be measured accurately and that, "For any potential problem, there is some best group of treatments to use and best allocation of persons to treatments" (p. 680). The allocation process meant matching individuals' aptitude strengths to treatments that utilize those strengths through differential stimulus properties or instructional methods.

Although the potential list of aptitudes that might be used in matching is nearly unlimited, the different instructional methodologies are much more limited. The typical matching procedure involves aptitudes such as auditory, visual, and kinesthetic processes, cognitive styles such as simultaneous and successive (Das, Kirby, & Jarman, 1979) or sequential and simultaneous (Kaufman, Goldsmith, & Kaufman, 1984), or neuropsychological constructs such as right hemisphere and left hemisphere functioning. The different instructional methodologies prescribed for children with these aptitude strengths are highly similar (Reschly & Gresham, 1989). Phonic methods

Table 8.1. Aptitude by Treatment Interaction

		Aptitude	
		Left Hemisphere Auditory-Vocal Successive	Right Hemisphere Visual-Motor Simultaneous
Instructional Method	Phonic	Match Method to Strength Presumed Maximum Benefit Actual-No Effect	Mismatch Presumed Minimal Effect Actual-No Effect
	Sight	Mismatch Presumed Minimal Effect Actual-No Effect	Match Method to Strength Presumed Maximum Benefit Actual-No Effect

Notes: Assumption: Matching aptitude with treatment (instruction) produces maximum benefits.
Empirical basis: Weak for underlying process or neuropsychological strenghts and weaknesses.

of teaching reading and overall emphasis on auditory cues typically are prescribed for children believed to have strengths in auditory processing, sequential or successive cognitive styles, or left hemisphere functions. Similarly, whole-word methods of teaching reading via visual cues are stressed for children with strengths in visual processing, simultaneous cognitive styles, or right hemisphere functioning.

For these models to have treatment validity, there must be an interaction between the presumed aptitude strength and instructional methodology. For example, children with right hemisphere strengths must learn more efficiently when instructional materials and methods are selected and presented in ways to utilize that strength, and conversely, such students must do less well if instructional methodology is not matched to strengths (see Table 8.1).

Unfortunately, the research-to-date does not support the existence of significant treatment by aptitude interactions. Some of the difficulties with ATI research were summarized by Cronbach (1975) in another *American Psychologist* article in which he expressed doubt about ever being able to use matching in clinical or educational settings. Based on 18 years of largely unsuccessful ATI research, Cronbach concluded that, "Once we attend to interactions, we enter a hall of mirrors that extends to infinity" (p. 119). The major problems in the ATI research were: (a) non-existent or very weak interactions; that is, matching strengths had, at best, small and inconsequential effects: (b) results over studies were enormously inconsistent; and (c) there were higher order interactions, that is, complex three- and four-way interactions, that would be impossible to apply to practical clinical and educational problems.

The current research on modality matching and neuropsychological assessment and prescriptions fits the pattern described by Cronbach (1975). Matching presumed strengths with instructional methodologies does not lead to demonstrable differential gains in academic achievement, regardless of whether the aptitude strengths are conceptualized as modality preferences (Kavale & Forness, 1987, 1990; Kavale, 1990), cognitive styles (Ayers & Cooley, 1986; Ayers, Cooley, & Severson, 1988), or neuropsychological functions (Reschly & Gresham, 1989; Teeter, 1987, 1989). Despite the negative evidence, the modality matching, cognitive style, and neuropsychological-functioning approaches to assessment and intervention continue to be used widely in a variety of settings by psychologists and educators.

The hall of mirrors that Cronbach described in 1975 continues to confound efforts to apply an inherently sensible idea, that is, selecting and implementing instructional methodology that utilizes an individual's cognitive-processing strengths. Reasons similar to those that may account for the negative-treatment validity evidence on the PL and PM models are relevant to the ATI models. The problem(s) may reside with the basic theory, the measures of aptitudes, or the presently available interventions. Reschly and Gresham (1989) noted deficiencies in all of these areas. The models often are predicated on rather primitive theories of neurological functioning or information processing. The determination of strengths usually involves intense analyses of profiles of scores on different subtests or measures from different standardized batteries. The profile-difference scores that are fundamental to determination of strengths and weaknesses typically have low reliabilities and other psychometric deficiencies (Macmann & Barnett, 1994a, 1994b; McDermott, Fantuzzo, Glutting, Watkins, & Baggaley, 1992; McDermott, Fantuzzo, & Glutting, 1992). Perhaps most important, the interventions are limited and not very powerful. Regardless of which explanation(s) is/are correct, and whether any of these deficiencies can be overcome, current use of any of the ATI models in diagnosis and treatment of learning problems is highly questionable.

Dynamic Assessment/Change Models

Dynamic assessment and mediated learning intervention models presume that learning abilities, or aptitudes, are modifiable rather than stable. Specifically, within this paradigm, what had been previously conceived of as largely genetically determined and stable within the human organism is believed to be open to change with human intervention. Through specific mediations, actual modification of cognitive structures and motivational factors are believed to influence the manner in which the individual is able to approach new situations; thus, the acquisition of knowledge. The function of assessment in this model is to identify efficient and inefficient cognitive and motivational parameters of the individual; the parameters accessible to mediation; and the kind and

Table 8.2. Contrasting Paradigms

STABILITY MODELS	CHANGE MODELS
Closed system	Open system
Stable invariant human characteristics	Modifiable human characteristics
Assume intelligence primarily inherited	Assume intelligence dependent on person-environment interactions
Relatively insensitive to educational and cultural influences	Specific investment in cognitive structures can produce enhanced functioning
Presumed upper limits; reduced environmental demand	Upper limits not presumed; demand designed to challenge evolving cognitive and knowledge structures
Focus on measurement, classification prediction, replication	Focus on dynamic interaction between context and cognition; intervention inherent in assessment
Passive acceptant approach	Active modification approach

intensity of mediated interventions needed to produce change in accessible parameters. Determining accessibility of cognitive parameters and the measurement of their changes occur as mediated intervention is provided. In contrast to conventional forms of assessment that typically produce classifications, labels, and predictions based on the belief in stable individual differences, dynamic assessments are designed to identify the targets and methods of intervention for enhancement of individual functioning across classroom, home, and community settings.

The notion of testing the ability to learn while observing learning-in-progress emerged near the early part of the century (Dearborn 1921; De Weerdt, 1927; Penrose, 1934). The idea that observing the results of deliberate stimulation of learning would yield important data and information for actually developing aptitude has its early roots in the work of Vygotsky (1934/1962) and Rey (1934). Vygotsky's (1962, 1978) cultural-historical theory of human mental development, the genetic epistemology of Piaget (1952), Luria's (1966a, 1966b) neuropsychological investigations of brain-behavior relationships, Schwartz's (1977, 1983) psychosocial-neurophysiological model of self-regulation, and Feuerstein's (1970; Feuerstein, Jensen, Hoffman, & Rand, 1985) theory of mediate-learning experiences and structural cognitive modifiability (Feuerstein, Rand, Jensen, Kaniel, & Tzuriel, 1987; Jensen & Feuerstein, 1987) were all part of this evolution of ideas (Jensen, Robinson-Zañartu, & Jensen, 1992).

The Range of Dynamic Models

A number of models of "dynamic assessment" have been developed, each of which attends to the assessment of intra-individual differences. However, there are significant differences between the models, which range from a modified testing-the-limits approach (Carlson & Wiedl, 1976, 1978, 1979) to the structural cognitive modifiability theory of Feuerstein and his colleagues (Feuerstein, 1970; Feuerstein, Haywood, Rand, Hoffman, & Jensen, 1985) and Jensen (1990; 1992). Underlying theoretical assumptions, measurement, examiner-examinee interactions, goals for change, number, and types of parameters targeted for intervention, and assumptions regarding transfer effects vary widely across these models, and may be conceptualized on a continuum (see Table 8.2).

On one end of the range are those that maintain psychometric standardization and insert training or direct instruction in problem solving between test trials (Budoff, 1987a, 1987b; Budoff & Friedman, 1964; Carlson & Wiedl, 1978, 1979). Campione & Brown (1987), strongly influenced by the work of Vygotsky (1978) and neo-Vygotskians in Russia, observed the effects of instruction on targeted tasks, focusing on "readiness to learn." Their work seems to presume that one cannot actually influence the readiness, but only locate it and determine who will profit most from instruction. The assessed measure of gain, which they refer to as dynamic assessment, is presumed to have greater predictive utility than the initial unaided level of performance. They remark that although more clinically based procedures may in fact yield richer

information, their choice was an approach that could yield strong quantitative data.

Feuerstein's (1970; Feuerstein, Haywood, Rand, Hoffman, & Jensen, 1985) structural cognitive modifiability outlined a far more clinical model. Here, using a series of nonacademic, or de-contextualized tools, designed to tap cognitive skills in various modalities, the examiner used both specific interactive behaviors known as mediation and some 30 specific cognitive functions to elicit and attempt to re-form cognitive habits. The functions identified as basic to that individual's enhanced cognitive functioning, and the specific mediations found effective in the enhancement process, then became the targets of intervention, assuming these new skills would then transfer to new tasks. Feuerstein's work has permeated the educational communities across North America, South America (e.g., Venezuela, Chile), Africa (South Africa) and Europe with training and applications of his Learning Potential Assessment Device (LPAD) and companion Instrumental Enrichment (IE) intervention program. These have been used not only with the low-performing population for which it was originally designed, but with bilingual and gifted children as well. Feuerstein's work has been perhaps the most controversial of these (change) models because of its radical departure from long-standing concepts of stable individual differences, assumptions about upper limits of individual potential, and reliance on psychometric measurement. Jensen (1990; 1992) extended this change model, addressing not only cognitive functions, but knowledge-structure development, as well. We will refer especially to the latter change models in the remaining discussions of the nature of dynamic assessment and mediated learning interventions.

Goals of Enhanced Cognitive Functioning

Interventions designed to accompany dynamic assessments presume to modify the nature of the individual's cognitive functioning or learning processes over time. Initially, mediated learning interventions depart from the use of (contextualized) tasks with specific academic context such as reading or arithmetic in the attempt to target underlying cognitive skills without the interference of motivational barriers. They rely instead on using problem-solving tasks which lend themselves to interaction with the examiner, and are designed to require specific, often progressively complex or abstract cognitive skills. Through specific "mediational" interactions, fragile areas of cognitive functioning (e.g., comparative behavior, systematic exploratory behavior; use of two or more sources of information) are gradually modified and new habits or skills formed. These new skills are then gradually tested and applied in settings of increasingly distant transfer, with application to the curriculum an example of far transfer.

A major goal in the mediation or training of cognitive skills is their transfer to new situations, thus enhancing the ease and flexibility with which learners approach new information and problem-solving. It is in the arena of near transfer to other problem-solving situations, including presumed measures of intelligence, that a fair amount of evidence supports the efficacy of mediated learning to date (Babad & Budoff, 1974; Budoff, 1987a; Carlson & Wiedl, 1978, 1980; Feuerstein, Rand, Hoffman, & Miller, 1980; Johnson, 1996; Klauer, 1989; Lidz & Peña, 1996). For instance, Thickpenny & Howie (1990) evaluated the effects of teaching thinking skills to deaf adolescents and found significant gains on two subtests of the WISC-R as well as on the Matching Familiar Figures test. Campione & Brown (1987) summarized three sets of studies of mediated learning, in which they found that near transfer was substantial for "low ability" students. Johnson (1996) reported significant increases in Full-scale IQ scores on the WISC-R following a semester of mediated learning instruction in a pilot study with low-functioning children.

In one of the most rigorous examinations of these issues, Jensen & Singer (1987) measured the effects of Feuerstein's IE Program, using the full three-year program with 234 experimental and 164 control low-functioning adolescents. They provided clear evidence for the acquisition and near transfer of new cognitive functions (Jensen & Singer, 1987). In addition, Jensen (1990) found that a factor analysis of these functions loaded into four categories; the first three, which he termed reception, transformation, and communication, closely paralleled Feuerstein's clinical grouping of input, elaboration, and output, thus contributing to the validity of Feuerstein's grouping and processes. The fourth, termed cognitive control, was associated with the role of impulsivity and control over the tempo of the mental act. However, and of great importance, far transfer to significant academic improvement was not found (Jensen & Singer, 1987). Shortly thereafter, Jensen (1990,

1991) postulated that a review of the research must lead us to conclude that cognitive enrichment programs by themselves have not been effective in producing better academic outcomes. This research led to his later postulations regarding the development of cognitive and knowledge structures (Jensen, 1992; Jensen, Robinson-Zañartu, & Jensen, 1992).

Goals of Enhancing School Achievement

Although researchers working within a dynamic assessment/mediated learning paradigm have found significant changes in the manifest levels of cognitive functioning of both school age and adult learners, evidence suggests that far transfer from enhanced problem-solving to enhanced academic performance does not automatically transfer to content areas, but must be deliberately taught. Jensen (1992) has defined this problem as the attempt to "proceduralize" knowledge, or actually infuse new cognitive processes into content or knowledge areas. Beginning in the late 1980s, other researchers who had worked with dynamic models of cognitive modifiability (e.g., Greenberg, 1990; Harth, 1982; Haywood, Towery-Woolsey, Arbitman-Smith, & Aldrudge, 1988; Perkins, 1987; Salema & Valente, 1990) were proposing that instructional models must be developed that actually infused and deliberately worked on transfer to the school curricula.

Some researchers have targeted specific cognitive skill development within specific content areas. Salema & Valente (1990), for instance, examined effects of systematic teaching of thinking skills and metacognition in developing composition skills in Portuguese among low achievers. They reported a significant difference in experimental and control groups in learning to write compositions. Consistent with the far-transfer question, that shift did not transfer into other subject matter. Krieglar & Kaplan (1990) used an abbreviated form of Feuerstein's Instrumental Enrichment (IE) Program to try to demonstrate a bridge to reading achievement, assuming that inattention, hyperactivity, and impulsivity were intervening variables that interfered with reading performance and could be mediated. They created cognitive links to a specific reading task and found significant differences between experimental and control groups on (a) teacher ratings of attention, (b) reading accuracy, and (c) the Porteus Maze test.

Perhaps the most broadly applied of these models to date is Greenberg's (1990, 1992, 1994; Greenberg, Coleman, & Rankin, 1993) Cognitive Enrichment Network (COGNET) Educational Model. COGNET gives teachers a set of cognitive strategies and mediating skills to infuse within and across the curriculum. Models, known as mini-lesson plans, are adapted for each teacher's curricular needs. In addition, the program encourages and models the incorporation of cooperative learning strategies concurrent with the mediated learning strategies. Greenberg's (1994) research on the effects of the program in four different types of schools reported that high-risk students in the COGNET schools made greater gains overall than comparison groups on standardized tests of basic skills.

The Dynamic Assessment and Mediation of Cognitive and Knowledge Structures

Mediated learning and dynamic assessment as described in Modifiability Enhancement Theory (MET), the change model proposed by Jensen (1992; Jensen, Robinson-Zañartu, & Jensen, 1992), is useful for exploring the integration of thinking skills into academic areas, as it provides: (a) a description of the relationship and rationale of the content-process link in learning, (b) a process by which assessment is linked with intervention, and (c) incorporation of the contextualizing culture of the child (Robinson-Zañartu & Cook-Morales, 1992). Further, MET posits that the proceduralization of knowledge structures is a highly specific process; thus, automatic transfer cannot be assumed, but must be facilitated.

Knowledge structures are proposed in MET to be formed via a process called proceduralization. In proceduralization, the factual knowledge base of any given content area is woven together with cognitive structures, motivational factors, and personality attributes to form a set of processes that enable efficient collection, transformation, and communication of information within that content area. Cognitive functions, such as comparative behavior, planning, and strategies for inferential thinking, are theorized to convert information in an analytical manner, generating an expected outcome from given input. Associators, such as the experience of disequilibrium, interiorization, formation of mental representations, and imaginative hypothetical thinking, are theorized to convert

Table 8.3. Continuum of Assumptions and parameters of Dynamic Models of Assessment

Adapted theoretical foundation	>>>	Comprehensive theoretical foundation
Single cognitive skill mediated	>>>	Mediation of multiple cognitive functions, personality and motivational factors
Standardized administration	>>>	Nonstandard administration guided by attempts to modify cognitive structure
Assumption of far transfer	>>>	Mediation for far transfer
Psychometric measurement	>>>	Measurement of process of change
Task-related changes sought	>>>	Structural change sought
Upper limits presumed	>>>	Upper limits not presumed

Note: Carlson & Wiedl >>>Campione & Brown >>>Budoff>>>Feuerstein>>>Jensen

information by an associative process that may generate a variety of potential outcomes. Together, the functions and associators form cognitive structures, contributing intellective capacity to human functioning. Motivational factors, such as a need for mastery, a desire for novelty, and the presence of aspirations, determine the inclination of the individual to engage in mental acts and, in turn, support those mental acts. Personality attributes such as self-confidence, frustration tolerance, and optimism, are seen in MET to determine aspects of the manner and style of the individual's cognitive and knowledge-structure development (Jensen, 1992). The variables in this model have been found to be sensitive to change, and thus able to contribute to enhancement of functioning. Approximately half were carefully investigated by Feuerstein and his colleagues (Feuerstein, Haywood, Rand, Hoffman, & Jensen, 1985; Jensen & Feuerstein, 1987). The additional variables are currently under investigation at Delphi Health & Science (Jensen, 1991).

Stability versus Change Models

The basic underlying assumptions and characteristics of dynamic versus static assessments are framed in the paradigms of stability versus change. Table 8.3 presents a synopsis of those differences

Stability models (e.g., measurement of I.Q.) presume stable individual differences that can be measured, yielding a valid indication of current functioning, as well as a prediction of future performance. Thus, stability models often limit the expectations set out for the individual measured, and are consistent with such practices such as labeling, classifying, and placement in situations such as classrooms, in which the environment is often modified to accommodate those expectations (e.g., simplified curricula). Stability models are characterized by an orientation toward products (e.g., numerical predictors), and assume that naturally occurring individual differences in ability exist. Their influence has been particularly strong in the field of psychometrics, where the search for stable individual differences yielded methods for the identification and classification of individuals based on their performance on standardized and normed tests (Jensen, 1992). The efficacy of this paradigm has been called into question for some time across subfields of psychology in which empirical findings point to context as a critical variable in understanding human functioning (Basic Behavioral Science Task Force of the National Advisory Mental Health Council, 1996; Bowlby, 1960; Harlow & Harlow, 1966; Kaufman & Rosenblum, 1967; Rogoff & Chavajay, 1995; Sackett, 1967).

Change models, as applied to human learning and modifiability, operate from the assumption that context is a critical variable that must be applied to the evaluation of human functioning. Although these models vary in their attention to such factors as age, etiology, and severity of impairment, they share the assumptions that the human organism is an open system and, therefore, that assessment should target the possibilities for cognitive and motivational enhancement (Jensen, 1992). They posit that human nature is cultural, and that learning involves the processing of contextually (e.g., culturally) meaningful symbols. Further, they propose that learning is a dynamic and open process in which active modification can be applied to the enhancement of functioning (Jensen, 1992; Robinson-Zañartu & Cook-Morales, 1992).

Dynamic Attention to Process versus Static Attention to Product

Dynamic assessment represents a significant departure from the static product-oriented model in which the only process involved is the posing of questions or situations, and recording of responses by the examiner, who remains deliberately separate from the process. In dynamic assessment the focus of attention is on the examinee's learning process, which is evaluated under conditions of learning. That is, the examiner not only observes but intervenes in the assessment process based on those observations, attempting to modify cognitive or learning approaches and then observe the results of the interventions. The targets of interventions are drawn from a repertoire of cognitive functions and motivational factors.

Dynamic assessment as outlined in both Jensen's and Feuerstein's models holds that learning begins with the primary caregiver, whose *context* is their primary culture and language. Thus, that caregiver is the first mediator of learning: he or she introduces the child to the elements of the environment with intentionality, establishes a reciprocal relationship, helps frame and shape behaviors associated with learning, and gives meaning and motivation to the interactions and emergent learning. These behaviors comprise essential characteristics of mediation. These behavioral features of mediation, in which the examiner engenders intentionality-reciprocity, the feeling of competence, regulation of behavior or cognitive pacing, transcendent value to the immediate experience, and a sense of meaning of change as it occurs, characterize the nature of the examiner-examinee interaction in a dynamic assessment

Issues of Measurement

Issues of measurement, validity, and reliability within change models are significant concerns of theoreticians operating from within as well as outside of the dynamic assessment paradigm. Although outcomes on certain dimensions of acquisition or even proceduralization may be measured against outcomes of alternative methods, using a variety of criteria (e.g., curriculum based measurement; authentic/portfolio assessment), the necessary level of individualization of each dynamic assessment precludes the use of some conventional measures of validity and reliability. The shift in basic assumptions will require a shift from the psychometric approach to mathematical constructs that are designed to measure change rather than stability. The focus is not to compare products of one individual with others, but to measure processes of change within that individual. Thus, they must depend on understanding those changes: what is presumed to undergo change, how that is accomplished, and when they are said to have occurred.

Jensen (1992) has proposed and is currently researching a mathematical measure of performance efficiency which utilizes computer-assisted touch-screen technology. This technology is able to record learner responses instantly and continually, and produces a within-subject measure of level and change of efficiency. The learning curve representing the changing value of efficiency provides data from which reliability and validity of those functional changes can be ascertained. He suggests that although individual and group statistics could be computed for this measure, its primary significance lies in its clinical application to the mediation process while the effort is being made to develop new modes of functioning.

Transfer of Effects

Transfer of the effects of dynamic assessment and mediated learning intervention have been discussed in some detail above. A substantial body of research supports the assumption that this model leads to enhanced cognitive functioning and problem-solving skills within training contexts and on tasks similar to those used in training. Some evidence also suggests that teachers perceive their students as better problem-solvers following mediated learning interventions (Greenberg, 1990; Kriegler & Kaplan, 1990). However, far transfer has yet to be established. Goals of education are currently being re-examined for their relevance to student needs in the 21st century. In the United States, mandates for educational reform now stress that the goals of education should reach beyond academic achievement. Significant attention is being directed toward enhanced thinking skills (Carnegie Council on Adolescent Development, 1989; National Educational Goals Report, 1991) as well as producing outcomes that *enable* the learner to profit from instruction (Ysseldyke & Thurlow, 1992).

While research with these models demonstrates enhanced problem-solving skills in near transfer, it does not yet support the assumption of transfer to broad academic enhancement, although a number of studies now indicate that when bridged to specific content, enhanced cognitive skills can enhance content acquisition (Greenberg, 1990; Kriegler & Kaplan, 1990; Salema & Valente, 1990). Promising preliminary studies and new theoretical directions may change these conclusions about far transfer in the future. Greenberg's (1990) COGNET model, for instance, has coupled mediated and cooperative learning strategies to produce demonstrable academic gains across multiple-content areas. Because of his attention to the multiple-dimensions variables in the enhancement of aptitude, Jensen's theoretical work on proceduralization of knowledge seems particularly worthy of rigorous scientific examination.

Applications to Decisions

Three types of applications to decision-making emerge from dynamic assessment: identification of misclassified students, preventive or developmental teaching methodologies, and strategic child interventions. The first identifies students misclassified on traditional intelligence measures as low in ability. Students regarded as misclassified are those who perform at levels on dynamic measures comparable to students in regular classes.

The second and third applications to decision making rely on the articulation of cognitive functions and motivational factors that are believed to enhance learning. The second application infuses mediated learning into the teaching methodology of the regular classroom. Here, cognitive functions implicit in the content of the curriculum are identified, and the functions are mediated within the existing academic content as a part of the lesson plan.

The third application, mediated intervention, specifies individual goals for enhancement determined through dynamic assessment. These targets will be prerequisite functions or factors designed to prepare the child to grasp and use the classroom-based functions. For instance, if the teacher has targeted the cognitive function of *categorizing* in conjunction with learning grouping in math, the prerequisite function of *comparative behavior* may need to be targeted to prepare the child. A broad variety of other functions, such as *inhibition of impulsive responding, conservation of constancies, precision and accuracy in data gathering, or awareness of a problem* would be examined during the dynamic assessment, and appropriate goals and sequences established for the child.

Treatment Validity

Treatment validity for dynamic assessment and mediated learning intervention models is enmeshed with the issues of near and far transfer discussed above. Most models of dynamic assessment have claimed that their aims are to enhance general cognitive functioning or problem-solving skills (e.g., Feuerstein, Haywood, Rand, Hoffman, & Jensen, 1985; Greenberg, 1990; Jensen, 1990, 1992). This criterion has been demonstrated on near-transfer tasks such as non-verbal measures of intelligence and teacher reports of new student approaches to problem solving (Johnson, 1996; Kriegler & Kaplan, 1990). In addition, attempts to deliberately pair learning or thinking skills with academic content have produced evidence of positive outcomes in reading accuracy, writing skills, and teacher ratings of attention (Greenberg, 1990, 1994; Salema & Valente, 1990). However, far transfer of enhanced cognitive functioning to either general cognitive functioning or enhanced broad academic performance has not yet been adequately demonstrated. Researchers are currently involved in new directions that apply the cognitive-skills enhancement directly to curriculum. These directions may lead to evidence of these models becoming more useful in the enhancement of student achievement.

SUMMARY

Reflecting the trends of the past decade, aptitude is examined from a perspective that includes any measure designed to predict individual differences in later learning. Specifically, attention is directed to three models of cognitive- and processing-aptitudes measures presumed to improve diagnosis, placement, and/or treatment of children and youth with learning problems: (1) psycholinguistic (PL) and perceptual-motor (PM) models, (2) aptitude by treatment interaction (ATI) models, and (3) change models of dynamic assessment (DA) and mediated learning intervention. All three models share the assumptions that a repertoire of cognitive pro-

cesses underlie the learning process, that examination of intra- rather than inter-individual differences should yield important data in the design of training or instruction, and that enhanced cognitive functioning and academic achievement should be ultimate goals in interventions based on these models.

Significant differences between the models lead to differing conclusions about their usefulness. The ATI models presume that aptitudes are stable individual traits, relatively insensitive to intervention. PL and PM models presume that deficient functions can be identified and changed with interventions. Consistent with stability models, ATI, PL, and PM models rely on cognitive responses (or products) as the primary unit of observation, and conventional standardized assessment procedures. In contrast, DA models assume that cognitive functions are not only modifiable, but interrelated, so that their enhancement would lead to the enhancement of overall cognitive skills. Consistent with change models, they place the unit of observation on the thinking processes used to produce a variety of responses, rather than the products. Procedures in DA are dynamic and interactive, guided by the thinking processes of the student and, therefore, are necessarily nonstandard.

Each model differs on presumed transfer of effects, leading to differences in the decisions associated with the models. ATI models assume that stable strengths should be matched with treatment or teaching to that strength, thus producing enhanced functioning. The PM and PL models assume that deficient functions can be trained, thus leading to enhanced functioning. Decisions related to these stability models usually relate to the kind of instruction that should be used, or to classifying a student for placement based on the assumption of certain cognitive or processing deficits. Dynamic change models assume that transfer can be achieved through the intervention processes identified during and integral to the assessment. Decisions relate to teaching methodology aimed at enhancing overall functioning. Interventions are interactive behaviors (mediations) applied by teachers, parents, or specialists to the targeted cognitive functions in the context of the curriculum.

When the standard of treatment validity based on enhanced academic outcomes is applied, two models currently fall short of expectations; the third is insufficiently researched to draw firm conclusions. Although the reasons for negative treatment validity evidence may be weaknesses in the basic theory, the measure of aptitude, or the available interventions, ATI research shows very weak interactions at best. In the case of PM, interventions appear to have limited effects on PM processes. PL interventions appear to have some minimal effect on PL processes, but no demonstrated effect on achievement. DA interventions appear to have a modest effect on cognitive enhancement and problem solving that is restricted to similar tasks (such as other non-verbal problem-solving tasks). However, no compelling evidence of transfer to broad areas of achievement exists to date. Recent theoretical and research developments in this area appear promising, and should be followed. Experiments that deliberately pair learning skills and academic content may produce better evidence of transfer of cognitive training effects. With increased attention in education to the role of thinking skills, assessment and intervention approaches that can provide evidence of enhanced thinking and enhanced achievement may become increasingly important.

REFERENCES

Anastasi, A. (1997). *Psychological testing* (7th ed.). New York: MacMillan.

Arter, J. A., & Jenkins, J. R. (1977). Examining the benefits and prevalence of modality considerations in special education. *Journal of Special Education, 11*, 281–298.

Ayres, R. R., & Cooley, E. J. (1986). Sequential versus simultaneous processing on the K-ABC: Validity in predicting learning success. *Journal of Psychoeducational Assessment, 4*, 211–220.

Ayres, R. R., Cooley, E. J., & Severson, H. H. (1988). Educational translation of the Kaufman assessment battery for children: A construct validity study. *School Psychology Review, 17*, 113–124.

Babad, E., & Budoff, M. (1974). Sensitivity and validity of learning potential measurement in three levels of ability. *Journal of Educational Psychology, 66*, 439–447.

Basic Behavioral Science Task Force of the National Advisory Mental Health Council (1996). NAMHC Report: Basic behavioral science research for mental health: Sociocultural and environmental processes. *American Psychologist, 51*, 722–731.

Bowlby, J. A. (1960). Separation anxiety. *International Journal of Psychoanalysis, 41*, 89–113.

Brown, F. G. (1983). *Principles of educational and psychological testing* (3rd ed.). New York: Holt, Rinehart, and Winston.

Budoff, M. (1987a). The validity of the learning potential assessment. In C. S. Lidz (Ed.), *Dynamic assessment: An interactional approach to evaluating learning potential*. New York: The Guilford Press.

Budoff, M. (1987b). Measures for assessing learning potential. In C. S. Lidz (Ed.), *Dynamic assessment: An interactional approach to evaluating learning potential*. New York: The Guilford Press.

Budoff, M., & Friedman, M. (1964). "Learning potential" as an assessment approach to the adolescent mentally retarded. *Journal of Consulting Psychology, 28*, 434–439.

Bush, W. J., & Giles, M. T. (1977). *Aids to psycholinguistic teaching* (2nd ed.) Columbus, OH: Charles E. Merrill.

Campione, J. C., & Brown, A. L. (1987). Linking dynamic assessment with school achievement. In C. S. Lidz, (Ed.), *Dynamic assessment: An interactional approach to evaluating learning potential*. (pp. 82–115). New York: Guilford.

Carlson, J., & Wiedl, K. H. (1976). Applications of "testing-the-limits: Towards a differential testing approach employing the Ravens Coloured Matrices. *Trier Psychologische Berichte, 3*, 1–80.

Carlson, J., & Wiedl, K. H. (1978). Use of testing the limits procedures in the assessment of intellectual capacities in children with learning difficulties. American *Journal of Mental Deficiency, 2*,(6), 559–564.

Carlson, J., & Wiedl, K. H. (1979). Towards a differential testing approach: Testing-the-limits employing the Ravens Matrices. *Intelligence, 3*, 323–344.

Carlson, J., & Wiedl, K. H. (1980). Applications of a dynamic testing approach in intelligence assessment: Empirical results and theoretical formulations. *Zeitschrift fur Differentielle und Diagnostiche Psychologie, 1*,(4), 303–318.

Carnegie Council on Adolescent Development. Task Force on Education of Young Adolescent (1989). *Turning points: Preparing American youth for the 21st century: The report of the Task Force on Education of Young Adolescents*. Washington, DC: Carnegie Council on Adolescent Development.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. *American Psychologist, 12*, 671–684.

Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist, 30*, 116–127.

Cronbach, L. J. (1990). *Essentials of psychological testing* (5th ed.). New York: Harper & Row.

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods*. New York: Wiley (Halstead Press).

Das, J. P., Kirby, J. R., & Jarman, R. F. (1979). *Simultaneous and successful cognitive processes*. New York: Academic Press.

Dearborn, W. F. (1921). Intelligence and its measurement. *Journal of Educational Psychology, 12*, 210–212.

De Weerdt, E. H. (1927). A study of the improvability of fifth grade children in certain mental functions. *Journal of Educational Psychology, 18*, 547–557.

Feuerstein, R. (1970). A dynamic approach to causation, prevention and alleviation of retarded performance. In H. C. Haywood (Ed.), *Social-cultural aspects of mental retardation*. New York: Appleton-Century-Crofts.

Feuerstein, R., Haywood, H. C., Rand, Y., Hoffman, M. B., & Jensen, M.R. (1985). The learning potential assessment device: Manual. Jerusalem, Israel: Hadassah WIZO Canada Research Institute.

Feuerstein, R., Jensen, M. R., Hoffman, M. B., & Rand, Y. (1985). Instrumental enrichment. An intervention program for structural cognitive modifiability: Theory and practice. In J. W. Sega., S. F. Chilpman, & R. Glaser (Eds.), *Thinking and learning skills, Vol.1. Relating instruction to research* (pp. 43–82). Hillsdale, New Jersey: Erlbaum.

Feuerstein, R., Rand, Y., Hoffman, M, & Miller, R. (1980). *Instrumental enrichment*. Glenview, IL: Scott Foresman & Company.

Feuerstein, R., Rand, Y., Jensen, M. R., Kaniel, S., Tzuriel, D. (1987). Prerequisites for assessment of learning potential: The LPAD mode. In C. S. Lidz (Ed.), *Dynamic assessment, an interactional approach to evaluating learning potential*. New York.: Guilford Press.

Glass, G. V. (1983). Effectiveness of special education. *Policy Studies Review, 2*, 65–78.

Greenberg, K. H. (1990). Mediated learning in the classroom. International *Journal of Cognitive Education and Mediated Learning, 1*, 33–44.

Greenberg, K. H. (1992, August). *Research and mediated learning: Implications for program implementation*. Paper presented at the Mediated Learning in Health and Education: Forging a New Alliance Conference, Tampa, FL.

Greenberg, K. H. (1994, November). *University/school partnerships in research and service: The Cognitive Enrichment Network National Follow-through Educational Model*. Paper presented at the annual meeting of the Mid-South Educational Research Association, Nashville, TN.

Greenberg, K. H., Coleman L., & Rankin, W. (1993). The Cognitive Enrichment Network Program: Goodness of fit with gifted underachievers. *Roeper Review, 16*, 91–95.

Hammill, D., & Larsen, S. (1974). The effectiveness of psycholinguistic training. *Exceptional Children, 41*, 5–14.

Hammill, D., & Larsen, S. (1978). The effectiveness of psycholinguistic training: A reaffirmation of position. *Exceptional Children, 44*, 402–414.

Harlow, H. F., & Harlow, M. K. (1966). Learning to love. *American Scientist, 54*, 244–272.

Harth, R. (1982). The Feuerstein perspective on the modification of cognitive performance. Focus on *Exceptional Children, 15*, 11–22.

Hartlage. C., & Reynolds, C. R. (1981). Neuropsychological assessment and the individualization of instruction. In G. W. Hynd & J. E. Obrzut (Eds.), *Neuropsychological assessment and the school age child: Issues and procedures*. New York: Grune & Stratton.

Haywood, H. C., Towery-Woolsey, J., Arbitman-Smith, R., & Aldrudge, A. (1988). Cognitive education with deaf adolescents: Effects of Instrumental Enrichment. *Topics in Language Disorders, 8*, 23–40.

Jensen, M. R. (1990). Change models and some evidence for phases and their plasticity in cognitive structures. *International Journal of Cognitive Educative & Mediated Learning, 1*(1), 5–16.

Jensen, M. R. (October, 1991). Keynote address presented at the California Association for Mediated Learning, Thousand Oaks, CA.

Jensen, M. R. (1992). Principles of change models in school psychology and education. In J. Carlson (Ed.), *Advances in cognition and educational practice, Vol. 1*, 47–72. Greenwich, CT: JAI Press.

Jensen, M. R., & Feuerstein, R. (1987). The learning potential assessment device: From philosophy to practice. In C. S. Lidz (Ed.), *Dynamic assessment: An interactional approach to evaluating learning potential* (pp. 379–402). New York: Guilford.

Jensen, M. R., Robinson-Zañartu, C., & Jensen, M. L. (1992). *Dynamic assessment and mediated learning. Assessment and intervention for developing cognitive and knowledge structures: An alternative in the era of reform*. Monograph for the Advisory Committee on the Reform of California's Assessment Procedures in Special Education. Sacramento: California Department of Education.

Jensen, M. R., & Singer, J. L. (1987). *Structural cognitive modifiability in low functioning adolescents: An evaluation of Instrumental Enrichment*. Report to the State of Connecticut Department of Education, Bureau of Special Education and Pupil Personnel Services, Hartford, CT.

Johnson, R. (1996, October). *Intervening with I.E. and raising IQ scores*. Paper presented at the California Association for Mediated Learning, San Diego, CA.

Kaufman, A. S., Goldsmith, B. Z., & Kaufman, N. L. (1984). *K-SOS: Kaufman sequential or simultaneous*. Circle Pines, MN: American Guidance Service.

Kaufman, A. S., & Kaufman, N. (1983). *Kaufman Assessment Battery for Children* (K-ABC). Circle Pines, MN: American Guidance Service.

Kaufman, I. C., & Rosenblum, L. A. (1967). The reaction to separation in infant monkeys: Anaclitic depression and conservation-withdrawal. *Psychosomatic Medicine, 29*, 648–675.

Kavale, K. A. (1981). Functions of the Illinois Test of Psycholinguistic Abilities (ITPA): Are they trainable? *Exceptional Children, 47*, 496–510.

Kavale, K. A. (1990). The effectiveness of special education. (pp. 868–898). In T. B. Gutkin & C. R. Reynolds (Eds.), (pp. 868–898). *The handbook of school psychology* (2nd ed). New York: Wiley.

Kavale, K. A., & Forness, S. R. (1987). Substance over style: Assessing the efficacy of modality testing and teaching. *Exceptional Children, 54*, 228–239.

Kavale, K. A., & Forness, S. R. (1990). Substance over style: A rejoinder to Dunn's animadversions. *Exceptional Children, 56*, 357–361.

Kavale, K. A., & Mattson, P. D. (1983). "One jumped off the balance beam:" Meta-analysis of perceptual-motor training. *Journal of Learning Disabilities, 16*, 165–173.

Kirk, S. A., & Kirk, W. (1971). *Psycholinguistic learning disabilities: Diagnosis and remediation*. Champaign, IL: University of Illinois Press.

Kirk, S. A., McCarthy, J., & Kirk, W. (1968). *Illinois Test of Psycholinguistic Abilities*. Champaign, IL: University of Illinois Press.

Klauer, K. J. (1989). Teaching for analogical transfer as a means of improving problem-solving, thinking and learning. *Instructional Science*, 179–192.

Kreigler, S. M., & Kaplan, M. E. (1990). Improving inattention and reading in inattentive children through MLE: A Pilot study. *International Journal of Cognitive Education & Mediated Learning, 1*(3), 185–192.

Lidz, C. S. (1987). *Dynamic assessment: An interactional approach to evaluation learning potential*. New York: Guilford Press.

Lidz, C. S., & Peña, L. (1996). Dynamic assessment: The model, its relevance as a non biased approach, and its application to Latino American preschool children. *Language, Speech and Hearing Services in Schools, 27*, 367–372.

Lund, K., Foster, G., & McCall-Perez, F. (1978). The effectiveness of psycholinguistic training: A reevaluation. *Exceptional Children, 44*, 310–321.

Luria, A. R. (1966a). *Higher cortical functions in man*. New York: Basic Books.

Luria, A. R. (1966b). *Human brain and psychological processes.* New York: Harper and Row.

Macmann, G. M., & Barnett, D. W. (1994a). Structural analysis of correlated factors: Lessons from the verbal-performance dichotomy of the Wechsler scales. *School Psychology Quarterly, 9,* 161–167.

Macmann, G. M., & Barnett, D. W. (1994b). Some additional lessons from the verbal-performance dichotomy of the Wechsler scales: A rejoinder to Kaufman and Keith. *School Psychology Quarterly, 9,* 223–226.

McDermott, P. A., Fantuzzo, J. W., & Glutting, J. J. (1992). Just say no to subtest analysis: A critique on Wechsler theory and practice. *Journal of Psychoeducational Assessment, 8,* 289–302.

McDermott, P. A., Fantuzzo, J. W., Glutting, J. J., Watkins, M. W., & Baggaley, A. R. (1992). Illusions of meaning in the ipsative assessment of children's ability. *Journal of Special Education, 25,* 504–526.

Mercer, C. D., King-Sears, P., & Mercer, A. R. (1990). Learning disabilities definitions and criteria used by state education departments. *Learning Disability Quarterly, 13,* 141–152.

Minskoff, E. H. (1975). Research on psycholinguistic training: Critique and guidelines. *Exceptional Children, 42,* 136–144.

Minskoff, E. H., Wiseman, D. E., & Minskoff, J. G. (1972). *The MWM program for developing language abilities.* Ridgeway, NJ: Educational Performance Associates.

Naglieri, J. A. (1989). A cognitive processing theory for the measurement of intelligence. *Educational Psychologist, 24,* 185–206.

Naglieri, J. A., & Das, J. P. (1997). *Cognitive Assessment System.* Itasca, IL: Riverside.

National Education Goals Report (1991): *Building a nation of learners.* Washington, DC: National Education Goals Panel.

Newcomer, R., Larsen, S., & Hammill, D. (1975). A response to Minskoff. *Exceptional Children, 42,* 144–148.

Osgood, C. E. (1957). Motivational dynamics of language behavior. In M. R. Jones (Ed.), *Nebraska Symposium on Motivation.* Lincoln, NE: University of Nebraska Press.

Penrose, L. S. (1934). *Mental defect.* New York: Farrar and Rinehart.

Perkins, D. N. (1987). Thinking frames: An integrative perspective on teaching cognitive skills. In J. B. Baron & R. J. Sternberg (Eds.), *Teaching thinking skills: Theory and practice* (pp. 41–61). New York: Freedman & Co.

Piaget, J. (1952). *The origins of intelligence in children.* (M. Cook, Trans.), New York: International Universities Press. (Original work published 1936)

Reschly, D. J., & Gresham, F. M. (1989). Current neuropsychological diagnosis of learning problems: A leap of faith. In C. R. Reynolds & E. Fletcher-Janzen (Eds.), *Handbook of Clinical Neuropsychology* (pp. 503–519). New York: Plenum Press.

Rey, A. (1934). D'un procédé pour évaluer l'éducabilité (quelques applications en psychopathologie). (A method for assessing educability and applications to psychopathology). *Archives de Psychologie, 24,* 297–337.

Reynolds, C. R. (1981). Neuropsychological assessment and the habilitation learning: Considerations in the search for aptitude x treatment interaction. *School Psychology Review, 10,* 343–349.

Reynolds, C. R. (1986). Transactional models of intellectual development, yes. Deficit models of process remediation, no. School Psychology Review, 15, 256–260.

Reynolds, C. R. (1992). Two key concepts in the diagnosis of learning disabilities and the habilitation of learning. *Learning Disability Quarterly, 15,* 2–12.

Robinson-Zañartu, C., & Cook-Morales, V. J. (1992). American Indian issues in school psychology. Working paper. San Diego, CA: San Diego State University.

Rogoff, B., & Chavajay, P. (1995). What's become of research on the cultural basis of cognitive development? *American Psychologist, 50,* 859–876.

Sackett, G. P. (1967). Some persistent effects of different rearing conditions on preadult social behavior of monkeys. *Journal of Comparative Physiological Psychology, 64,* 363-365.

Salema, M. H., & Valente, M. O. (1990). Learning to think: Metacognition in written composition. *International Journal of Cognitive Education and Mediated Learning, 1*(2), 161–170.

Salvia, J., & Ysseldyke, J. (1988). *Assessment in special and remedial education* (4th ed.). Boston: Houghton-Mifflin Company.

Schwartz, G. (1977). Psychosomatic disorders and biofeedback: A psychobiological model of disregulation. In J. A. Maser and M. E. P. Seligman (Eds.), *Psychopathology: Experimental Models* (pp. 271–307). San Francisco: W. H. Freeman.

Schwartz, G. (1983). Disregulation theory and disease: Applications to the repression/cerebral disconnection/cardiovascular disorder hypothesis. *International Review of Applied Psychology, 32,* 95–118.

Snow, R. E. (1980). Aptitude and achievement. In W. B. Schrader (Ed.), *Measuring achievement: Progress over a decade. New directions for testing and measurement.* San Francisco: Jossey-Bass.

Snow, R. E. (1992). Aptitude theory: Yesterday, today, and tomorrow. *Educational Psychologist, 27,* 5–32.

Teeter, P. A. (1987). Review of neuropsychological assessment and intervention with children and adolescents. *School Psychology Review, 16*, 582–583.

Teeter, P. A. (1989). Neuropsychological approaches to the remediation of educational deficits. In C. R. Reynolds & E. Fletcher-Janzen (Eds.), *Handbook of Clinical Child Neuropsychology* (pp. 357–376). New York: Plenum Press.

Thickpenny, J. P., & Howie, D. R. (1990). Teaching thinking skills to deaf adolescents: The implementation and education of instrumental enrichment. *International Journal of Cognitive Education and Mediated Learning, 1*(3), 193–209.

Vygotsky, L. S. (1962). *Thought and language.* (E. Hanfman & G. Vakar, Trans.), Cambridge, MA: MIT Press. (Original work published 1934)

Vygotsky, L. S. (1978). *Mind in Society: The development of higher psychological processes.* (M. Cole, V. John-Steiner, S. Scribner, & E. Souberman, Eds. and Trans.). Cambridge, MA: Harvard University Press. (Original work published 1935)

Wechsler, D. (1974). *Manual for the Wechsler intelligence scale for children-revised.* San Antonio, TX: Psychological Corporation.

Wechsler, D. (1991). *Wechsler intelligence scale for children-third edition: Manual.* San Antonio, TX: Psychological Corporation.

Woodcock, R. W., & Mather, N. (1989). WJ-R tests of cognitive ability-standard and supplemental batteries: Examiner's Manual. In R. W. Woodcock & M. B. Johnson, (Eds.), *Woodcock-Johnson, psychoeducational battery-revised.* Allen, TX: DLM Teaching Resources.

Ysseldyke, J. E., & Thurlow, M. L. (1992). Educational outcomes: Do we consider all students? *Communiqué, 20*(7), 20–21.

SUGGESTED READINGS

Jensen, M. R. (1992). Principles of change models in school psychology and education. In J. Carlson (Ed.), *Advances in cognition and educational practice,* Vol. 1 (pp. 47-72).

Kavale, K., & Forness, S. R. (1999). The effectiveness of special education. In T. B. Gutkin & C. R. Reynolds (Eds.), *The handbook of school psychology* (3rd Ed.). (pp. 984–1024). New York: Wiley.

Lidz, C. S. (1987). *Dynamic assessment: An interactional approach to evaluation learning potential.* New York: Guilford Press.

Reschly, D. J., & Gresham, F. M. (1989). Current neuropsychological diagnosis of learning problems: A leap of faith. In C. R. Reynolds & E. Fletcher-Janzen (Eds.), *Handbook of Clinical Neuropsychology* (pp. 503–519). New York: Plenum Press.

Snow, R. E. (1992). Aptitude theory: Yesterday, today, and tomorrow. *Educational Psychologist, 27,* 5–32.

Teeter, P. A. (1989). Neuropsychological approaches to the remediation of educational deficits. In C. R. Reynolds & E. Fletcher-Janzen (Eds.), *Handbook of Clinical Child Neuropsychology* (pp. 357–376). New York: Plenum Press.

This Page Intentionally Left Blank

CHAPTER 9

INTEREST INVENTORIES

Jo-Ida C. Hansen, Ph.D.

INTRODUCTION

The study of interests and the development of interest inventories emerged from applied psychology. The importance of an individual's interests in job selection was first recognized by educators in the 1900s and shortly thereafter by industry. Early theorists in the field, such as Parsons (1909), hypothesized that occupational adjustment was enhanced if an individual's characteristics and interests matched the requirements of the occupation. As E. K. Strong, Jr. (1943) pointed out in *Vocational Interests of Men and Women*, interests provide additional information, not available from analyses of abilities or aptitudes, for making career decisions. Consideration of interests, along with abilities, values, and personality characteristics, provides a thorough evaluation of an individual that is superior to consideration of any trait in isolation.

The earliest method for assessing interests was *estimation*, accomplished by asking individuals to indicate how they felt about various activities. To improve on the accuracy of their estimation, people were encouraged to *try out* activities before making their estimates. However, try-out techniques for evaluating interests were time consuming and costly; the search for a more economical assessment method led to development of interest *checklists* and *rating scales* (Kitson, 1925; Miner, 1922) and eventually to *interest inventories* that used statistical procedures to summarize an individual's responses to a series of items representing various activities and occupations.

The Earliest Item Pool

The first item pool of interest activities was accumulated in a seminar taught by Clarence S. Yoakum at Carnegie Institute of Technology in 1919. The 1,000-item pool was developed using a *rational sampling approach* designed to represent the entire domain of interests. Over the years, statistical analyses were performed to determine the worth of each item, and numerous test authors used that original item pool as the foundation for development of their inventories (e.g., Occupational Interest Inventory [Freyd, 1922-1923], Interest Report Blank [Cowdery, 1926]; General Interest Survey [Kornhauser, 1927]; Vocational Interest Blank [Strong, 1927]; Purdue Interest Report [Remmers, 1929]; Interest Analysis Blank [Hubbard, 1930]; Minnesota Interest Inventory [Paterson, Elliott, Anderson, Toops, & Heidbreder, 1930]).

Characteristics of Good Items

Interest inventory items should be evaluated periodically because societal changes can make items obsolete as well as create the need for new items. Several qualities that contribute to the excellence of items, and ultimately to the excellence of an interest inventory, can be used to assess the value of each item.

First, items should differentiate among groups because the purpose of interest inventories is to distinguish people with similar interests from those with dissimilar interests. The item in Figure 9.1,

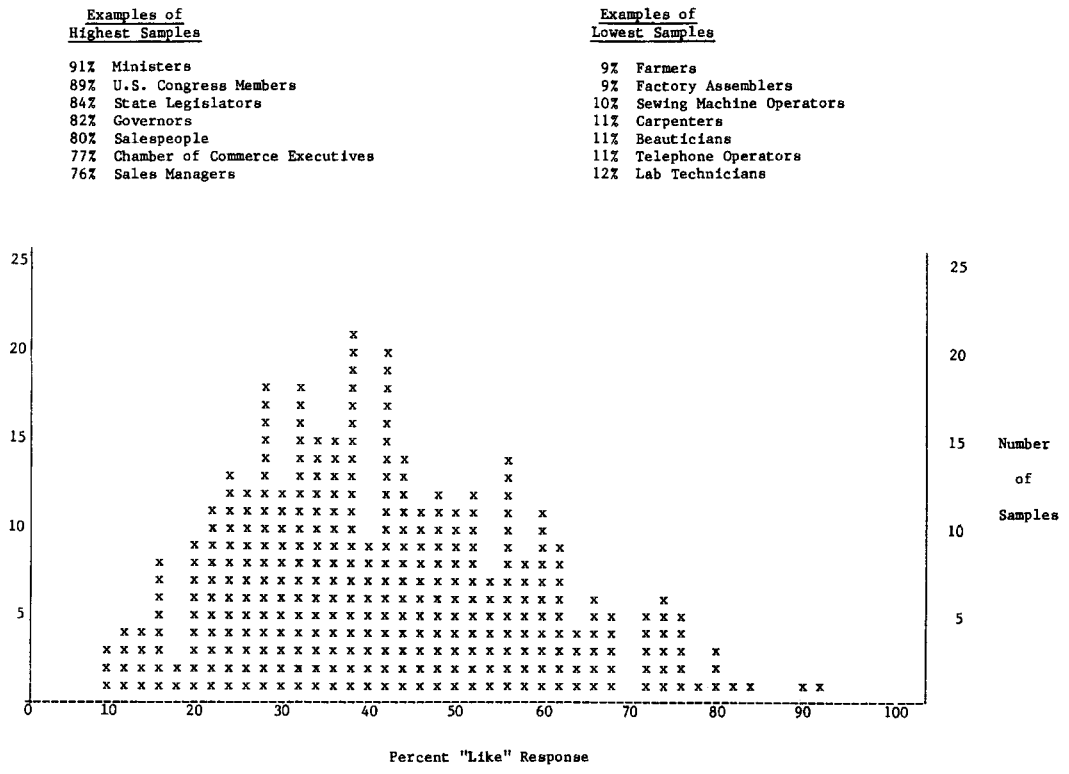


Figure 9.1. Percent "Like" responses to the item *Making a speech* for 350 occupational samples.

for example, has the power to spread 350 occupations over a wide range of response percentages. The lowest "Like" response rate for this item, *Making a Speech*, is 9 percent (meaning that few people in the sample answered "Like" to the item), and response rates range up to a high of 91 percent (meaning that the majority of the sample responded "Like").

Samples or groups with similar interests should have similar item-response rates, and clusters of groups with high or low response rates should make sense. In Figure 9.1, for example, the samples of ministers, members of the U.S. Congress, state legislators, governors, salespeople, and Chamber of Commerce executives had high "Like" response rates of 77-91 percent. Farmers, factory assemblers, sewing machine operators, carpenters, beauticians, and laboratory technicians, however, had low "Like" response rates to the same item. Those clusters of high and low response-rate samples are intuitively satisfying and illustrate the item's content validity; one expects ministers and politicians, for example, to enjoy making a speech.

Items also should be sex-fair; no item should suggest that any occupation or activity is more appropriate for one sex than the other. In addition to sex-fair items, all interpretive and instructional materials for interest inventories should be sex-fair.

To facilitate adaptation of inventories for use with ethnic minorities or for international use, interest items should be unambiguous and culture-fair. Straightforward items also are more likely to have the same meaning for everyone taking the inventory regardless of cultural or occupational orientation, and they will be easier to translate into several languages.

All items should be revised periodically to ensure that they are current and familiar to the respondents. The face validity, as well as content validity, of an interest inventory is affected if the item pool contains obsolete items that are unfamiliar to the general population. On the other hand, as

new technologies develop, new items should be generated to ensure that the entire domain of interests is represented in the item pool.

Finally, items should be easy to read. All materials that accompany interest inventories (e.g., instructions, profile, interpretive information) and the item pool itself should be easy to read to make the inventory useful for a wide educational and age- range in the population.

Theories of Vocational Interests

The earliest interest inventories were developed using the atheoretical, *empirical method of contrast groups* that is based on an assumption that people with similar interests can be clustered together and, at the same time, be differentiated from groups with dissimilar interests. Inventories that still incorporate this method of scale construction are the Strong Interest Inventory[1] (Harmon, Hansen, Borgen & Hammer, 1994), the Career Assessment Inventory™[2] (Johansson, 1975; 1986) and the Campbell Interest and Skill Survey™ (CISS®[3]; Campbell, Hyne, & Nilsen, 1992).

Results from the early empirical investigations of interests later were used to develop hypotheses about the structure of interests. Anne Roe (1956) and John Holland (1959), for example, used the factor analysis of Guilford and his colleagues (Guilford, Christensen, Bond, & Sutton, 1954), who found seven interest factors: (a) mechanical, (b) scientific, (c) social welfare, (d) aesthetic expression, (e) clerical, (f) business, and (g) outdoor work, to guide the development of their theories about interests.

CONSTRUCTION OF INTEREST INVENTORY SCALES

Construction of interest inventories is based on several assumptions:

1. First, a person can give informed responses of degree of interest (e.g., like, indifferent, dislike) to familiar activities and occupations.
2. Then, unfamiliar activities have the same factor structure as do familiar activities.
3. Therefore, familiar activities and occupations can be used as items in interest inventories to identify unfamiliar occupational interests.

Early interest inventories typically featured either *homogeneous* or *heterogeneous* scales. Now, however, many inventories—Campbell Interest and Skill Survey™, the Career Assessment Inventory™, the Kuder Occupational Interest Survey (KOIS) (Form DD) (Kuder, 1966), and the Strong Interest Inventory—combine homogeneous and heterogeneous scales. Generally, heterogeneous scales are more valid for predictive uses of interest inventories (e.g., predicting future job entry or college major), but homogeneous scales are more useful for providing parsimonious descriptions of the structure of a sample's interests (Edwards & Whitey, 1972).

Homogeneous Scale Development

One method of scale construction involves clustering together items based on internal consistency or homogeneous scaling. Items chosen in this manner have high intercorrelations. Empirical methods, such as cluster or factor analyses, can be used to identify the related items. The scales of the Vocational Interest Inventory (VII) (Lunneborg, 1976), for example, were constructed using factor analysis. The scales also may be based on rational selection of items; this method uses a theory to determine items appropriate for measuring the construct represented by each scale. For example, the General Occupational Themes of the Strong Interest Inventory were rationally constructed using Holland's theoretical definition of the six vocational types to guide item selection (Campbell & Holland, 1972; Hansen & Johansson, 1972).

Heterogeneous Scale Development

The Occupational Scales of the Campbell Interest and Skill Survey™, the Career Assessment Inventory™, Strong Interest Inventory™$_1$, and the Kuder Occupational Interest Survey (Form DD) are composed of items with low intercorrelations, and therefore, are called heterogeneous scales. Heterogeneous scales are atheoretical: in other words, the choice of items is based on empirical results rather than an underlying theory. The CISS®, the Career Assessment Inventory™, and the Strong Interest Inventory use™ the empirical method of contrast groups to select items; this technique compares the item-response rates of occupational criterion groups and contrast groups,

representing the interests of people in general, to identify items that significantly differentiate the two samples. The KOIS uses a different empirical method that compares an individual's item-response pattern directly to the item-response patterns of criterion samples that represent the interests of various occupations and college majors.

CURRENT INTEREST INVENTORIES

One of the most recently developed interest inventories is the Campbell Interest and Skill Survey™ (Campbell, 1995). Other widely used inventories include the Vocational Preference Inventory (Holland, 1985c), the Self-Directed Search (SDS) (Holland, 1971, 1987a, 1994), various forms of the Kuder, the Strong Interest Inventory™, the Career Assessment Inventory™, the Jackson Vocational Interest Survey (JVIS) (Jackson, 1977), the unisex version of American College Testing's Interest Inventory (UNIACT) (Lamb & Prediger, 1981; Swaney, 1995), and the Vocational Interest Inventory (VII) (Lunneborg, 1976).

Campbell Interest and Skill Survey

David Campbell, author of the Campbell Interest and Skill Survey™ (CISS®, describes the instrument as a product of 90 years of psychometric evolution influenced to a large extent by Campbell's work with the Strong Interest Inventory™ in the 1960s, 1970s and 1980s (Campbell, 1995). The CISS® is unique among interest inventories in that the instrument is designed to assess not only an individual's interest in academic and occupational topics but also an individual's estimation of her or his skill in a wide range of occupational activities. The profile includes 98 scales on which two scores are provided—an interest score and a skill score.

Item Pool and Profile

The item pool for the CISS® includes 200 interest items and 120 items designed to assess self-reported skills. The response format for the interest items is a six-point scale ranging from "Strongly Like" to "Strongly Dislike". The skill items also have a six-point response scale that includes self evaluations of Expert, Good, Slightly Above Average, Slightly Below Average, Poor, and None (have no skills in this area).

Scales

The CISS® profile includes three types of scales: seven Orientation Scales, 29 Basic Scales, and 60 Occupational Scales. The Orientation Scales capture the major interest factors that have been identified through various statistical clustering procedures and include Influencing (business and politics), Organizing (managing and attention to detail), Helping (service and teaching), Creating (the arts and design), Analyzing (science and math), Producing (hands-on and mechanical), and Adventuring (physical activities and competition). The Orientation Scales are used as the organizational frame of reference for the CISS® profile (see Figures 9.2 and 9.3).

The 29 Basic Scales were developed by clustering together homogeneous items in content areas such as Sales, Supervision, Adult Development, International Activities, Science, Woodworking, and Military/Law Enforcement. These scales are grouped on the profile under the Orientation with which they correlate most highly (see Figure 9.2).

The Occupational Scales were constructed using the empirical method of contrast groups originally refined for interest measurement by E. K. Strong, Jr. Successful, satisfied workers in each of 60 occupations were surveyed. Their responses to each of the CISS® items were compared to the item responses of a general reference sample composed of employed workers from a variety of occupations. Items that substantially differentiated the occupational criterion sample from the general reference sample were selected for the occupation's scale.

The first step to determine the location of the Occupational Scales on the profile, was to compute the mean score for each occupational criterion sample on the Orientation Scales. The occupation's highest Orientation score then was used to locate the Occupational Scale on the profile. For example, the test pilot and ski instructor criterion samples scored highest on the Adventuring Orientation, and therefore, the Occupational Scales representing their interests are clustered with the Adventuring Orientation on the profile (see Figure 9.3).

Two additional scales on the CISS® profile are Academic Focus and Extraversion. The Aca-

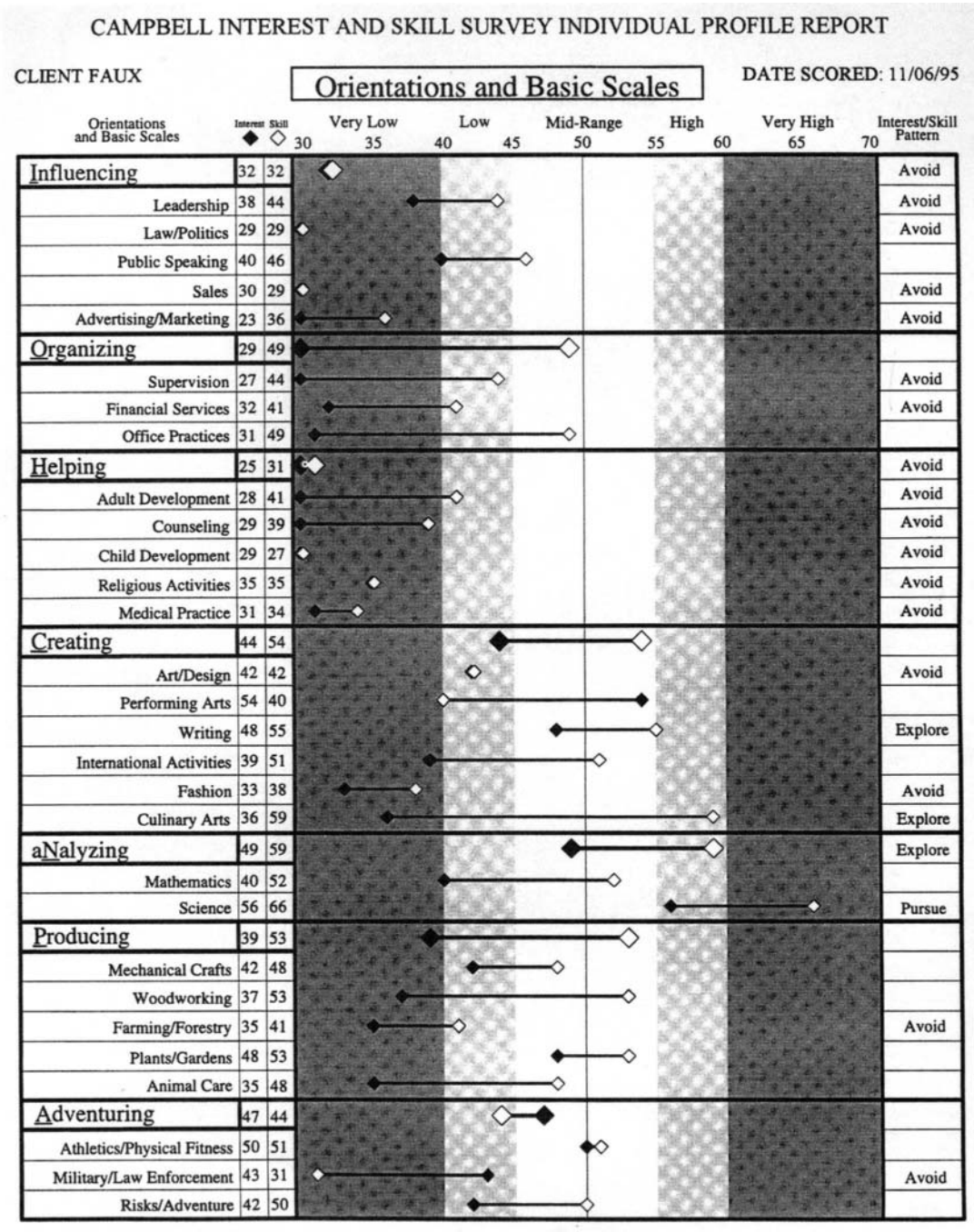


Figure 9.2. Profile for the Orientations and Basic Scales for the CISS(R).

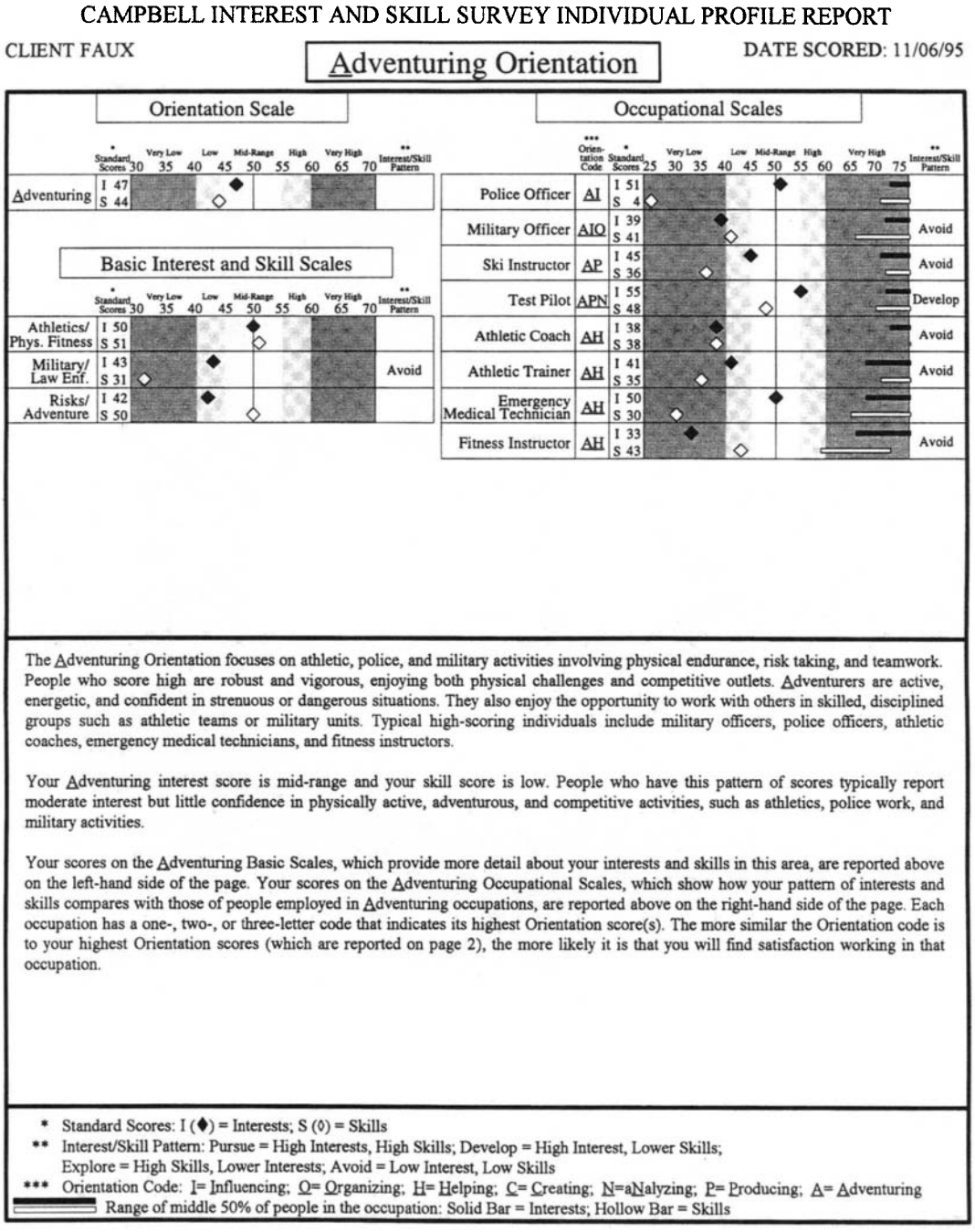


Figure 9.3. Profile for the Adventuring Orientation of the CISS.

demic Focus Scale measures interest and confidence in academic pursuits especially science and the arts. The Extraversion Scale measures interest and confidence in activities that require high levels of personal interaction.

Norming and Profile Report

All of the scales on the CISS(R) are normed on a general reference sample of women and men. The scales also are standardized with the result that the mean score for the general reference sample is about 50 and the standard deviation about 10. The sample used to norm the scales included 1,790 female and 3,435 male respondents from 65 occupational samples. The raw score means for the two samples were averaged to give the sexes equal weighting in the raw-score-to-standard-score conversion.

The CISS(R) Report is an 11-page document that includes one page that reports the Orientation and Basic Scale Interest and Skills scores as illustrated in Figure 9.2. An additional seven pages summarize scores for all of the scales related to each of the seven Orientations as illustrated in Figure 9.3 for the Adventuring Orientation. The additional three pages include one page for the special scales and procedural checks, and finally, a two page summary.

In addition to presenting an interest and a skill score for each scale, the profile also includes a graph that plots the interest and skill scores to provide interpretive comments ranging from very low to very high. An interpretive bar representing the middle 50 percent of scores for each criterion sample on its own Interest and Skill scales also is provided on the profile (solid bar = Interests; hollow bar = Skills) for the Occupational Scales. Finally, each of the seven Orientation pages includes a short interpretive report that summarizes the individual's results.

Measurement of both interests and confidence in skills enriches the interpretive information that can be gleaned from the CISS(R) scores. Based on a comparison of the level of the interest and skills scores for each scale, the individual is advised to *Pursue* the area if both the interest and skill scores are high, to *Develop* the area if the interest score is high but the skill score is low, to *Explore* the area if interest is low and skill high, or to *Avoid* if both interest and skill scores are low.

Validity and Reliability

Substantial evidence of the construct validity of the interest and skill scales is presented in the manual of the CISS(R) (Campbell, Hyne, & Nilsen, 1992). Test-retest correlations over a 90-day interval are .87, .83, and .87 for the Orientation, Basic, and Occupational interest scales, respectively, and .81, .79, and .79 for the Orientation, Basic, and Occupational skill scales.

Holland's Interest Inventories

Emergence of John Holland's theory of careers (Holland, 1959, 1966, 1992) began with the development of the Vocational Preference Inventory (VPI) (Holland, 1958). Based on interest data collected with the VPI as well as data from other interest, personality, and values inventories and from analyses of the structure of interests, Holland formulated his theory of vocational life and personality. According to Holland, people can be divided into six types or some combination of six types: Realistic, Investigative, Artistic, Social, Enterprising, and Conventional. Holland indicates that the types can be organized in the shape of a hexagon in the R-I-A-S-E-C order; the types adjacent to one another on the hexagon (e.g., Realistic-Investigative or Enterprising-Conventional) are more related than types that are diametrically opposed to one another (e.g., Realistic-Social or Artistic-Conventional). Attempts to verify Holland's hexagonal representation of the world of work show, in general, that the structure of interests approximates the theoretical organization proposed by Holland (Campbell & Hansen, 1981; Cole & Hanson, 1971; Hansen, Collins, Swanson, & Fouad, 1993; Haverkamp, Collins & Hansen, 1994; Prediger, 1982; Rounds, 1995).

Holland's theory has led to development of inventories and sets of scales to measure his six types, for example, his own Self-Directed Search, the ACT Interest Inventory, the Career Decision-Making System-Revised (CDM-R) (Harrington & O'Shea, 1993), the General Occupational Themes of the Strong Interest Inventory™ (Campbell & Holland, 1972;

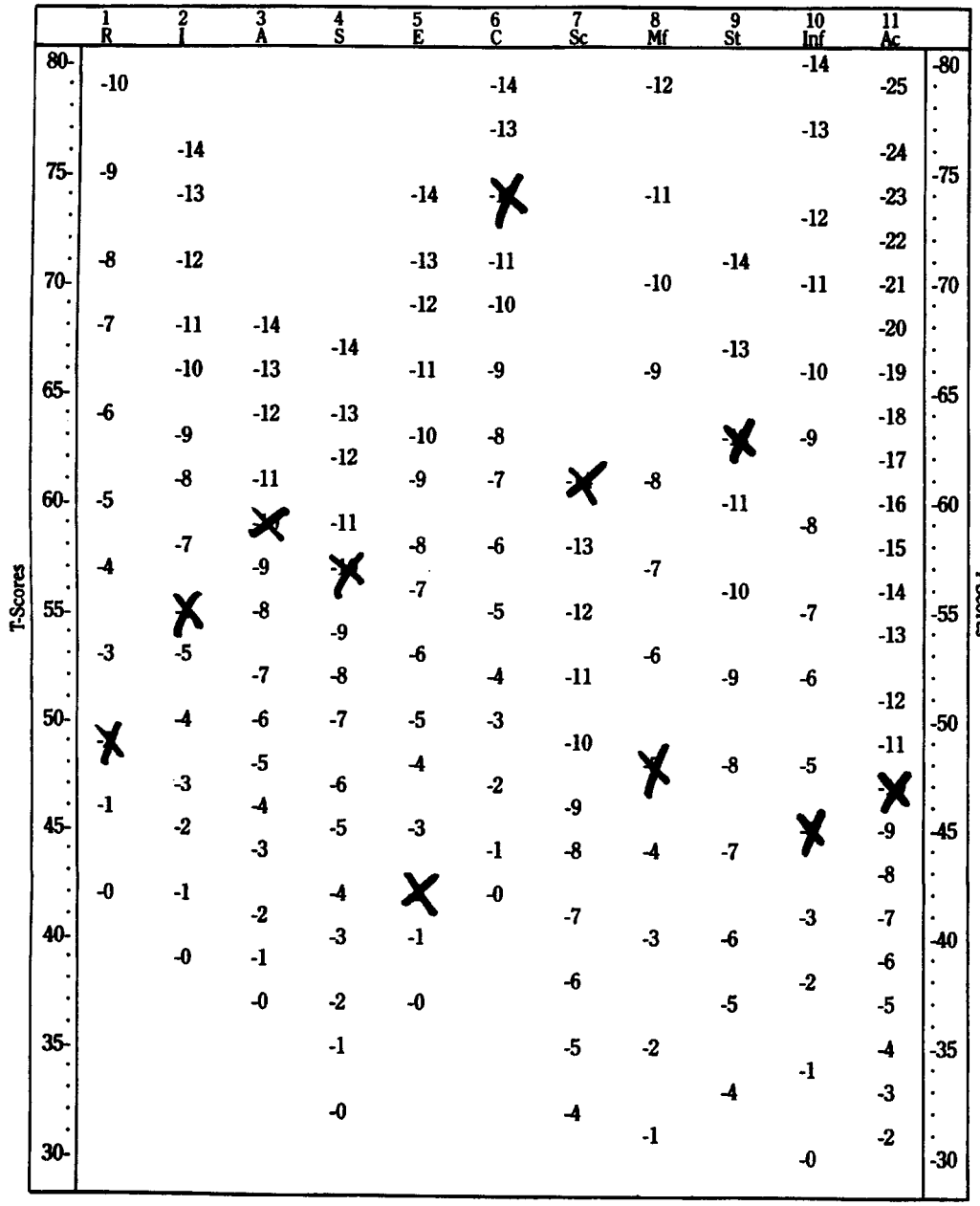


Figure 9.4. Profile for the Vocational Preference Inventory. Reproduced and adapted by special permission of the Publisher, Psychological Assessment Resources, Inc., Odessa, FL 33556, from the Vocational Preference Inventory by Dr. John L. Holland, Ph.D., Copyright 1978, 1985 by PAR, Inc. Further reproduction is prohibited without permission from PAR, Inc.

How To Organize Your Answers

Start on page 4. Count how many times you said L for "Like." Record the number of L's or Y's for each group of Activities, Competencies, or Occupations on the lines below.

	R	I	A	S	E	C
Activities (pp. 4-5)	4	6	5	6	9	0
Competencies (pp. 6-7)	9	8	1	10	5	6
Occupations (p. 8)	2	6	10	1	2	0
Self-Estimates (p. 9) (What number did) you circle?)	6	7	1	6	1	4
	6	6	4	4	4	4
Total Scores (Add the five R scores, the five I scores, the five A scores, etc.)	25	33	21	27	19	14

The letters with the three highest numbers indicate your summary code. Write your summary code below. (If two scores are the same or tied, put both letters in the same box.)

Summary Code

Highest	2nd	3rd
I	S	R

Figure 9.5. Summary code and scores for the SDS. Adapted and reproduced by special permission of the Publisher, Psychological Assessment Resources, Inc., Odessa, FL 33556, from the Self-Directed Search Assessment Booklet by John L. Holland, Ph.D. Copyright 1970, 1977, 1985, 1990, 1994 by PAR, Inc. Further reproduction is prohibited without permission from PAR, Inc.

Hansen & Johansson, 1972), and the General Themes of the Career Assessment Inventory™.

Vocational Preference Inventory

Development of the Vocational Preference Inventory (VPI) was based on a series of theoretical and empirical reports. Holland surveyed personality, vocational choice, and vocational interest literature; identified interest-personality factors; and hypothesized how they related to one another. Then, he used 160 occupational titles to develop an item pool that represented the interest factors or types.

The current version of the VPI (Holland, 1985c) has seven homogeneous scales, constructed in a series of rational-empirical steps that measure Self-Control (Sc) plus the six types hypothesized in Holland's theory: Realistic (R), Investigative (I), Artistic (A), Social (S), Enterprising (E), and Conventional (C). Other VPI scales developed using empirical methods of scale construction include: Acquiescence (Ac), measuring willingness to say "yes" to items; Status (St), indicating interest in occupational status; Masculinity-Femininity (Mf), measuring interest in occupations traditionally preferred by men or women; and Infrequency (Inf), assessing the tendency to answer items in an atypical direction.

The VPI may be hand scored; raw scores are plotted either on the female profile shown in Figure 9.4 or a male profile. Even though Holland is a strong proponent of the use of raw scores for predicting occupational membership, the profile is calibrated to provide standard scores based on either 378 female or 354 male college students and employed adults to provide comparisons across scales.

Self-Directed Search

The Self-Directed Search (SDS) (Holland, 1985b, 1987a, 1994), similar to the VPI, was developed to measure Holland's six types. It may be self-administered, self-scored, and to a limited degree, self-interpreted. The 228-item assessment booklet includes four sections: Activities the respondent would like to do; Competencies; Occupations; and Self-Estimates.

The reading level of the SDS is estimated to be at the seventh- or eighth-grade level; Form Easy (E), which has only 203 items, is rated at the fourth-grade level. As illustrated in Figure 9.5, the most important feature of the SDS profile is the summary codes. The three highest raw scores represent the respondent's primary, secondary, and tertiary code assignments. Holland (1979) suggests flexibility in using the three summary codes for occupational exploration, since the codes are approximate, not precise

A series of materials has been developed to assist in the interpretation of the SDS. The *1987 Manual Supplement* (Holland, 1987a) explains the use of the SDS in individual- and group-career assistance. *The Occupations Finder* (Holland, 1985a) and *The College Majors Finder* (Holland, 1987b) provide three-letter Holland codes for 1,156 occupations and more than 900 college majors, respectively. Occupational and educational alternatives can be identified by surveying the two booklets to find possibilities with summary codes which are similar to the individual's summary code.

Reliability and Validity. The median test-retest reliability coefficient for the 11 VPI scales over a two-week interval is .72; over the same period the median reliability coefficient for the six SDS scales is .82 for high school students and over 7 to 10 months, .92 for college students (Holland, 1978, 1979, 1985b, 1985c). Studies of the predictive validity of the VPI and SDS, for choice of occupation and college major over one-, two-, and three-year intervals, range from 35 percent to 66 percent accuracy (Holland, 1962, 1979, 1985c, 1987a; Holland & Lutz, 1968).

Strong Interest Inventory

The earliest version of the Strong Vocational Interest Blank®[4] (1927) used the empirical method of contrast groups to construct occupational scales representing the interests of men in 10 occupations. The first form for women was published in 1933, and until 1974 the instrument was published with separate forms for women and men. In 1974 (Campbell, 1974), the two forms were combined by selecting the 325 best items from the previous women's (TW398) and men's (T399) forms and in 1981 (Campbell & Hansen, 1981) another revision was completed in an effort to provide matched-sex Occupational Scales (e.g., male- and female-normed Forester Scales, male- and

female-normed Flight Attendant Scales, male- and female-normed Personnel Director Scales). The 1985 revision (Hansen & Campbell, 1985) marked the end of the sex-equalization process which began in 1971. One additional major change in the 1985 revision was the expansion of the breadth of the profile to include more nonprofessional and vocational/technical occupational scales. The most recent revision of the Strong was completed in 1994 (Harmon, Hansen, Borgen, & Hammer, 1994).

Item Pool and Profile

The item booklet for the 1994 revision of the Strong Interest Inventory includes 317 items, divided into eight sections: Part 1, Occupational Titles; Part 2, School Subjects; Part 3, Activities; Part 4, Leisure Activities; Part 5, Types of People; Part 6, Forced-choice Preference Between Two Activities; Part 7, Self-Description Characteristics, and Part 8, Preference in the World of Work. The item format requires respondents to indicate the degree of their interest in each item by responding "Like," "Indifferent," or "Dislike."

The profile includes four sets of scales: six General Occupational Themes, 25 Basic Interest Scales, 211 Occupational Scales that represent professional and nonprofessional occupations (e.g., farmers, geographers, photographers, social workers, buyers, credit managers), and four Personal Styles Scales.

Occupational Scales

The Occupational Scales of the Strong are another example of test construction using the empirical method of contrast groups. The response-rate percentage of the occupational criterion sample to each item is compared to the response-rate percentage of the appropriate-sex contrast sample (i.e., General Reference Sample of females or males) to identify items that differentiate the two samples. Usually 30 to 50 items are identified as the interests ("Likes") or the aversions ("Dislikes") of each occupational criterion sample. The raw scores for an individual scored on the Occupational Scales are converted to standard scores based on the occupational criterion sample, with mean set equal to 50 and standard deviation of 10 (see Figure 9.6).

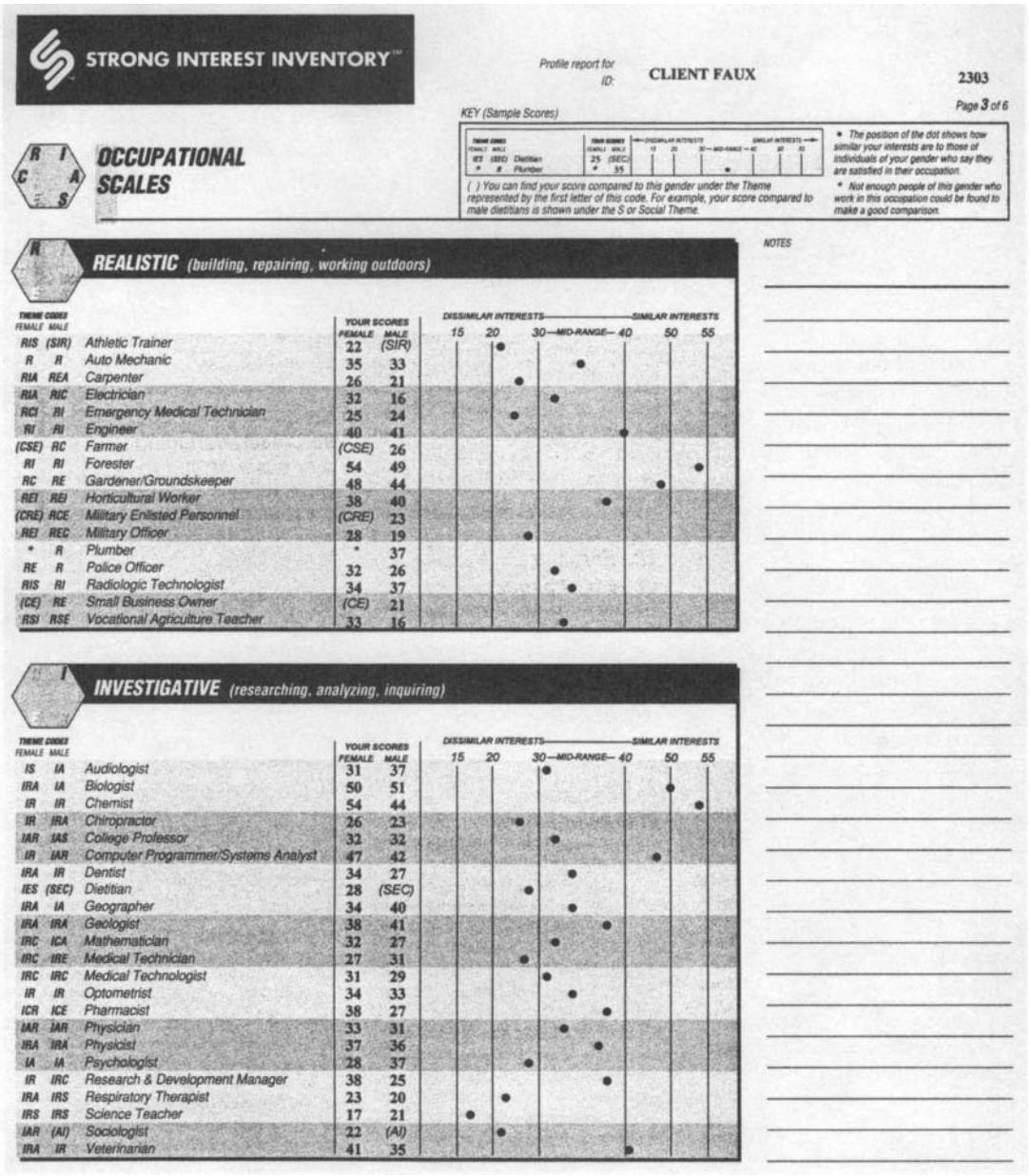
For most occupations, matched-sex scales are presented on the Strong Interest Inventory profile. However, seven of the 109 occupations (211 Scales) are represented by just one scale (e.g., f Child Care Provider, f Dental Assistant, f Dental Hygienist, f Home Economics Teacher, f Secretary, m Agribusiness Manager, and m Plumber).

General Occupational Themes

The General Occupational Themes (GOT) are a merger of Strong's empiricism with Holland's theory of vocational types. The six homogeneous Themes contain items selected to represent Holland's definition of each type—Realistic, Investigative, Artistic, Social, Enterprising, and Conventional. Data comparing the enhanced 1994 GOT to Holland's Vocational Preference Inventory or Self-Directed Search are not available. However, the 1985 GOT correlated highly (.72 to .79) with same-named Vocational Preference Inventory scales (Hansen, 1983). Correlations between the GOT indicate that the hexagonal order that Holland proposes to describe the relationship between his types (adjacent types have more in common than do diametrically opposed types) also describes the relationship between the Strong Interest Inventory Themes (Harmon et al., 1994).

Figure 9.7 illustrates the score information provided for the General Occupational Themes on the profile; the same information is presented for the Basic Interest Scales. The standard scores are based on a General Reference Sample composed of women and men with mean set equal to 50 and standard deviation of 10. In addition to standard scores, interpretive bars provide a visual representation of the distribution of the female General Reference Sample (upper bar) and male General Reference Sample (lower bar), respectively.

The integration of Holland's theory with Strong's empiricism provides the organizational framework for the current Strong profile. The Occupational Scales are coded with one to three Holland types based on the criterion sample's highest scores on the General Occupational Themes. The codes, in turn, are used to categorize the Occupational Scales on the profile (see Figure 9.6). The Basic Interest Scales (BIS) also are clustered according to Holland types by identifying the



STRONG INTEREST INVENTORY™

Profile report for ID: CLIENT FAUX 2303

Page 3 of 6

KEY (Sample Scores)

THEME CODES FEMALE	MALE		YOUR SCORES FEMALE	MALE
IES	(SEC)	Dietitian	25	(SEC)
*	R	Plumber	*	35

( ) You can find your score compared to this gender under the Theme represented by the first letter of this code. For example, your score compared to male dietitians is shown under the S or Social Theme.

* The position of the dot shows how similar your interests are to those of individuals of your gender who say they are satisfied in their occupation.

* Not enough people of this gender who work in this occupation could be found to make a good comparison.

OCCUPATIONAL SCALES

REALISTIC (building, repairing, working outdoors)

THEME CODES FEMALE	MALE		YOUR SCORES FEMALE	MALE
RIS	(SIR)	Athletic Trainer	22	(SIR)
R	R	Auto Mechanic	35	33
RIA	REA	Carpenter	26	21
RIA	RIC	Electrician	32	16
RCI	RI	Emergency Medical Technician	25	24
RI	RI	Engineer	40	41
(CSE)	RC	Farmer	(CSE)	26
RI	RI	Forester	54	49
RC	RE	Gardener/Groundskeeper	48	44
REI	REI	Horticultural Worker	38	40
(CRE)	RCE	Military Enlisted Personnel	(CRE)	23
REI	REC	Military Officer	28	19
*	R	Plumber	*	37
RE	R	Police Officer	32	26
RIS	RI	Radiologic Technologist	34	37
(CE)	RE	Small Business Owner	(CE)	21
RSI	RSE	Vocational Agriculture Teacher	33	16

DISSIMILAR INTERESTS — 15 20 30—MID-RANGE—40 50 55 — SIMILAR INTERESTS

INVESTIGATIVE (researching, analyzing, inquiring)

THEME CODES FEMALE	MALE		YOUR SCORES FEMALE	MALE
IS	IA	Audiologist	31	37
IRA	IA	Biologist	50	51
IR	IR	Chemist	54	44
IR	IRA	Chiropractor	26	23
IAR	IAS	College Professor	32	32
IR	IAR	Computer Programmer/Systems Analyst	47	42
IRA	IR	Dentist	34	27
IES	(SEC)	Dietitian	28	(SEC)
IRA	IA	Geographer	34	40
IRA	IRA	Geologist	38	41
IRC	ICA	Mathematician	32	27
IRC	IRE	Medical Technician	27	31
IRC	IRC	Medical Technologist	31	29
IR	IR	Optometrist	34	33
ICR	ICE	Pharmacist	38	27
IAR	IAR	Physician	33	31
IRA	IRA	Physicist	37	36
IA	IA	Psychologist	28	37
IR	IRC	Research & Development Manager	38	25
IRA	IRS	Respiratory Therapist	23	20
IRS	IRS	Science Teacher	17	21
IAR	(AI)	Sociologist	22	(AI)
IRA	IR	Veterinarian	41	35

DISSIMILAR INTERESTS — 15 20 30—MID-RANGE—40 50 55 — SIMILAR INTERESTS

NOTES

Figure 9.6. Profile for the General Occupational Themes and the Basic Interest Scales of the Strong. Modified and reproduced by special permission of the Publisher, Consulting Psychologists Press, Inc., Palo Alto, CA 94303 from the *Strong Interest Inventory*™ of the *Strong Vocational Interest Blanks*® Form T317. Copyright 1933, 1938, 1945, 1946, 1966, 1968, 1974, 1981,1985, 1994 by The Board of Trustees of the Leland Stanford Junior University. All rights reserved. Printed under license from Stanford University Press, Stanford, CA 94305. Further reproduction is prohibited without the Publisher's written consent. *Strong Interest Inventory* is a trademark and *Strong Vocational Interest Blanks* is a registered trademark of the Stanford University Press.

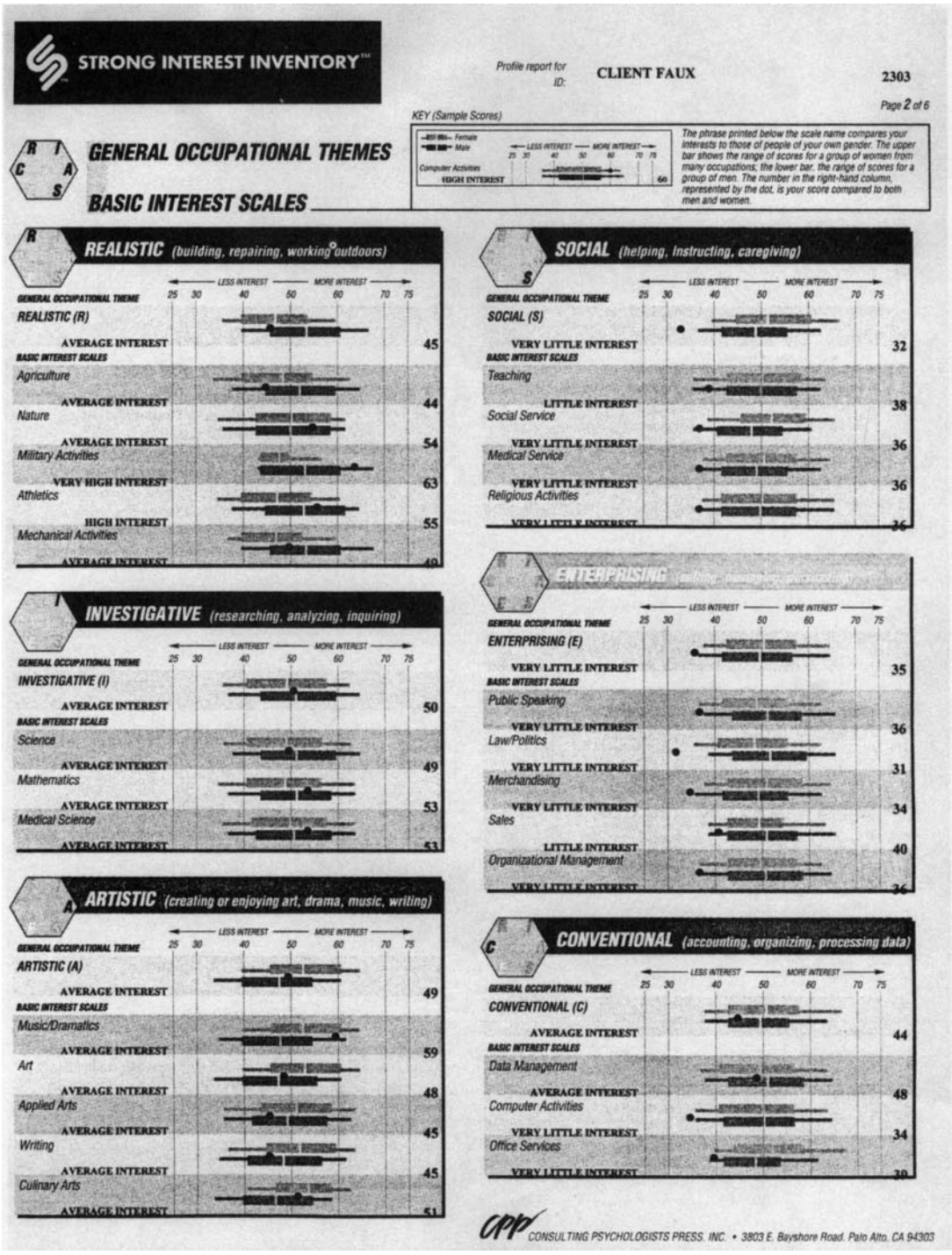


Figure 9.7. Profile for the Realistic and Investigative Occupational Scales of the Strong. Modified and reproduced by special permission of the Publisher, Consulting Psychologists Press, Inc., Palo Alto, CA 94303 from the *Strong Interest Inventory*™ of the *Strong Vocational Interest Blanks*® Form T317. Copyright 1933, 1938, 1945, 1946, 1966, 1968, 1974, 1981, 1985, 1994 by The Board of Trustees of the Leland Stanford Junior University. All rights reserved. Printed under license from Stanford University Press, Stanford, CA 94305. Further reproduction is prohibited without the Publisher's written consent. *Strong Interest Inventory* is a trademark and *Strong Vocational Interest Blanks* is a registered trademark of the Stanford University Press.

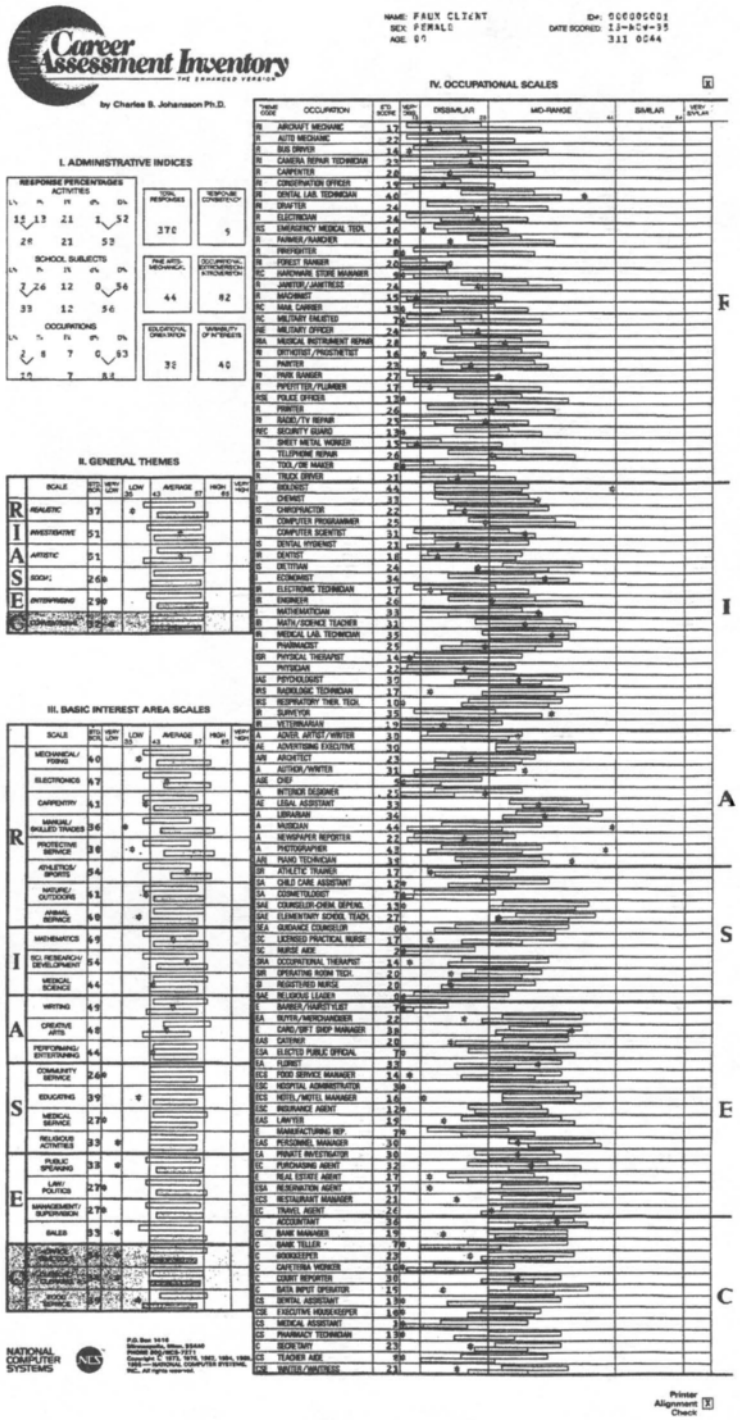


Figure 9.8. Profile for the Career Assessment Inventory. Copyright © 1973, 1986 NATIONAL COMPUTER SYSTEMS, INC. All rights reserved. Used here by permission.

Theme with which each Basic Interest Scale has its highest correlation.

Basic Interest Scales

The 25 Basic Interest Scales (BIS) were constructed using the statistical technique of cluster analysis to identify highly correlated items (Campbell, Borgen, Eastes, Johansson, & Peterson, 1968). The BIS were developed to focus on the measurement of only one interest factor per scale and, consequently, are easier to interpret than the heterogeneous Occupational Scales that incorporate items representing several interest factors as well as likes and aversions in each scale.

The BIS-scale names, as indicated in Figure 9.7, describe the homogeneous item content and the interest trait measured by each scale. Like the GOT, standard scores based on a combined-sex General Reference Sample, and interpretive bars based on female and male General Reference Samples are presented on the profile.

Personal Styles Scales

Four Personal Styles Scales—Work Style, Learning Environment, Leadership, and Risk Taking/Adventure—also are reported on the Strong profile. All four of these bi-polar scales were standardized using the combined-sex General Reference Sample; interpretive bars based on female and male General Reference Samples are presented on the profile.

The Work Style Scale is intended to identify people who prefer to work with ideas, data, and things (low scores) and those who prefer to work with people (high scores). The Learning Environment Scale distinguishes between people who prefer academic learning environments (high scores) and those who prefer practical training (low scores). Similarly, the Leadership Scale is meant to identify those who prefer to do a task themselves or to lead by example (low scores) and those who like to be in charge of others (high scores). The Risk-Taking/Adventure Scale, as the scale name implies, measures the extent to which an individual is willing to take risks.

Reliability and Validity

The test-retest reliability of the scales on the Strong profile is substantial over short and long intervals. Median reliabilities over one month and three month periods for the General Occupational Themes were .86, and .81; for the Basic Interest Scales were .85, and .80; and for the Occupational Scales were .87 and .85 (Harmon, et al., 1994).

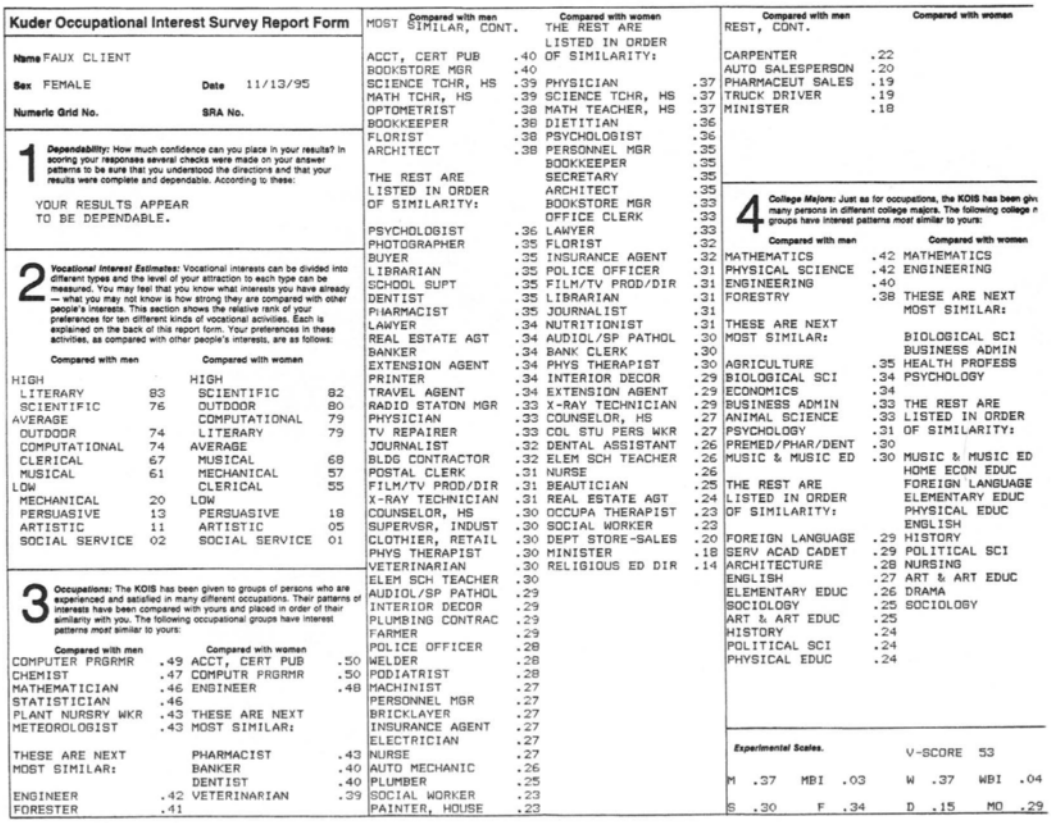
Because interest inventories are used to make long-term decisions, predictive validity is important. The Strong Interest Inventory has a long history of predictive validity studies for its various editions, however, no predictive validity data are available at this time for the 1994 Form. Data from earlier forms of the Strong show that, at least in the past, high scores on the Occupational Scales are related to occupations eventually entered; generally, between one-half and three-fourths of the subjects in predictive validity studies enter occupations predictable from their earlier scores (Campbell, 1966; Dolliver, Irvin & Bigley, 1972; Hansen, 1986; Spokane, 1979). Studies assessing the usefulness of the Strong Interest Inventory for predicting college majors have found hit rates similar to those reported for occupational entry (Hansen & Swanson, 1983; Hansen & Tan, 1992).

Career Assessment Inventory

The first edition of the Career Assessment Inventory™ (Johansson, 1975; Johansson & Johansson, 1978) was developed for use with individuals considering immediate career entry, community college education, or vocational-technical training, and was modeled after the Strong Interest Inventory™. In 1982, the decision was made to move from separate-sex to combined-sex Occupational Scales. The enhanced version of the Career Assessment Inventory™ published in 1986 (Johansson, 1986) has been expanded to include several Occupational Scales representing professional occupations.

The enhanced Career Assessment Inventory™ test booklet includes 370 items, and the profile reports three sets of scales: six homogeneous General Themes, 25 homogeneous Basic Interest Areas and 111 heterogeneous Occupational Scales. The Career Assessment Inventory™ uses Holland's theory to organize the Basic Interest Areas and Occupational Scales on the profile, clustering together those that represent each of Holland's six types (see Figure 9.8).

The General Themes and Basic Interest Areas are normed on a combined-sex reference sample composed of employed adults and students drawn



Kuder Occupational Interest Survey Report Form

Name FAUX CLIENT
Sex FEMALE Date 11/13/95
Numeric Grid No. SRA No.

1 *Dependability:* How much confidence can you place in your results? In scoring your responses several checks were made on your answer patterns to be sure that you understood the directions and that your results were complete and dependable. According to these:

YOUR RESULTS APPEAR TO BE DEPENDABLE.

2 *Vocational Interest Estimates:* Vocational interests can be divided into different types and the level of your attraction to each type can be measured. You may feel that you know what interests you have already — what you may not know is how strong they are compared with other people's interests. This section shows the relative rank of your preferences for ten different kinds of vocational activities. Each is explained on the back of this report form. Your preferences in these activities, as compared with other people's interests, are as follows:

Compared with men		Compared with women	
HIGH		HIGH	
LITERARY	83	SCIENTIFIC	82
SCIENTIFIC	76	OUTDOOR	80
AVERAGE		COMPUTATIONAL	79
OUTDOOR	74	LITERARY	79
COMPUTATIONAL	74	AVERAGE	
CLERICAL	67	MUSICAL	68
MUSICAL	61	MECHANICAL	57
LOW		CLERICAL	55
MECHANICAL	20	LOW	
PERSUASIVE	13	PERSUASIVE	18
ARTISTIC	11	ARTISTIC	05
SOCIAL SERVICE	02	SOCIAL SERVICE	01

3 *Occupations:* The KOIS has been given to groups of persons who are experienced and satisfied in many different occupations. Their patterns of interests have been compared with yours and placed in order of their similarity with you. The following occupational groups have interest patterns *most* similar to yours:

Compared with men		Compared with women	
COMPUTER PRGRMR	.49	ACCT, CERT PUB	.50
CHEMIST	.47	COMPUTR PRGRMR	.50
MATHEMATICIAN	.46	ENGINEER	.48
STATISTICIAN	.46		
PLANT NURSRY WKR	.43	THESE ARE NEXT	
METEOROLOGIST	.43	MOST SIMILAR:	
THESE ARE NEXT		PHARMACIST	.43
MOST SIMILAR:		BANKER	.40
		DENTIST	.40
ENGINEER	.42	VETERINARIAN	.39
FORESTER	.41		

MOST SIMILAR, CONT.	Compared with men	Compared with women THE REST ARE LISTED IN ORDER OF SIMILARITY:
ACCT, CERT PUB	.40	
BOOKSTORE MGR	.40	
SCIENCE TCHR, HS	.39	PHYSICIAN .37
MATH TCHR, HS	.39	SCIENCE TCHR, HS .37
OPTOMETRIST	.38	MATH TEACHER, HS .37
BOOKKEEPER	.38	DIETITIAN .36
FLORIST	.38	PSYCHOLOGIST .36
ARCHITECT	.38	PERSONNEL MGR .35
		BOOKKEEPER .35
THE REST ARE		SECRETARY .35
LISTED IN ORDER		ARCHITECT .35
OF SIMILARITY:		BOOKSTORE MGR .33
		OFFICE CLERK .33
PSYCHOLOGIST	.36	LAWYER .33
PHOTOGRAPHER	.35	FLORIST .32
BUYER	.35	INSURANCE AGENT .32
LIBRARIAN	.35	POLICE OFFICER .31
SCHOOL SUPT	.35	FILM/TV PROD/DIR .31
DENTIST	.35	LIBRARIAN .31
PHARMACIST	.35	JOURNALIST .31
LAWYER	.34	NUTRITIONIST .31
REAL ESTATE AGT	.34	AUDIOL/SP PATHOL .30
BANKER	.34	BANK CLERK .30
EXTENSION AGENT	.34	PHYS THERAPIST .30
PRINTER	.34	INTERIOR DECOR .29
TRAVEL AGENT	.34	EXTENSION AGENT .29
RADIO STATON MGR	.33	X-RAY TECHNICIAN .29
PHYSICIAN	.33	COUNSELOR, HS .27
TV REPAIRER	.33	COL STU PERS WKR .27
JOURNALIST	.32	DENTAL ASSISTANT .26
BLDG CONTRACTOR	.32	ELEM SCH TEACHER .26
POSTAL CLERK	.31	NURSE .26
FILM/TV PROD/DIR	.31	BEAUTICIAN .25
X-RAY TECHNICIAN	.31	REAL ESTATE AGT .24
COUNSELOR, HS	.30	OCCUPA THERAPIST .23
SUPERVSR, INDUST	.30	SOCIAL WORKER .23
CLOTHIER, RETAIL	.30	DEPT STORE-SALES .20
PHYS THERAPIST	.30	MINISTER .18
VETERINARIAN	.30	RELIGIOUS ED DIR .14
ELEM SCH TEACHER	.30	
AUDIOL/SP PATHOL	.29	
INTERIOR DECOR	.29	
PLUMBING CONTRAC	.29	
FARMER	.29	
POLICE OFFICER	.28	
WELDER	.28	
PODIATRIST	.28	
MACHINIST	.27	
PERSONNEL MGR	.27	
BRICKLAYER	.27	
INSURANCE AGENT	.27	
ELECTRICIAN	.27	
NURSE	.27	
AUTO MECHANIC	.26	
PLUMBER	.25	
SOCIAL WORKER	.23	
PAINTER, HOUSE	.23	

REST, CONT.	Compared with men	Compared with women
CARPENTER	.22	
AUTO SALESPERSON	.20	
PHARMACEUT SALES	.19	
TRUCK DRIVER	.19	
MINISTER	.18	

4 *College Majors:* Just as for occupations, the KOIS has been giv many persons in different college majors. The following college n groups have interest patterns most similar to yours:

Compared with men		Compared with women
MATHEMATICS	.42	MATHEMATICS
PHYSICAL SCIENCE	.42	ENGINEERING
ENGINEERING	.40	
FORESTRY	.38	THESE ARE NEXT MOST SIMILAR:
THESE ARE NEXT MOST SIMILAR:		BIOLOGICAL SCI
		BUSINESS ADMIN
AGRICULTURE	.35	HEALTH PROFESS
BIOLOGICAL SCI	.34	PSYCHOLOGY
ECONOMICS	.34	
BUSINESS ADMIN	.33	THE REST ARE
ANIMAL SCIENCE	.33	LISTED IN ORDER
PSYCHOLOGY	.31	OF SIMILARITY:
PREMED/PHAR/DENT	.30	
MUSIC & MUSIC ED	.30	MUSIC & MUSIC ED
		HOME ECON EDUC
THE REST ARE		FOREIGN LANGUAGE
LISTED IN ORDER		ELEMENTARY EDUC
OF SIMILARITY:		PHYSICAL EDUC
		ENGLISH
FOREIGN LANGUAGE	.29	HISTORY
SERV ACAD CADET	.29	POLITICAL SCI
ARCHITECTURE	.28	NURSING
ENGLISH	.27	ART & ART EDUC
ELEMENTARY EDUC	.26	DRAMA
SOCIOLOGY	.25	SOCIOLOGY
ART & ART EDUC	.25	
HISTORY	.24	
POLITICAL SCI	.24	
PHYSICAL EDUC	.24	

Experimental Scales.

			V-SCORE	53			
M	.37	MBI	.03	W	.37	WBI	.04
S	.30	F	.34	D	.15	MO	.29

Figure 9.9. Profile for the KOIS (Form DD). The Kuder Occupational Interest Survey Report Form is reproduced with the permission of National Career Assessment Services, Inc. All Rights Reserved.

from the six Holland interest areas; 75 females and 75 males from each of the six groups, for a total of 900 subjects, compose the sample. In addition to the standard scores based on a combined-sex sample, however, the Career Assessment Inventory™ profile presents bars for each scale representing the range of scores for females and males in the reference sample. These additional data help to circumvent the problem of gender differences on some of the homogeneous scales.

The Occupational Scales of the enhanced version of the Career Assessment Inventory™ were developed using the empirical method of contrast samples to select items that differentiated combined-sex criterion and general reference samples from each other. The combined-sex criterion samples fall short of the goal of equally representing females and males within the sample (e.g., 0 female aircraft mechanics, 0 male medical assistants, 0 female purchasing agents, and 0 male secretaries). The author attempted to improve the psychometrics of the scales by doing separate-sex item analyses if the separate-sex samples were large enough. In most instances, however, the sample representing one sex or the other was too small to produce reliable item analyses (e.g., 22 male bank tellers, 16 female dental laboratory technicians, 12 male bookkeepers, and 20 female enlisted personnel). In fact, 64 of the 111 Occupational Scales were developed using criterion samples that included less than 50 subjects representing one sex or the other (45 scales with < 50 female subjects and 19 scales with < 50 male subjects). Consequently, the exploration validity, for the scales developed with unbalanced female-male ratio criterion samples, is questionable for the underrepresented gender.

Kuder's Interest Inventories

The Personal Preference Record (Form A) was published in 1939 by Frederic Kuder and included seven almost independent homogeneous scales. Kuder added two more homogeneous scales in 1943 (Form B) and another homogeneous scale in 1948 (Form C). The Kuder General Interest Survey (Form E) (Kuder, 1988) measures the 10 interest areas of Form C but expresses the items in language that is easier to understand. The first edition of the Kuder Occupational Interest Survey (Form DD) was published in 1966; the latest additions and revisions are reported in the *Kuder Occupational Interest Survey Form DD, General Manual* (Kuder & Zytowski, 1991).

General Interest Survey (Form E)

The General Interest Survey (Form E) is composed of homogeneous scales that measure interest in 10 broad areas: Outdoor, Mechanical, Computational, Scientific, Persuasive, Artistic, Literary, Musical, Social Service, and Clerical. Kuder originally developed the scales by grouping related items on the basis of content validity; later he used item analyses to determine groups of items (scales) with high internal consistency.

The item booklet contains 168 forced-choice triads reported to be at the sixth grade reading level. The respondent compares each of the three activities with the other two and ranks them as most preferred (M) and least preferred (L).

The General Interest Survey (GIS) may be hand-scored or machine-scored; both techniques produce raw scores that are entered on a profile sheet. The respondent's raw scores, in turn, are compared with percentile distributions of either norm groups of girls or boys in grades 6 through 8 or grades 9 through 12.

The Kuder Preference Record-Vocational (Form C) is an earlier form that was designed for use with students in grades 9 to 12 and with adults. It uses more difficult vocabulary than does Form E but measures the same 10 areas of interest.

Kuder Occupational Interest Survey (Form DD)

The Kuder Occupational Interest Survey (KOIS) (Form DD) is composed of 100 triads of activities similar to those of the Kuder-Form E already described. The profile includes 104 Occupational Scales and 39 College Major Scales that, like the Strong Interest Inventory, compare the respondent's interests to those of people in criterion samples. Unlike the Strong, the KOIS does not use the empirical method of contrast groups for scale construction. Instead, the individual's responses are compared directly to those of the criterion samples, and scores are reported as Lambda coefficients, which do not allow comparison of scores across different persons' profiles as can be done with standard scores. Thus, a respondent's KOIS scores derive meaning only from the rank each scale occupies among all of the scales.

This form of the Kuder must be machine scored; the respondent receives the profile illustrated in Figure 9.9. The 109 Occupational Scales represent: (a) 33 occupations (66 scales) that were developed using both female and male criterion samples, (b) 32 that are based on male criterion samples only, and (c) 11 based on female criterion samples only. The 40 College Major Scales represent 14 majors (28 scales) that are based on female and male criterion samples, eight based on male samples only, and five on female samples.

In 1985 (Zytowski), a new profile for the KOIS was designed and 10 Vocational Interest Estimates (VIE scales) were added to the existing Occupational Scales. The VIE section of the profile is described as a short form (i.e., fewer items are included on each scale) of the earlier Kuder instruments that measure homogeneous or global areas of interests. Reliabilities of the new scales are acknowledged by Zytowski (1985) as less than those for Form E or Form C, precipitating the decision to call the scales "estimates" of interests.

The VIE scales are reported on the profile in rank order with divisions into high (75th percentile), average, and low (25th percentile) portions, as are the Occupational Scales and College Majors, based on percentile ranks (See Figure 9.9). The separate-sex norm samples for the VIE are composed of high school and college students and individuals from private agencies ($N = 1631$ women and 1583 men). The profile also offers instructions for converting the VIE percentiles to Holland codes by combining the various scales. For example, the Outdoor and Mechanical Scales are combined to estimate Holland's Realistic type, and Computational and Clerical are combined to represent the Conventional type.

Reliability and Validity. An inventory, such as the KOIS, which provides rank-ordered results

intended to discriminate interests within the respondent rather than to discriminate among people, has special requirements for analyses of reliability. Test-retest reliability can be assessed only in terms of the consistency of the order of scores for each subject from one testing to the next. Kuder and Diamond (1979) reported individual two-week test-retest Occupational Scale reliabilities computed for high school and college-age students; the median reliability for all cases was .90. Zytowski (1985), using college students (N = 192), reported profile stability of .80 for the VIEs over a two-week interval.

A large predictive validity study for the KOIS (Zytowski, 1976) involved over 800 women and men who were located 12 to 19 years after taking the Kuder. Fifty-one percent were employed in an occupation predicted by their scores on the KOIS.

Jackson Vocational Interest Inventory

The Jackson Vocational Interest Survey (JVIS) (Jackson, 1977), appropriate for high school and college students and adults who need assistance with educational and career planning, is composed of 289 forced-choice items describing occupational activities. The 34 homogeneous scales that measure *work roles* and *work styles* each contain 17 items estimated to be at the seventh-grade reading level. The work-role scales include five that characterize specific occupations (e.g., Engineering, Elementary Education) and 21 that represent a cluster of jobs (e.g., Creative Arts, Social Science). The eight work-style scales measure preferences for environments that require certain behaviors (e.g., Dominant Leadership, Accountability). The hand-scored JVIS profile includes only the 34 Basic Interest Scales; the machine-scored profile also includes 10 General Occupational Themes measuring broad patterns of interests that reflect the respondent's *orientation toward work* rather than interests (e.g., Logical, Enterprising); 17 broad clusters of university major fields (Educational Classifications) and 32 occupational clusters (Occupational Classifications).

Development of the 34 homogeneous Basic Interest Scales relied on a theory-based technique of scale construction. The process began with identification of the interests to be measured from previous research in vocational psychology. Then 3,000 items were written to represent the interest constructs. Finally, the item pool was submitted to a series of factor analyses to identify the 289 items that had high correlations with factor scores on their own scales and low correlations with other JVIS scales. The 10 General Occupational Themes later were constructed by factor analyzing the 34 Basic scales.

Standard score norms for the Basic and Theme scales are based on a combined-sex sample of female and male high school and college students. Interpretive bars representing the percentile distributions of scores of the females and males on each scale allow individuals to infer how their scores compare with that of other people. The Educational and Occupational Classifications involve analyses of an individual's entire profile of Basic scales compared to model profiles of college students in various academic majors and of people employed in a wide variety of occupations.

Vocational Interest Inventory

The Vocational Interest Inventory (VII) (Lunneborg, 1976, 1981), designed for use with young people, is similar to the JVIS on several dimensions. First, the interests to be measured were selected on theoretical considerations. The eight homogeneous scales of the VII were developed to represent the eight groups described in Roe's theory of occupational classifications: Service, Business Contact, Organization, Technical, Outdoor, Science, General Culture, and Arts and Entertainment. Second, the scales were constructed using a series of factor analyses that reduced the initial item pool to the final 112 forced-choice items. The eight scales each contain 28 response choices that have high correlations with factor scores on their own scales and low correlations with other VII scales. Third, the scales were normed on a combined-sex sample of students. According to the author (Lunneborg, 1981), scores on only two scales were unaffected by gender, and thus, the VII may have the problem of bias in interpretation for one sex or the other.

UNIACT

The revised edition of the unisex version of the ACT Interest Inventory (UNIACT-R) (Swaney, 1995) is a component in several of American College Testing's programs including the ACT Assessment Program used by college-

bound students in planning for college and in DISCOVER, a computer-based career-planning system for high school and college students and adults. The test booklet includes 90 items that are evenly distributed across six scales (15 items per scale) that are intended to assess interest in Holland's six types: Technical (Realistic), Science (Investigative), Arts (Artistic), Social Service (Social), Business Contact (Enterprising), and Business Operations (Conventional). In addition, 60 of the 90 items are used in the Data/Ideas and Things/People Summary Scales (30 items per scale).

The item pool for the UNIACT-R was developed with an emphasis on identifying items that (a) represented Holland's six types and (b) had a 10 percent or smaller sex difference in the percentages of "like" responses. Rational scale-construction techniques were used to initially assign items to each scale, and empirical analyses were used as a follow-up to make final refinements in the item composition of each scale.

Three sets of norms are provided for the UNIACT-R—Grade 8 (N = 4,631), Grade 10 (N = 4,133), and Grade 12 (N = 4,679)—with the intention that users will select the norm group that most closely resembles the age range of the students in their program. The median three-week test-retest reliability coefficient for the six Basic Interest Scales is .82. The coefficients for Data/Ideas and Things/People over the same interval are .87 and .82, respectively. Evidence of convergent and divergent validity and criterion-related validity contribute to the construct validity of the UNIACT-R and are reported in the manual (Swaney, 1995).

Interpretation of the UNIACT-R incorporates ACT's World-of-Work Map which arranges groups of similar jobs into 12 regions that are analogous to 12 pieces of a pie. The 12 regions represent various combinations of data, ideas, things, and people work-tasks that proceed around the circle in the same order hypothesized by Holland: Technical (R), Science (I), Arts (A), Social Service (S), Business Contact (E), and Business Operations (C). Clients are encouraged to explore occupations in the region indicated by their six Basic Interest Scale scores as well as in adjacent regions.

STABILITY OF INTERESTS

The degree to which interests are stable is important to the predictive power of inventories. If interests are fickle and unstable, interest inventory scores will not explain any of the prediction variance.

Stability of interests was one of the earliest concerns of researchers in interest measurement (Strong, 1943). Cross-sectional and longitudinal methods have been used in a plethora of studies to document that interests are stable even at relatively young ages of 15 or 16 years. By age 20 years, the stability of interests is obvious even over test-retest intervals of 5 to 10 years, and by age 25, interests are very stable (Hansen & Swanson, 1983; Johansson & Campbell, 1971; Swanson & Hansen, 1986).

During the long history of the Strong Interest Inventory, over 30 occupations have been tested at least three times: in the 1930s, 1960s, and 1970s/1980s. Analyses of these data have shown that interests of randomly sampled occupational groups are stable (Hansen, 1988a). Figure 9.10, a profile of interests for lawyers collected in the 1930s, 1960s, and 1970s, illustrates the typical finding for all the occupations:

1. The configuration of the interests of an occupation stays the same over long periods of time, and
2. even when interests change to some small extent, the relative importance of various interests stays the same. (Hansen, 1988a)

USE OF INTEREST INVENTORIES

Interest inventories are used to efficiently assess interests by a variety of institutions including high school and college advising offices, social service agencies, employment agencies, consulting firms, corporations, and community organizations such as the YWCA.

Career Exploration

The major use of assessed interests, usually reported as interest-inventory scores, is in career counseling that leads to decisions such as choosing a major, selecting an occupation, making a mid-career change, or preparing for retirement. First, counselors use the interest-inventory profiles

General Occupational Themes

Theme	30s	60s	70s
R-Theme	46	47	50
I-Theme	48	50	50
A-Theme	46	49	50
S-Theme	46	51	49
E-Theme	46	52	52
C-Theme	48	47	45

Basic Interest Scales

Scale	30s	60s	70s
Nature	44	46	48
Adventure	48	52	54
Mechanical Activities	47	47	48
Science	50	49	48
Mathematics	51	50	48
Medical Science	46	48	50
Music Dramatics	47	50	52
Art	44	45	48
Writing	50	52	53
Teaching	45	50	49
Social Service	47	49	46
Athletics	50	52	50
Public Speaking	54	59	58
Law Politics	57	59	59
Merchandising	44	49	49
Sales	45	49	49
Business Management	46	51	50
Office Practices	46	47	46

Figure 9.10. Mean interest profile for male lawyers tested in the 1930s (•—•), the 1960s (x—x), and the 1970s (o—o).

to develop hypotheses about clients that may be discussed, confirmed, or discarded during career exploration. Then, the interest scores and profile provide a framework for interest exploration and a mechanism for helping the client to integrate her or his past history with current interests.

Inventory results serve as a starting point for evaluating interests, as an efficient method for objectively identifying interests, and as a structure for the counseling process. Inventory results help some counselees to increase the number of options they are considering; some use the results to begin to narrow the range of possible choices. Others only want to confirm educational or vocational decisions that they already have made.

Selection and Placement

Interest inventories also are used to assess interests during employment selection and placement evaluations. Among qualified candidates, interest inventories help to identify those most likely to complete the training program and stay in the profession (Berdie & Campbell, 1968; Reeves & Booth, 1979). Even after initial selection, interest inventories may be used to help an employee find the right job within the company (Dunnette & Kirchner, 1965; Hansen, 1994).

Research

Researchers use measures of interests (e.g., check-lists, self-estimates, rating scales, interest inventories) to operationalize interest traits, investigate the origin and development of interests, explore changes or stability in society, and understand the relationship between interests and other psychological variables such as abilities, satisfaction, success, and personality. Studies assessing the structure of interests and also the interests of various occupational groups provide information for understanding the organization of the world of work and the relationships among occupations.

Most interest inventories are constructed to measure vocational interests. Recent research, however, indicates that instruments such as the Strong Interest Inventory™ measure not only vocational interests but also leisure interests (Cairo, 1979; Varca & Shaffer, 1982). Holland (1973) has proposed that instruments measuring his six personality types also can identify a respondent's preferences for environments and types of people as well as job activities.

FUTURE DIRECTIONS

The frequency of test use in counseling has not changed appreciably in the last 30 years; however, use of interest inventories has increased while use of other tests (e.g., ability, aptitude, achievement) has decreased (Engen, Lamb, & Prediger, 1982; Watkins, Campbell & McGregor, 1988; Zytowski & Warman, 1982). A wide variety of new interpretive materials, career-guidance packages, and interactive computerized systems for inventory interpretation and career exploration is available. Thus far, evaluations of the use of interest inventories indicate that various modes and mediums of presentation are equally effective (Hansen, Neuman, Haverkamp, & Lubinski, 1997; Johnson, Korn, & Dunn, 1975; Maola & Kane, 1976; Miller & Cochran, 1979; Rubinstein, 1978; Smith & Evans, 1973; Vansickle & Kapes, 1993; Vansickle, Kimmel & Kapes, 1989). The trend in the future, with decreasing budgets and personnel in educational institutions, will be toward even greater use of computers for interest- inventory administration and interpretation and for integration into computerized career-counseling modules.

Techniques for developing reliable and valid interest inventories are available now, and the construction methods have reached a plateau of excellence in reliability and validity. Therefore, publishers can direct their efforts toward an increased emphasis on interpretation and counselor competency. Test manuals traditionally were written to provide data required by the American Psychological Association's *Standards for Educational and Psychological Testing* (1985); now, interpretive manuals are prepared in addition to technical manuals to help the professional maximize the usefulness of inventory results (Hansen, 1992; Holland, 1971, 1987a; Zytowski, 1981). Increasingly publishers are attempting to develop testing packages that integrate interest inventories with other psychological measures such as personality inventories or self-efficacy measures (e.g., the Strong and the Myers-Briggs Type Indicator). Unfortunately, these packages have been released by publishers without expending much effort to

collect data to assess the validity of using the instruments as a package.

As the use of interest inventories expands to new populations, research must also move in that direction to aid in understanding the characteristics of the populations as well as the best methods for implementing interest inventories with them. The cross-cultural use of interest inventories also is increasing the demand for valid translations of inventories and for data on the predictive accuracy of inventories normed on U.S. populations for non-English-speaking respondents (Fouad & Spreda, 1995).

SUMMARY

Interest inventories will be used in the future as in the past to operationalize the trait of interests in research. Attempts to answer old questions, such as the interaction of interests and personality, success, values, satisfaction, and ability will persevere.

Holland's theory undoubtedly will continue to evoke research in the field. Studies designed to understand educational and vocational dropouts and changers, to analyze job satisfaction, to understand the development of interests, and to predict job or academic success will draw on Holland's theoretical constructs for independent variables and on interest inventories to identify interests. Exploration of vocational interests always has been a popular topic in counseling psychology; the increased use of inventories and career guidance programs indicates that interest inventories will continue to be an important component in psychological research.

NOTES

1. Strong Interest Inventory is a trademark of the Standford University Press.

2. Career Assessment Inventory is a trademark of NATIONAL COMPUTER SYSTEMS, INC.

3. Campbell Interest and Skill Survey is a trademark and "CISS" is a registered trademark of David P. Campbell, Ph.D.

4. Strong Vocational Interest Blanks is a registered trademark of the Stanford University Press.

REFERENCES

American Psychological Association. (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Berdie, R. F., & Campbell, D. P. (1968). Measurement of interest. In D. K. Whitla (Ed.), *Handbook of measurement and assessment in behavioral sciences.* Reading, MA: Addison-Wesley.

Cairo, P. C. (1979). The validity of the Holland and Basic Interest Scales of the Strong Vocational Interest Blank: Leisure activities versus occupational membership as criteria. *Journal of Vocational Behavior, 15,* 68–77.

Campbell, D. P. (1966). Occupations ten years later of high school seniors with high scores on the SVIB life insurance salesman scale. *Journal of Applied Psychology, 50,* 369-372.

Campbell, D. P. (1974). *Manual for the SVIB-SCII.* Stanford, CA: Stanford University Press.

Campbell, D. P. (1995). The Campbell Interest and Skill Survey (CISS): A Product of ninety years of psychometric evolution. *Journal of Career Assessment, 3,* 391–410.

Campbell, D. P., Borgen, F. H., Eastes, S. H., Johansson, C. B., & Peterson, R. A. (1968). A set of basic interest scales for the Strong Vocational Interest Blank for Men. *Journal of Applied Psychology Monograph, 52,* 1-54.

Campbell, D. P., & Hansen, J. C. (1981). *Manual for the SVIB-SCII* (3rd ed.). Stanford, CA: Stanford University Press

Campbell, D. P., & Holland, J. L. (1972). Applying Holland's theory to Strong's data. *Journal of Vocational Behavior, 2,* 353-376.

Campbell, D. P., Hyne, S. A., & Nilsen, D. L. (1992). *Manual for the Campbell Interest and Skill Survey.* Minneapolis, MN: National Computer Systems.

Cole, N. S., & Hansen, G. (1971). *An analysis of the structure of vocational interests* (ACT Research Report No. 40). Iowa City, IA: American College Testing Program.

Cowdery, K. M. (1926). Measurement of professional attitudes: Differences between lawyers, physicians, and engineers. *Journal of Personnel Research, 5,* 131-141.

Dolliver, R. H., Irvin, J. A., & Bigley, S. E. (1972). Twelve-year follow-up of the Strong Vocational Interest Blank. *Journal of Counseling Psychology, 19,* 212-217.

Dunnette, M. D., & Kirchner, W. K. (1965). *Psychology applied to industry.* New York: Appleton-Century-Crofts.

Edwards, K. J., & Whitney, D. R. (1972). A structural analysis of Holland's personality types using fac-

tor and configural analysis. *Journal of Counseling Psychology, 19*, 136-145.

Engen, H. B., Lamb, R. R., & Prediger, D. J. (1982). Are secondary schools still using standardized tests? *Personnel and Guidance Journal, 60*, 287-290.

Fouad, N. A., & Spreda, S. L. (1995). Use of interest inventories with special populations: Women and minority groups. *Journal of Career Assessment, 4*, 453–468.

Freyd, M. (1922-1923). The measurement of interests in vocational selection. *Journal of Personnel Research, 1*, 319-328.

Guilford, J. P., Christensen, P. R., Bond, N. A., Jr., & Sutton, M. A. (1954). A factor analysis study of human interests. *Psychological Monographs*, Whole No. 375, *68*, 1-38.

Hansen, J. C. (1983). *Correlation between VPI and SCII scores*. Unpublished manuscript. Center for Interest Measurement Research, University of Minnesota.

Hansen, J. C. (1986, August). *12-Year longitudinal study of the predictive validity of the SVIB-SCII*. Paper presented at the meetings of the American Psychological Association, Washington, DC.

Hansen, J. C. (1988). Changing interests: Myth or reality? *Applied Psychology: An International Review. 37*, 133-150

Hansen, J. C. (1992). *User's guide to the SII* (2nd edition). Palo Alto, CA: Consulting Psychologists Press.

Hansen, J. C. (1994). The measurement of vocational interests. In M. G. Rumsey, C. B. Walk, & J. H. Harris (Eds., *Personnel selection and classification*. Hillsdale, NJ: Lawrence Erlbaum.

Hansen, J. C., & Campbell, D. P. (1985). *Manual for the SVIB-SCII* (4th ed.). Stanford, CA: Stanford University Press.

Hansen, J. C., Collins, R., Swanson, J. L., & Fouad, N. A. (1993). Gender differences in the structure of interests. *Journal of Vocational Behavior, 42*, 200–211.

Hansen, J. C., & Johansson, C. B. (1972). The application of Holland's vocational model to the Strong Vocational Interest Blank for Women. *Journal of Vocational Behavior, 2*, 479-493.

Hansen, J. C., Neuman, J., Haverkamp, B. E., & Lubinski, B. R. (1997). Comparison of user reaction to two methods of SII administration and report feedback. *Measurement and Evaluation in Counseling and Development, 30*, 115-117.

Hansen, J. C., & Swanson, J. L. (1983). Stability of interests and the predictive and concurrent validity of the 1981 Strong-Campbell Interest Inventory. *Journal of Counseling Psychology, 30*, 194-201.

Hansen, J. C., & Tan, R. N. (1992). Concurrent validity of the 1985 Strong Interest Inventory for college major selection. *Measurement and Evaluation in Counseling and Development, 25*, 53–57.

Harmon, L., Hansen, J. C., Borgen, F., & Hammer, A. (1994). *Strong Interest Inventory applications and technical guide*. Stanford, CA: Stanford University Press.

Harrington, T. F., Jr., & O'Shea, A. J. (1993). *Manual for the Career Decision-Making System- Revised*. Circle Pines, MN: American Guidance Service.

Haverkamp, B. E., Collins, R. C., & Hansen, J. C. (1994). Structure of interests of Asian-American college students. *Journal of Counseling Psychology, 41*, 256–264.

Holland, J. L. (1958). A personality inventory employing occupational titles. *Journal of Applied Psychology, 42*, 36-342.

Holland, J. L. (1959). A theory of vocational choice. *Journal of Counseling Psychology, 6*, 35-45.

Holland, J. L. (1962). Some explorations of a theory of vocational choice: I. One- and two-year longitudinal studies. *Psychological Monographs, 76*, 26.

Holland, J. L. (1966). *The psychology of vocational choice*. Waltham, MA: Blaisdell.

Holland, J. L. (1971). *The counselor's guide to the self-directed search*. Palo Alto, CA: Consulting Psychologists Press.

Holland, J. L. (1973). *Making vocational choices: A theory of careers*. Englewood Cliffs, NJ: Prentice-Hall.

Holland, J. L. (1978). *Manual for the Vocational Preference Inventory* (3rd ed.). Palo Alto, CA: Consulting Psychologists Press.

Holland, J. L. (1979). *The Self-Directed Search professional manual*. Palo Alto, CA: Consulting Psychologists Press.

Holland, J. L. (1985a). *The occupations finder*. Odessa, FL: Psychological Assessment Resources.

Holland, J. L. (1985b). *Professional manual for The Self-Directed Search*. Odessa, FL: Psychological Assessment Resources.

Holland, J. L. (1985c). *Vocational Preference Inventory (VPI) manual—1985 edition*. Odessa, FL: Psychological Assessment Resources.

Holland, J. L. (1987a). *1987 manual supplement for the Self-Directed Search*. Odessa, FL: Psychological Assessment Resources.

Holland, J. L. (1987b). *The college majors finder*. Odessa, FL: Psychological Assessment Resources.

Holland, J. L. (1992). *Making vocational choices: A theory of vocational personalities and work environments* (2nd edition). Odessa, FL: Psychological Assessment Resources.

Holland, J. L. (1994). *Self-Directed Search Form R: 1994 edition.* Odessa, FL: Psychological Assessment Resources.

Holland, J. L., & Lutz, S. W. (1968). Predicting a student's vocational choice. *Personnel and Guidance Journal, 46,* 428-436.

Hubbard, R. M. (1930). Interest Analysis Blank. In D. G. Paterson, R. M. Elliott, L. D. Anderson, H. A. Toops, & E. Heidbrider (Eds.), *Minnesota Mechanical Ability Test.* Minneapolis, MN: University of Minnesota Press.

Jackson, D. N. (1977). *Jackson Vocational Interest Survey manual.* London, Ontario: Research Psychologists Press.

Johansson, C. B. (1975). *Manual for the Career Assessment Inventory.* Minneapolis, MN: National Computer Systems.

Johansson, C. B. (1986). *Career Assessment Inventory: Enhanced version.* Minneapolis, MN: National Computer Systems.

Johansson, C. B., & Campbell, D. P. (1971). Stability of the Strong Vocational Interest Blank for Men. *Journal of Applied Psychology, 55,* 24-26.

Johansson, C. B., & Johansson, J. C. (1978). *Manual supplement for the Career Assessment Inventory.* Minneapolis, MN: National Computer Systems.

Johnson, W. F., Korn, T. A., & Dunn, D. J. (1975). Comparing three methods of presenting occupational information. *Vocational Guidance Quarterly, 24,* 62-65.

Kitson, H. D. (1925). *The psychology of vocational adjustment.* Philadelphia: Lippincott.

Kornhauser, A. W. (1927). Results from a quantitative questionnaire of likes and dislikes used with a group of college freshmen. *Journal of Applied Psychology, 11,* 85-94.

Kuder, G. F. (1939). *Kuder Preference Record—Form A.* Chicago: University of Chicago Bookstore.

Kuder, G. F. (1943). *Vocational Preference Record—Form B.* Chicago: Science Research Associates.

Kuder, G. F. (1948). *Kuder Preference Record—Form C (Vocational).* Chicago: Science Research Associates.

Kuder, G. F. (1966). *General manual: Occupational Interest Survey, Form DD.* Chicago: Science Research Associates.

Kuder, G. F. (1988). *Kuder General Interest Survey, Form E,* Chicago: Science Research Associates.

Kuder, G. F., & Diamond, E. E. (1979). *Kuder Occupational Interest Survey: General manual* (2nd ed.). Chicago: Science Research Associates.

Kuder, G. F., & Zytowski, D. G. (1991). *Kuder Occupational Interest Survey Form DD, general manual.* Monterey, CA: California Testing Bureau.

Lamb, R. R., & Prediger, D. J. (1981). *Technical report for the unisex edition of the ACT Interest Inventory.* Iowa City, IA: American College Testing Program.

Lunneborg, P. W. (1976). *Manual for the Vocational Interest Survey.* Seattle, WA: University of Washington, Educational Assessment Center.

Lunneborg, P. W. (1981). *Vocational Interest Inventory manual.* Los Angeles: Western Psychological Services.

Maola, J., & Kane, G. (1976). Comparison of computer-based versus counselor-based occupational information systems with disadvantaged vocational students. *Journal of Counseling Psychology, 23,* 163-165.

Miller, M. J., & Cochran, J. R. (1979). Evaluating the use of technology in reporting SCH results to students. *Measurement and Evaluation in Guidance, 12,* 166–173.

Miner, J. B. (1922). An aid to the analysis of vocational interests. *Journal of Educational Research, 5,* 311-323.

Parsons, F. (1909). *Choosing a vocation.* Boston: Houghton Mifflin.

Paterson, D. G., Elliott, R. M., Anderson, L. D., Toops, H. A., & Heidbreder, E. (Eds.). (1930). *Minnesota Mechanical Abilities Test.* Minneapolis, MN: University of Minnesota Press.

Prediger, D. J. (1982). Dimensions underlying Holland's hexagon: Missing link between interests and occupations? *Journal of Vocational Behavior, 15,* 155-163.

Reeves, D. J., & Booth, R. F. (1979). Expressed versus inventoried interests as predictors of paramedical effectiveness. *Journal of Vocational Behavior, 15,* 155-163.

Remmers, H. H. (1929). The measurement of interest differences between students of engineering and agriculture. *Journal of Applied Psychology, 13,* 105-119.

Roe, A. (1956). *The psychology of occupations.* New York: Wiley.

Rounds, J. B. (1995). Vocational interests. Evaluating structural hypotheses. In R. V. Dawis & D. Lubinski (Eds.), *Individual differences and assessment.* Palo Alto, CA: Consulting Psychologists Press.

Rubenstein, M. R. (1978). Integrative interpretation of vocational inventory results. *Journal of Counseling Psychology, 25,* 306-309.

Smith, R. D., & Evans, J. (1973). Comparison of experimental group and individual counseling facilitators of vocational development. *Journal of Counseling Psychology, 20,* 202-208.

Spokane, A. R. (1979). Occupational preferences and the validity of the Strong-Campbell Interest Inventory for college women and men. *Journal of Counseling Psychology, 26,* 312-318.

Strong, E. K., Jr. (1927). *Vocational Interest Blank.* Stanford, CA: Stanford University Press.

Strong, E. K., Jr. (1943). *Vocational interests of men and women*. Stanford, CA: Stanford University Press.

Swaney, K. B. (1995). *Technical manual: Revised unisex edition of the ACT Interest Inventory (UNIACT)*. Iowa City, IA: American College Testing Program.

Swanson, J. L., & Hansen, J. C. (1986). A clarification of Holland's construct of differentiation: The importance of score elevation. *Journal of Vocational Behavior, 28*, 163–173.

Vansickle, T. R., & Kapes, J. T. (1993). Comparing paper-pencil and computer-based versions of the Strong-Campbell Interest Inventory. *Computers in Human Behavior, 9*, 441–449.

Vansickle, T. R., Kimmel, C., & Kapes, J. T. (1989). A comparison of the computer-based and paper-pencil versions of the Strong Campbell Interest Inventory. *Measurement and Evaluation in Counseling and Development, 22*, 88–93.

Varca, P. E., & Shaffer, G. S. (1982). Holland's theory: Stability of avocational interests. *Journal of Vocational Behavior, 21*, 288-298.

Watkins, C. E., Jr., Campbell, V. C., & McGregor, P. (1988). Counseling psychologists' use of and opinions about psychological tests. *The Counseling Psychologist, 16*, 476–486.

Zytowski, D. G. (1976). Predictive validity of the Kuder Occupational Interest Survey: A 12- to 19-year follow-up. *Journal of Counseling Psychology, 3*, 221-233.

Zytowski, D. G. (1981). *Counseling with the Kuder Occupational Interest Survey*. Chicago: Science Research Associates.

Zytowski, D. G. (1985). *Kuder DD manual supplement*. Chicago: Science Research Associates.

Zytowski, D. G., & Warman, R. E. (1982). The changing use of tests in counseling. *Measurement and Evaluation in Guidance, 15*, 147-152.

This Page Intentionally Left Blank

PART V

NEUROPSYCHOLOGICAL ASSESSMENT

This Page Intentionally Left Blank

CHAPTER 10

COMPREHENSIVE NEUROPSYCHOLOGICAL ASSESSMENT BATTERIES

Gerald Goldstein

INTRODUCTION

This chapter is the first of three covering the area of neuropsychological assessment. It will therefore provide a general introduction to the field of neuropsychological assessment and deal specifically with the extensive standard test batteries used with adults. Neuropsychological assessment is a relatively new term that has essentially replaced the older terms "testing for brain damage" or "testing for organicity." Lezak (1995) indicates that these procedures are used for three purposes: diagnosis, provision of information important for patient care, and research. A significant component of the patient care function is rehabilitation planning and monitoring (Goldstein, 1978; Goldstein & Beers, 1998; Goldstein & Ruthven, 1983; Meier, Benton, & Diller, 1987). The focus of neuropsychological assessment has traditionally been on the brain-damaged patient, but there have been major extensions of the filed to psychiatric disorders (Goldstein, 1986; 1991; Yozawitz, 1986), functioning of non-brain-damaged individuals with medical disorders (Ryan, 1998) and normal aging (Goldstein & Shelly, 1975; Nussbaum, 1997).

Perhaps the best definition of a neuropsychological test has been offered by Ralph Reitan, who describes it as a test that is sensitive to the condition of the brain. If performance on a test changes with a change in brain function, then the test is a neuropsyhoclogical test. However, it should be pointed out that the comprehensive neuropsychological test batteries should not only contain neuropsychological tests. They should also contain some tests that are generally insensitive to brain dysfunction, primarily because such tests are often useful for providing a baseline against which extent of impairment associated with acquired brain damage can be measured. Most neuropsychological assessment methods are formal tests, but some work has been done with rating scales and self-report measures. Neuropsychological assessment is rarely conducted through a structured interview outside of a test situation.

A comprehensive neuropsychological test battery is a procedure that assesses all of the major functional areas generally affected by structural brain damage. We use the term ideally because none of the standard, commonly available procedures entirely achieves full comprehensiveness. Some observers have described the comprehensive procedures as screening batteries, because feasibility and time constraints generally require a sacrifice of detailed investigations of specific areas in order to achieve comprehensiveness. In Dr. Larrabee's chapter, we will learn more about what a clinical neuropsychologist does when asked to explore a particular area in detail rather than do a comprehensive evaluation. While the term *screening* may be justifiable in certain respects, the

extensive standard batteries in common use should not be grouped with the brief, paper-and-pencil screening tests used in many clinical and industrial settings. That is, they do not simply screen for presence or absence of brain damage, but also evaluate a number of functional areas that may be affected by brain damage. Since brain damage most radically affects cognitive processes, most neuropsychological tests assess various areas of cognition, but perception and motor skills are also frequently evaluated. Thus, neuropsychological tests are generally thought of as assessment instruments for a variety of cognitive, perceptual, and motor skills. That is not to say that brain damage does not affect other aspects of the personality, but traditionally the standard neuropsychological tests do not typically assess these other areas. Perhaps the most important reason for this preference is that cognitive tests have proven to be the most diagnostic ones. While personality changes may occur with a wide variety of psychiatric, general medical, and neurological conditions, cognitive changes appear to occur most dramatically in individuals with structural brain damage.

Numerous attempts have been made to classify the functional areas typically affected by brain damage, but the scheme proposed in what follows is a reasonably representative one. Perhaps the most ubiquitous change is general intellectual impairment. Following brain damage, the patient is not as bright as he or she was before. Problems are solved less effectively, goal-directed behavior becomes less well organized, and there is impairment of a number of specific skills such as solving arithmetic problems or interpreting proverbs. Numerous attempts have been made to epitomize this generalized loss, perhaps the most effective one being Goldstein and Scheerer's (1941) concept of impairment of the abstract attitude. The abstract attitude is a phenomenological concept having to do with the way in which the individual perceives the world. Some consequences of its impairment involve failure to form concepts or to generalize from individual events, failure to plan ahead ideationally, and inability to transcend the immediate stimulus situation. While the loss is a general one involving many aspects of the individual's life, it is best observed in a testing setting where the patient is presented with a novel situation in which some problem must be solved. Typically these tests involve abstraction or concept formation, and the patient is asked to sort or categorize in some way. The Goldstein-Scheerer tests (1941), perhaps the first neuropsychological battery, consist largely of sorting tests, but also provide the patient with other types of novel problem-solving tasks.

Probably the next most common manifestation of structural brain damage is impairment of memory. Sometimes memory impairment is associated with general intellectual impairment, sometimes it exists independently, and sometimes it is seen as an early sign of a progressive illness that eventually impairs a number of abilities other than memory. In most, but not all cases, recent memory is more impaired than remote memory. That is, the patient may recall his or her early life in great detail, but may be unable to recall what happened during the previous day. Often, so-called primary memory is also relatively well preserved. That is, the patient may be able to immediately repeat back what was just presented to him, such as a few words or a series of digits, but will not retain new information over a more extended period of time, particularly after intervening events have occurred. In recent years, our capacity to examine memory has benefited from a great deal of research involving the various amnesic syndromes (e.g., Baddeley, Wilson, & Watts, 1995; Butters & Cermak, 1980), and we have become quite aware that not all brain damaged patients experience the same kind of memory disorder (Butters, 1983). It generally requires a detailed assessment to specifically identify the various types of memory disorder, and the comprehensive batteries we will be discussing here generally can only detect the presence of a memory disorder and provide an index of its severity.

Loss of speed in performing skilled activities is an extremely common symptom of brain damage. Generally, this loss is described in terms of impaired psychomotor speed or perceptual-motor coordination. While its basis is sometimes reduction of pure motor speed, in many instances pure speed is preserved in the presence of substantial impairment on tasks involving speed of some mental operation or coordination of skilled movement with perceptual input. Thus, the patient may do well on a simple motor task such as finger tapping, but poorly on a task in which movement must be coordinated with visual input, such as a cancellation or substitution task. Tasks of this latter type are commonly performed slowly and laboriously by many kinds of brain damaged patients. Aside from slowness, there may be other disturbances of purposive movement that go under the general heading of apraxia. Apraxia may be manifested as simple clumsiness or awkwardness, an inability to

carry out goal directed movement sequences as would be involved in such functional activities as dressing, or as an inability to use movement ideationally as in producing gestures or performing pretended movements. While apraxia in one of its pure forms is a relatively rare condition, impairment of psychomotor speed is quite common and seen in a variety of conditions.

A set of abilities that bridge movement and perception may be evaluated by tasks in which the patient must produce some form of construction or copy a figure from a model. Among the first tests used to test brain-damaged patients was the Bender-Gestalt (Bender, 1938) a procedure in which the patient must copy a series of figures devised by Werthheimer (1923) to study perception of visual gestalten. It was found that many patients had difficulty copying these figures, although they apparently perceived them normally. These difficulties manifested themselves in reasonably characteristic ways, including various forms of distortion, rotation of the figure, simplification, or primitivation and perseveration. The copying task has continued to be used by neuropsychologists, either in the form of the Bender-Gestalt or a variety of other procedures. Variations of the copying- task procedure have involved having the patient draw the figure from memory (Bention, 1963; Rey, 1941), from a verbal command, e.g., "Draw a Circle" (Luria, 1973), or copy a figure that is embedded in an interfering background pattern (Canter, 1970). Related to the copying task is the constructional task, in which the patient must produce a three-dimensional construction from a model. The most popular test for this purpose is the Kohs Blocks or Block Design subtest of the Wechsler Scales (Wechsler, 1997). While in the timed versions of these procedures the patient may simply fail the task by virtue of running out of time, at least some brain-damaged patients make errors on these procedures comparable to what is seen on the copying tasks. With regard to block-design type tasks, the errors might involve breaking the contour of the model or incorrectly reproducing the internal structure of the pattern (Kaplan, 1979). Thus, a constructional deficit may not be primarily associated with reduction in psychomotor speed, but rather the inability to build configurations in three-dimensional space. Often, the ability is referred to as visual-spatial skill.

Visual-spatial skills also form a bridge with visual perception. When one attempts to analyze the basis for a patient's difficulty with a constructional task, the task demands may be broken down into movement, visual, and integrative components. Often, the patient has no remarkable impairment of purposive, skilled movement and can recognize the figure. If it is nameable, the patient can tell you what it is or if it is not, it can be correctly identified on a recognition task. However, the figure cannot be accurately copied. While the difficulty may be with the integration between the visual percept and the movement, it has also been found that patients with constructional difficulties, and indeed patients with brain damage in general, frequently have difficulties with complex visual perception. For example, they do poorly at embedded figures tasks (Teuber, Battersby, & Bender, 1951) or at tasks in which a figure is made difficult to recognize through displaying it in some unusual manner, such as overlapping it with other figures (Golden, 1981) or presenting it in some incomplete or ambiguous form (Mooney, 1957; Warrington & James, 1991). Some brain-damaged patients also have difficulty when the visual task is made increasingly complex through adding elements in the visual field. Thus, the patient may identify a single element, but not two. When two stimuli are presented simultaneously, the characteristic error is that the patient reports only seeing one. The phenomenon is known as extinction (Bender, 1952) or neglect (Jeannerod, 1987).

Many brain-damaged patients also have deficits in the areas of auditory and tactile perception. Sometimes, the auditory impairment is such that the patient can hear, but sounds cannot be recognized or interpreted. The general condition is known as agnosia and can actually occur in the visual, auditory, or tactile modalities. Agnosia has been defined as "perception without meaning," implying the intactness of the primary sense modality but loss of the ability to comprehend the incoming information. Auditory agnosia is a relatively rare condition, but there are many disturbances of auditory perception that are commonly seen among brain-damaged patients. Auditory neglect can exist and it is comparable to visual neglect; sounds to either ear may be perceived normally, but when a sound is presented to each ear simultaneously, only one of them may be perceived. There are a number of auditory verbal problems that we will get to when we discuss language. Auditory attentional deficits are common and may be identified by presenting complex auditory stimuli, such as rhythmic patterns, which the patient must recognize or reproduce immediately

after presentation. A variety of normal and abnormal phenomena may be demonstrated using a procedure called dichotic listening (Kimura, 1961). It involves presenting two different auditory stimuli simultaneously to each ear. The subject wears earphones, and the stimuli are presented using stereophonic tape. Higher level tactile deficits generally involve a disability with regard to identifying symbols or objects by touch. Tactile neglect may be demostrated by touching the patient over a series of trials with single and double stimuli, and tactile recognition deficits may be assessed by asking the patient to name objects placed in his or her hand or to identify numbers or letters written on the surface of the skin. It is particularly difficult to separate primary sensory functions from higher cognitive processes in the tactile modality, and many neuropsychologists perform rather detailed sensory examinations of the hands, involving such matters as light touch thresholds, two-point discrimination, point localization, and the ability to distinguish between sharp and dull tactile stimuli (Golden, Purisch, & Hammeke, 1985; Semmes, Weinstein, Ghent, & Teuber, 1960).

The neuropsychological assessment of speech and language has in some respects become a separate discipline involving neuropsychologists, neurologists, and speech and language pathologists. There is an extensive interdisciplinary literature in the area (Albert, Goodglass, Helm, Rubens, & Alexander, 1981; Benson & Ardila, 1996), and several journals that deal almost exclusively with the relationships between impaired or normal brain function and language (e.g., *Brain and Language*). Aphasia is the general term used to denote impairment of language abilities as a result of structural brain damage, but not all brain-damaged patients with communicative difficulties have aphasia. While aphasia is a general term covering numerous subcategories, it is now rather specifically defined as an impairment of communicative ability associated with focal damage to the left hemisphere in most people. Stroke is probably the most common cause of aphasia.

Historically, there have been numerous attempts to categorize the subtypes of aphasia (Goodglass, 1983), but in functional terms, the aphasias involve a rather dramatic impairment of the capacity to speak, to understand the speech of others, to find the names for common subjects, to read (alexia), write (agraphia), calculte (acalculia), or to use or comprehend gestures. However, a clinically useful assessment of these functional disorders must go into their specific characteristics. For example, when we say the patient has lost the ability to speak, we may mean that he or she has become mute or can only produce a few utterances in a halting, labored way. On the other hand, we may mean that the patient can produce words fluently, but the words and sentences being uttered make no sense. When it is said that the patient does not understand language, that may mean that spoken but not written language is understood, or it may mean that all modalities of comprehension are impaired. Thus, there are several aphasic syndromes, and it is the specific syndrome that generally must be identified in order to provide some correlation with the underlying localization of the brain damage and to make rational treatment plans. We may note that the standard comprehensive neuropsychological test batteries do not include extensive aphasia examinations. There are several such examinations available, such as the *Boston Diagnostic Aphasia Examination* (Goodglass & Kaplan, 1983) and the *Western Aphasia Battery* (Kertesz, 1979). Even though they may be used in conjunction with a neuropsychological assessment battery, they are rather lengthy procedures in themselves and require special expertise to administer and interpret.

For various reasons, it is often useful to assess attention as part of the neuropsychological examination. Sometimes, an attention deficit is a cardinal symptom of the disorder, but even if it isn't, the patient's level of attention may influence performance on tests of essentially all of the functional areas we have been discussing. A discussion of attention may be aided by invoking a distinction between *wide-aperture* and *narrow-aperture* attention (Kinsbourne, 1980). Wide-aperture attention has to do with the individual's capacity to attend to an array of stimuli at the same time. Attention may be so narrowly focused that the total picture is not appreciated. Tests for neglect may in fact be assessing wide aperture attention. narrow-aperture attention has to do with the capacity to sustain attention to small details. Thus, it can be assessed by vigilance tasks or tests like the Picture Completion subtest of the Wechsler scales. Brain-damaged patients may manifest attentional deficits of either type. They may fail to attend to a portion of their perceptual environment, or they may be unable to maintain sufficient concentration to successfully complete tasks requiring sustained occupation with details. Individuals with attentional deficits are often described as distractible or impul-

sive, and in fact, many brain-damaged patients may be accurately characterized by those terms. Thus, the assessment of presence and degree of attention deficit is often a highly clinically relevant activity. Recently, Mirsky and collaborators (1991) have proposed a useful division, based on a factor analytic study of attentional tasks, dividing them into tests that evaluate encoding, sustaining concentration, focusing, and shifting attention from one aspect of a task to another.

In summary, neuropsychological assessment typically involves the functional areas of general intellectual capacity, memory, speed and accuracy of psychomotor activity, visual-spatial skills, visual, auditory, and tactile perception, language, and attention. Thus, a comprehensive neuropsychological assessment may be defined as a procedure that at least surveys all of these areas. In practical terms, a survey is all that is feasible if the intent of the assessment is to evaluate all areas. It is obviously generally not feasible to do an in-depth assessment of each of these areas in every patient, nor is it usually necessary to do so.

SPECIAL PROBLEMS IN THE CONSTRUCTION AND STANDARDIZATION OF NEUROPSYCHOLOGICAL TEST BATTERIES

It will be assumed here that neuropsychological tests share the same standardization requirements as all psychological tests. That is, there is the need for appropriate quantification, norms, and related test-construction considerations, as well as the need to deal with issues related to validity and reliability. However, there are some special considerations regarding neuropsychological tests, and we will turn our attention to them here.

Practical Concerns in Test Construction

Neuropsychological test batteries must of necessity be administered to brain-damaged patients, many of whom may have severe physical disability, cognitive impairment, or a combination of the two. Thus, stimulus and response characteristics of the tests themselves, as well as the stimulus characteristics of the test instructions, become exceedingly important considerations. Neuropsychological test material should, in general, be constructed with salient stimuli that the patient can readily see or hear and understand. Material to be read should not require high levels of literacy, nor should grammatical structures be unduly complex. With regard to test instruction, the potential for multimodal instruction-giving should ideally be available. If the patient cannot see or read, it should be possible to say the instructions, without jeopardizing one's opportunity to use established test norms. The opportunity should be available to repeat and paraphrase instructions until it is clear that they are understood. It is of crucial importance in neuropsychological assessment that the examiner achieve maximum assurance that a test was failed because the patient could not perform the task being assessed, not because the test instructions were not understood. This consideration is of particular importance for the aphasic patient, who may have a profound impairment of language comprehension. With regard to response parameters, efforts should be made to assure that the test response modality is within the patient's repertoire.

In neuropsychological assessment, it is often not failure to perform some specific task that is diagnostic, but failure to perform some component of a series of tasks in the presence of intact function in other areas. As an example, failure to read a passage is not specifically diagnostic, since the inability to read my be associated with a variety of cognitive, perceptual, and learning difficulties. However, failure to be able to transfer a grapheme or a written symbol to a phoneme or sound in the presence of other manifestations of literacy could be quite diagnostic. Individuals with this type of deficit may be able to "sight-read" or recognize words as perceptual patterns, but when asked to read multisyllabic, unfamiliar words, they are unable to break the word down into phonemes and sound it out. In perhaps its most elegant form, neuropsychological assessment can produce what is called a double dissociation (Teuber, 1959); a task consistently failed by patients with a particular type of brain disorder accompanied by an equally difficult corresponding task that is consistently passed, and the reverse in the case of patients with some other form of brain disorder. ideally, then, neuropsychological assessment aims at detailed-as-possible specification of what functional deficits exist in a manner that allows for mapping of these deficits onto known systems in the brain. There are several methods of achieving this gaol, and not all neuropsychologists agree with regard to

the most productive route. In general, some prefer to examine patients in what may be described as a linear manner, with a series of interlocking component abilities, while others prefer using more complex tasks in the form of standard, extensive batteries and interpretation through examination of performance configurations. The linear approach is best exemplified in the work of A. R. Luria and various collaborators (Luria, 1973), while the configural approach is seen in the work of Ward Halstead (Halstead, 1947) Ralph Reitan (Reitan & Wolfson, 1993) and their many collaborators. In either case, however, the aim of the assessment is largely that of determining the pattern of the patient's preserved and impaired functions and inferring from this pattern what the nature might be of the disturbed brain function. The difficulty with using complex tasks to achieve that end is that such tasks are really only of neuropsychological interest if they can be analyzed by one of the two methods described here.

Issues Related to Validity and Reliability

Neuropsychological assessment has the advantage of being in an area where the potential for development of highly sophisticated validation criteria has been very much realized in recent years and will surely achieve even fuller realization in the near future. We will begin our discussion with this consideration, and so we will first be occupied with the matters of concurrent and predictive validity. A major review of validation studies was accomplished by Klove (1974) and updated by Boll (1981). A recent review was done by Reed and Reed (1997). Reitan and Wolfson (1993) have written an entire volume on the Halstead-Reitan battery (HRB) which contains a brief review of pertinent research findings in addition to extensive descriptions of the tests themselves and case materials. These reviews essentially only covered the Wechsler scales and the HRB, but there are several reviews of the work with the Luria-Nebraska Neuropsychological Battery as well (e.g., Moses & Purisch, 1997).

We will not deal with the content of those reviews at this point, but rather focus on the methodological problems involved in establishing concurrent or predictive validity of neuropsychological tests. With regards to concurrent validity, the criterion used in most cases is the objective identification of some central-nervous-system lesion arrived at independently of the neuropsychological test results. Therefore, validation is generally provided by neurologists or neurosurgeons. Identification of lesions of the brain is particularly problematic because, unlike many organs of the body, the brain cannot usually be visualized directly in the living individual. The major exceptions occur when the patient undergoes brain surgery or receives the rarely used procedure of brain biopsy. In the absence of the procedures, validation is dependent upon autopsy data or the various brain-imaging techniques. Autopsy data are not always entirely usable for validation purposes, in that numerous changes may have taken place in the patient's brain between time of testing and time of examination of the brain. Of the various imaging techniques, magnetic resonance imaging (MRI) is currently the most fruitful one. Cooperative among neuroraudiologists, neurologists, and neuropsychologists has already led to the accomplishment of several important studies correlating quantitative magnetic-resonance data with neuropsychological-test results (e.g., Minshew, Goldstein, Dombrowski, Panchaligam, & Pettegrew, 1993). Beyond MRI, however, we can see the beginnings of even more sensitive indicators, including measures of cerebral metabolism such as the PET scan (Positron Emission Tomography), and functional MRI. Recently, more generally available and even more sensitive measures of cerebral metabolism have appeared, including more recent generations of the PET scan, allowing for greatly improved resolution, SPECT (Single Photon Emission Computerized Tomography), which allows for studying brain metabolism in settings in which a cyclotron is not available, and the evolving methods of magnetic-resonance spectroscopy. These exciting new developments in brain imaging and observation of brain function will surely provide increasingly definitive criteria for neuropsychological hypothesis and assessment methods.

Within neuropsychological assessment, there has been a progression regarding the relationship between level of inference and criterion. Early studies in the filed as well as the development of new assessment batteries generally addressed themselves to the matter of simple presence or absence of structural brain damage. Thus, the first question raised had to do with the accuracy with which an assessment procedure could discriminate between brain-damaged and non-brain-damaged patients, as independently classified by the criterion procedure. In the early studies, the criterion utilized was generally clinical diagnosis, perhaps

supported in some cases by neurosurgical data or some laboratory procedure such as a skull X-ray or an EEG. It soon became apparent, however, that many neuropsychological tests were performed at abnormal levels, not only by brain-damaged patients, but by patients with several of the functional psychiatric disorders. Since many neuropsychologists worked in neuropsychiatric rather than general medical settings, this matter became particularly problematic. Great efforts were then made to find tests that could discriminate between brain-damaged and psychiatric patients or, as sometimes put, between "functional" and "organic" conditions. There have been several early reviews of this research, (Goldstein, 1978; Heaton, Baade, & Johnson, 1978; Heaton & Crowley, 1981; Malec, 1978), all of which were critical of the early work in this field in light of current knowledge about several of the functional psychiatric disorders. The chronic schizophrenic patient was particularly problematic, since such patients often performed on neuropsychological tests in a manner indistinguishable from the performance of patients with generalized structural brain damage. By now, this whole issue has been largely reformulated in terms of looking at the neuropsychological aspects of many of the functional psychiatric disorders (e.g., Goldstein, 1991; Henn & Nasrallah, 1982), largely under the influence of the newer biological approaches to psychopathology.

Neuropsychologists working in neurological and neurosurgical settings were becoming increasingly interested in validating their procedures against more refined criteria, notably in the direction of localization of brain function. The question was no longer only whether a lesion was present or absent, but if present, whether or not the tests could predict its location. Major basic research regarding this matter was conducted by H.-L. Teuber and various collaborators over a span of many years (Teuber, 1959). This group had access to a large number of veterans who had sustained open head injuries during World War II and the Korean conflict. Because the extent and site of their injuries were exceptionally well documented by neurosurgical and radiological data, and the lesions were reasonably well localized, these individuals were used productively in a long series of studies in which attempts were made to related both site of lesion and concomitant neurological defects to performance on an extensive series of neuropsychological procedures ranging from measures of basic sensory functions (Semmes, Weinstein, Ghent, & Teuber, 1960) to complex cognitive skills (Teuber & Weinstein, 1954). Similar work with brain-wounded individuals was accomplished by Freda Newcombe and collaborators at Oxford (Newcombe, 1969). These groups tended to concentrate on the major lobes of the brain (frontal, temporal, parietal, and occipital), and would, for example, do contrasts between the performances of patients with frontal and occipital lesions on some particular test or test series (e.g., Teuber, 1964). In another setting, but at about the same time as the Teuber group was beginning its work, Ward Halstead and collaborators conducted a large-scale neuropsychologically oriented study of frontal lobe function (Halstead, 1947). Ralph M. Reitan, who was Halstead's student, adopted several of his procedures, supplemented them, and developed a battery of tests that were extensively utilized in localization studies. Reitan's early work in the localization area was concerned with differences between the two cerebral hemispheres more than with regional localization (Reitan, 1955). The now well-known Wechsler-Bellevue studies of brain lesion lateralization (see review in Reitan, 1966) represented some of the beginnings of this work. The extensive work of Roger Sperry and various collaborators (Sperry, Gazzaniga, & Bogen, 1969) with patients who had undergone cerebral commisurotomy also contributed greatly to validation of neuropsychological tests with regard to the matter of differences between the two hemispheres; particularly the functional asymmetries or cognitive differences. Since the discoveries regarding the major roles of subcortical structures in the mediation of various behaviors (Cummings, 1990), neuropsychologists have also been studying the relationships between test performance and lesions in such structures and structure complexes as the limbic system (Scoville & Milner, 1957) and the basal ganglia (Butters, 1983).

The search for validity criteria has become increasingly precise with recent advances in the neurosciences as well as increasing opportunities to collect test data from various patient groups. One major conceptualization largely attributable to Reitan and his co-workers is that localization does not always operate independently with regard to determination of behavioral change, but interacts with type of lesion or the specific process that produced the brain damage. The first report regarding this matter related to differences in performance between patients with recently acquired lateralized brain damage and those who sustained lateralized

brain damage at some time in the remote past (Fitzhugh, Fitzhugh, & Reitan, 1961, 1962). Patients with acute lesions were found to perform differently on tests than patients with chronic lesions. It soon became apparent, through an extremely large number of studies (e.g., Goldstein, Nussbaum, & Beers, 1998) that there are many forms of type-locus interactions, and that level and pattern of performance on neuropsychological tests may vary greatly with the particular nature of the brain disorder. This development paralleled such advances in the neurosciences as the discovery of neurotransmitters and the relationship between neurochemical abnormalities and a number of the neurological disorders that historically had been of unknown etiology. We therefore have the beginnings of the development of certain neurochemical validating criteria (Davis, 1983; Freedman, 1990). There has also been increasing evidence for a genetic basis for several mental and neurological disorders. The gene for Huntington's disease has been discovered, and there is growing evidence for a significant genetic factor contributing to the acquisition of certain forms of alcoholism (Steinhauer, Hill, & Zubin, 1987). In general, the concurrent validity studies have been quite satisfactory, and many neuropsychological test procedures have been shown to be accurate indicators of many parameters of brain dysfunction.

A persistent problem in the past has been the possible tendency of neuropsychological tests to be more sensitive than the criterion measures. In fact, a study by Filskov and Goldstein (1974) demonstrated that neuropsychological tests may predict diagnosis more accurately than many of the individual neurodiagnostic procedures commonly used in assessment of neurological and neurosurgical patients (e.g., skull x-ray). It would appear that with the advent of the MRI scan and the even more advanced brain-imaging procedures this problem will be diminishing. A related problem involves the establishment of the most accurate and reliable external criterion. We have always taken the position (Goldstein & Shelly, 1982; Russell, Neuringer, & Goldstein, 1970) that no one method can be superior in all cases, and that the best criterion is generally the final medical opinion based on a comprehensive but pertinent evaluation, excluding, of course, behavioral data. In some cases, for example, the MRI scan may be relatively noncontributory, but there may be definitive laboratory findings based on examination of blood or cerebral spinal fluid. In some cases (e.g., Huntington's disease) the family history may be the most crucial part of the evaluation. It is not being maintained here that the best criterion is a doctor's opinion, but rather that no one method can stand out as superior in all cases when dealing with a variety of disorders. The diagnosis is often best established through the integration by an informed individual of data coming from a number of sources. A final problem to be mentioned here is that objective criteria do not yet exist for a number of neurological disorders, but even this problem appears to be undergoing a rapid stage of solution. Most notable in this regard is the *in vivo* differential diagnosis of the degenerative diseases of old age, such as Alzheimer's disease. There is also no objective laboratory marker for multiple sclerosis, and diagnosis of that disorder continues to be made on a clinical basis. Only advances in the neurosciences will lead to ultimate solutions to problems of this type.

In clinical neuropsychology, predictive validity has mainly to do with course of illness. Will the patient get better, stay the same, or deteriorate? Generally, the best way to answer questions of this type is through longitudinal studies, but very few such studies have actually been done. Even in the area of normal aging, in which many longitudinal studies have been accomplished, there really have been no extensive neuropsychologically oriented longitudinal studies. There is, however, some literature on recovery from stroke, much of which is attributable to the work of Meier and collaborators (Meier, 1974). Levin, Benton, and Grossman (1982) provide a discussion of recovery from closed head injury. Of course, it is generally not possible to do a full neuropsychological assessment immediately following closed head injury, and so prognostic instruments used at that time must be relatively simple ones. In this regard, a procedure known as the Glasgow Coma Scale (Teasdale & Jennett, 1974) has well-established predictive validity. Perhaps one of the most extensive efforts directed toward establishment of the predictive validity of neuropsychological tests was accomplished by Paul Satz and various collaborators, involving the prediction of reading achievement in grade school based on neuropsychological assessments accomplished during kindergarten (Fletcher & Satz, 1980; Satz, Taylor, Friel, & Fletcher, 1978). At the other end of the age spectrum, there are currently several ongoing longitudinal studies contrasting normal elderly individuals with dementia patients (Colsher & Wallace, 1991;

Evans et al., 1993). However, we do not yet know from these studies and other ongoing longitudinal investigations what the best prognostic instruments are for predicting the course of dementia or for determining whether or not an elderly individual suspected of having dementia will deteriorate or not.

An important aspect of predictive validity has to do with prediction of treatment and rehabilitation outcome. There have been several studies (reviewed by Allen, Goldstein, & Seaton, 1997) concerned with predicting outcome of alcoholism treatment on the basis of neuropsychological test performance. The results of these studies are mixed, but in general it would appear that test performance during the early stages of treatment may bear some relationship to outcome as evaluated by follow-up. Guilmette & Kastner (1996) reviewed prediction of vocational functioning from neuropsychological testing. before leaving this area, it should be mentioned that there are several not fully documented but apparently reasonable clinical principles related to prediction of treatment outcome. In general, patients with relatively well-circumscribed deficits and perhaps underlying structural lesions, tend to do better in treatment than do patients with more global deficits. There are some data that suggest that early intervention for aphasic adults, perhaps with two months post-onset in conjunction with spontaneous recovery, is more effective than treatment initiated later (Stein, 1988). Many years ago, Ben-Yishay, Diller, Gerstman, and Gordon (1970) reported that initial level of competence on a task to be trained is related to ability to profit from cues utilized in the training procedure.

In general, studies of predictive validity in neuropsychological assessment have not been as extensive as studies involving concurrent validity. However, the data available suggest that neuropsychological tests can predict degree of recovery or deterioration to some extent and have some capacity to predict treatment outcome. Since many neurological disorders change over time, getting better or worse, and the treatment of neurological disorders is becoming an increasingly active field (Zimmer & Grosberg, 1997), it is often important to have some foreknowledge of what will happen to the patient in the future in a specific rather than general way and to determine whether or not the patient is a good candidate for some form of treatment. Efforts have also been made to predict functional abilities involved in personal self-care and independent living on the basis of neuropsychological test performance, particularly in the case of elderly individuals (McCue, 1997). The extent to which neuropsychological assessment can provide this prognostic information will surely be associated with the degree of its acceptance in clinical settings.

Studies of the construct validity of neuropsychological tests represent a great amount of the corpus of basic clinical neuropsychological research. Neuropsychology abounds with constructs: short-term memory, attention, visual-spatial skills, psychomotor speed, motor engrams, and cell-assemblies. Tests are commonly characterized by the construct they purport to measure; Test A is a test of long-term memory; Test B is a test of attention; Test C is a test of abstraction ability; Test D is a measure of biological intelligence, etc. Sometimes we fail to recognize constructs as such because they are so well established, but concepts like memory, intelligence, and attention are in fact theoretical entities used to describe certain classes of observable behaviors. Within neuropsychology, the process of construct validation generally begins with an attempt to find a measure that evaluates some concept. Let us begin with a simple example, say the desire to develop a test for memory. Memory, as a neuropsychological construct, would involve a brain-behavior relationship. That is, neuropsychologists are concerned with how impaired brain function affects memory. There are memory tests available, notably the newly revised Wechsler Memory Scale (WMS-III) (Wechsler, 1997), but without experimental studies, that scale would only have face validity; that is, it appears to be a test of memory on the basis of the nature of the test items. However, if we ask the related questions, "Does the patient who does well on the scale have a normal memory?" We would have to know more about the test in regard to how well it assesses memory as a construct. Reasonable alternative hypotheses might be that the scale measures intelligence, educational level, or attention, or that these influences confound the test such that impairment of memory *per se* cannot be unequivocally identified.

The problem may be approached in numerous ways. A factor-analytic strategy may be used in which subtests of the Wechsler Memory Scale are placed into a factor analysis along with educational level and tests of intelligence and attention. It may be found that memory-scale subtests load on their own factor or on factors that receive high loadings

from the intelligence and attention tests or from educational level. Another approach may involve giving the test to patients with amnesia and to nonamnesic brain-damaged patients. A more sophisticated study may involve administering the Wechsler Memory Scale to these subjects along with other tests. Studies of these types may reveal some of the following hypothetical findings. The Wechsler Memory Scale is highly correlated with IQ, and so it is not possible to tell whether it measures the construct memory specifically or intellectual ability. Some patients cannot repeat stories read to them because they are aphasic and cannot produce words, not because of poor memories. Therefore, interpretation of the measure as an indicator of memory ability cannot be made unequivocally in certain populations. Certain amnesic patients do exceedingly poorly on certain components of the Wechsler Memory Scale, but well on other components. Such a finding would suggest that memory, as a neuropsychological construct, requires further refinement, since there appears to be a dissociation in patients known to have profound loss of memory between certain memory skills that are intact and others that are severely impaired. Still another approach, suggested by Cronbach (1960), is correlation with practical criteria. Individuals given the Wechsler Memory Scale could be asked to perform a number of tasks, all of which involve practical memory in some way, and the obtained data could be analyzed in terms of what parts of the scale predict success or failure at the various tasks.

Another important way of establishing the construct validity of neuropsychological test batteries involves determining capacity to classify cases into meaningful subtypes. In recent years, several groups of investigators have utilized classification statistics, notably R-type factor analysis and cluster analysis in order to determine whether combinations of test scores from particular batteries classify cases in accordance with established diagnostic categories or into types that are meaningful from the standpoint of neuropsychological considerations. A great deal of effort has gone into establishing meaningful, empirically derived subtypes of learning disability (Rourke, 1985), and there has also been work done in the neuropsychologically based empirical classification of neuropsychiatric patients (Goldstein, 1994; Schear, 1987).

It is particularly important to note that, at least in recent years, the construct validation of neuropsychological tests has involved a multidisciplinary effort with colleagues in cognitive psychology, the experimental psychology of memory and learning (utilizing both human studies and animal models), linguistics, and sensory and perceptual processes. For example, aphasia testing and other forms of language assessment have been profoundly influenced by research in neurolinguistics (Blumstein, 1981; Crary, Voeller, & Haak, 1988), while memory testing has been correspondingly influenced by recent developments in information theory and the experimental psychology of memory and learning (Baddeley, Wilson, & Watts, 1995; Butters & Cermak, 1980). These experimental foundations have aided significantly in the interpretation of clinical tests, and indeed, many new clinical tests are actually derived from laboratory procedures.

While neuropsychological tests should ideally have reliability levels commensurate with other areas of psychometrics, there are some relatively unique problems. These problems are particularly acute when the test-retest method is used to determine the reliability coefficients. The basic problem is that this method really assumes the stability of the subject over testing occasions. When reliability coefficients are established through the retesting of adults over a relatively brief time period, that assumption is a reasonable one, but it is not as reasonable in samples of brain-damaged patients who may be rapidly deteriorating or recovering. Indeed, it is generally thought to be an asset when a test reflects the appropriate changes. Another difficulty with the test-retest method is that many neuropsychological tests are not really repeatable because of substantial practice effects. The split-half method is seldom applicable, since most neuropsychological tests do not consist of lengthy lists of items, readily allowing for odd-even or other split-half comparisons. In the light of these difficulties, the admittedly small number of reliability studies done with the standard neuropsychological test batteries have yielded perhaps surprisingly good results. Boll (1981) has reviewed reliability studies done with the HRB, and Goldstein and Watson (1989) did a test-retest reliability study with several clinical groups. The test manual (Golden, Hammeke, & Purisch, 1980) reports reliability data for the Luria-Nebraska Battery. The details of these matters will be discussed later in our reviews of these two procedures. In any event, it seems safe to say that most neuropsychological test developers have not been greatly preoccupied with the reliabilities of their procedures, but those who have studied the matter appear to have provided sufficient data to permit the conclusion that the standard, commonly used

procedures are at least not so unreliable as to impair the validities of those procedures.

AN INTRODUCTION TO THE COMPREHENSIVE BATTERIES

The number of generally available comprehensive standard neuropsychological test batteries for adults is not entirely clear. The *Handbook of Clinical Neuropsychology* (Filskov & Boll, 1981) only contains chapters on two batteries; the Halstead-Reitan and Luria-Nebraska. Lezak (1995) lists the Halstead-Reitan, the Smith Neuropsychological Battery, and two versions of batteries dervived from Luria's work; one by Christensen (1975a, 1975b, 1975c) and Golden, Hammeke, and Purisch's Luria-Nebraska (originally South Dakota) Battery (1980). Jones and Butters (1983) reviewed the Halstead-Reitan, Lura-Nebraska, and Michigan batteries. Benton, Sivan, Hamsher, Varney and Spreen (1994) have produced a manual containing descriptions and instructions for tests these neuropsychologists have been associated with over the years, and the Spreen and Strauss (1998) have published a "compendium" of neuropsychological tests, but there was clearly no intention in either case to present these collections of tests as standard batteries. In this chapter, we will only consider the Halstead-Reitan and Luria-Nebraska procedures. The Michigan Battery (Smith, 1975) will not be reviewed, primarily because it consists largely of a series of standardized tests, all of which have their own validity and reliability literature.This literature is thoroughly reviewed by Lezak (1995).

The Halstead-Reitan Battery (HRB)

History

The history of this procedure and its founders has been reviewed by Reed (1983) and more recently by Reed and Reed (1997). These authors trace the beginnings of the battery to the special laboratory established by Halstead in 1935 for the study of neurosurgical patients. The first major report on the findings of this laboratory appeared in a book called *Brain and Intelligence: A Quantitative Study of the Frontal Lobes* (Halstead, 1947), the title of which suggests that the original intent of Halstead's tests was describing frontal lobe function. In this book, Halstead proposed his theory of "biological intelligence" and presented what was probably the first factor analysis done with neuropsychological test data. Perhaps more significantly, however, the book contains descriptions of many of the tests now contained in the HRB. As Reed (1983) suggests, the theory of biological intelligence never was widely accepted among neuropsychologists, and the factor analysis had its mathematical problems. But several of the tests that went into that analysis survived, and many of them are commonly used at present. In historical perspective, Halstead's major contributions to neuropsychological assessment, in addition to his very useful tests, include the concept of the neuropsyhoclogical laboratory in which objective tests are administered in standard fashions and quantitatively scored, and the concept of the impairment index, a global rating of severity of impairment and probability of the presence of structural brain damage.

Ralph M. Reitan was a student of Halstead at Chicago and was strongly influenced by Halstead's theories and methods. Reitan adopted the methods in the form of the various test procedures and with them established a laboratory at the University of Indiana. He supplemented these tests with a number of additional procedures in order to obtain greater comprehensiveness and initiated a clinical research program that is ongoing. The program began with a cross-validation of the battery and went on into numerous areas, including validation of new tests added to the battery (e.g., the Trail Making test), lateralization and localization of function, aging, and neuropsychological aspects of a wide variety of disorders such as alcoholism, hypertension, disorders of children, and mental retardation. Theoretical matters were also considered. Some of the major contributions included the concept of type-locus interaction (Reitan, 1966), the analysis of quantitative as opposed to qualitative deficits associated with brain dysfunction (Reitan, 1958, 1959), the concept of the brain-age quotient (Reitan, 1973), and the scheme for levels and types of inference in interpretation of neuropsychological test data (Reitan & Wolfson, 1993). In addition to the published research, Reitan and his collaborators developed a highly sophisticated method of blind clinical interpretation of the HRB that continues to be taught at workshops conducted by Dr. Reitan and associates. The HRB, as the procedure came to be known over the years, also has a

history. It has been described as a "fixed battery", but that is not actually the case. Lezak (1976) says in reference to this development, "This set of tests has grown by accretion and revision and continues to be revised" (p. 440). Halstead's original battery, upon which the factor analyses were based, included the Carl Hollow Square test, the Dynamic Visual Field Test, the Henmon-Nelson tests of mental ability, a flicker fusion procedure, and the Time Sense test. None of these procedures are now widely used, although the Time Sense and Flicker Fusion tests were originally included in the battery used by Reitan. The tests that survived included the Category test, the Tactual Performance test, the Speech Perception test, and Finger Tapping. Halstead also used the Seashore Rhythm test, which is included in the current version of the battery, but was not included in subbattery used by Halstead in his factor analyses. There have been numerous additions, including the various Wechsler Intelligence Scales, the Trail Making test, a subbattery of perceptual tests, the Reitan Aphasia Screening test, the Klove Grooved Pegboard, and other tests that are used in some laboratories but not in others. Alternative methods have also been developed for computing impairment indexes. (Reitan, 1991; Russell, Neuringer, & Goldstein, 1970).

Bringing this brief history into the present, the HRB continues to be widely used as a clinical and research procedure. Numerous investigators utilize it in their research, and there have been several successful cross-validations done in settings other than Reitan's laboratory (Goldstein & Shelly, 1972; Vega & Parson, 1967). In addition to the continuation of factor-analytic work with the battery, several investigators have applied other forms of multivariate analysis to it in various research applications. Several investigators have applied other forms of multivariate analysis to it in various research applications. Several investigations have been conducted relative to objectifying and even computerizing interpretation of the battery, the most well-known efforts probably being the Selz Reitan rules for classification of brain function in older children (Selz & Reitan, 1979) and the Russell, Neuringer and Goldstein "Neuropsychological Keys" (Russell, et al., 1970). The issue of reliability of the battery has recently been addressed, with reasonably successful results. Clinical interpretation of the battery continues to be taught at workshops and in numerous programs engaged in training of professional psychologists. The most detailed description of the battery available will be found in Reitan and Wolfson (1993).

Since the publication of the earlier editions of this chapter, much work has been done on the psychometrics of the HRB. We now have available a manual that provides normative information for adults including corrections for age, education, and gender (Heaton, Grant, & Matthews, 1991), and a number of elegant scoring and rating systems, reviewed in Russell (1997). These include the Neuropsychological Deficit Scale developed by Reitan himself (1987; 1991), the Halstead-Russell Neuropsychological Evaluation System (HRNES) (Russell, 1993), and the Comprehensive Norms for an Extended halstead-Reitan Battery (CNEHRB) system presented in the Heaton, Grant, & Matthews manual. These systems are all devoted to scaling of the HRB and providing new summary and index scores that are clinically useful. It may be noted that the CNEHRB system has stimulated some controversy revolving around the matter of whether it is appropriate to correct neuropsychological test scores for age and education (Reitan & Wolfson, 1995).

Structure and Content

Although there are several versions of the HRB, the differences tend to be minor, and there appears to be a core set of procedures that essentially all versions of the battery contain. The battery must be administered in laboratory containing a number of items of equipment and generally cannot be completely administrated at bedside. Various estimates of length of administration are given, but it is probably best to plan on about six to eight hours of patient time. Each test of the battery is independent and may be administered separately from the other tests. However, it is generally assumed that a certain number of the tests must be administered in order to compute an impairment index.

Scoring for the HRB varies with the particular test, such that individual scores may be expressed in time to completion, errors, number correct, or some form of derived score. For research purposes, these scores are generally converted to standard scores so that they may be profiled. Matthews (1981) routinely uses a T-score profile in clinical practice, while Russell and colleagues (1970) rate all of the tests contributing to the impairment index on a six-point scale, the data being displayed as a profile of the ratings. In their system the impair-

ment index may be computed by calculating the proportion of tests performed in the brain-damaged range according to published cut-off scores (Reitan, 1955) or by calculating the average of the ratings. This latter procedure provides a value called the Average Impairment Rating. Russell and coworkers (1970) have also provided quantitative scoring systems for the Reitan Aphasia Screening test and for the drawing of a Greek cross that is part of that test. However, some clinicians do not quantify those procedures, except in the form of counting the number of aphasic symptoms elicited. As indicated above, there is the recent development of a number of new indices and scoring systems. The General Neuropsychological Deficit Scale (GNDS) (Reitan & Wolfson, 1993) provides a substantial extension of the Impairment Index and Average Impairment Rating. The system utilizes 42 variables, and is based on four methods of inference, level of performance, pathognomonic signs, pattern, and right-left differences. It provides both a global score and scores for each inference method category. We will return to other aspects of the battery's structure after the following description of the component tests.

A. Halstead's Biological Intelligence Tests

1. The Halstead Category Test: This test is a concept-identification procedure in which the subject must discover the concept or principle that governs various series of geometric forms, verbal and numerical material. The apparatus for the test includes a display screen with four horizontally arranged numbered switches placed beneath it. The stimuli are on slides, and the examiner uses a control console to administer the procedure. The subject is asked to press the switch that the picture reminds him or her of, and is provided with additional instructions to the effect that the point of the test is to see how well he or she can learn the concept, idea, or principle that connects the pictures. If the correct switch is pressed, the subject will hear a pleasant chime, while wrong answers are associated with a rasping buzzer. The conventionally used score is the total number of errors for the seven groups of stimuli that form the tests. Booklet (Adams & Trenton, 1981; DeFillippis, McCampbell, and Roger, 1979) and abbreviated (Calsyn, O' Leary, & Chaney, 1980; Russell & Levy, 1987; Sherrill, 1987) forms of this test have been developed.

2. The Halstead Tactual Performance Test: This procedure utilizes a version of the Sequin-Goddard Formboard, but it is done blindfolded. The subject's task is to place all of the 10 blocks into the board, using only the sense of touch. The task is repeated three times, once with the preferred hand, once with the nonpreferred hand and once with both hands, following which the board is removed. After removing the blindfold, the subject is asked to draw a picture of the board, filling in all of the blocks he or she remembers in their proper locations on the board. Scores from this test include time to complete the task for each of the three trials, total time, number of blocks correctly drawn and number of blocks correctly drawn in their proper locations on the board.

3. The Speech Perception Test: The subject is asked to listen to a series of 60 sounds, each of which consist of a double e digraph with varying prefixes and suffixes (e.g., geend). The test is given in a four-alternative multiple-choice format, the task being that of underlining on an answer sheet the sound heard. The score is number of errors.

4. The Seashore Rhythm Test: This test consists of 30 pairs of rhythmic patterns. The task is to judge whether the two members of each pair are the same as or different from each other and to record the response by writing an S or a D on an answer sheet. The score is either number correct or number of errors.

5. Finger Tapping: The subject is asked to tap his or her extended index finger on a typewriter key attached to a mechanical counter. Several series of 10-second trials are run, with both the right and left hand. The scores are average number of taps generally over five trials, for the right and left hand.

B. Tests Added to the Battery by Reitan

1. The Wechsler Intelligence Scales: Some clinicians continue to use the Wechsler-Bellevue, some the WAIS, and some the WAIS-R. In any event, the test is given according to manual instructions and is not modified in any way. It is too soon to determine how many HRB users will switch over to the recently published WAIS III (Wechsler, 1997).

2. The Trail Making Test: In Part A of this procedure the subject must connect in order a series of circled numbers randomly scattered over a sheet of 8 1/2 × 11 paper. In Part B, there are circled numbers and letters, and the subject's task involves alternating between numbers and letters in serial order (e.g., 1 to A to 2 to B, etc.). The score is time to completion expressed in seconds for each part.

3. The Reitan Aphasia Screening Test: This test serves two purposes in that it contains both copying and language- related tasks. As an aphasia-screening procedure, it provides a brief survey of the major language functions: naming, repetition, spelling, reading, writing, calculation, narrative speech, and right-left orientation.

The copying tasks involve having the subject copy a square, Greek cross, triangle, and key. The first three items must each be drawn in one continuous line. The language section may be scored by listing the number of aphasic symptoms or by using the Russell and colleagues' quantitative system. The drawings are not formally scored or are rated through a matching to model system also provided by Russell and colleagues (1970).

4. Perceptual Disorders: These procedures actually constitute a sub-battery and include tests of the subject's ability to recognize shapes and identify numbers written on the fingertips, as well as tests of fingers discrimination and visual, auditory, and tactile neglect. Number of errors is the score for all of these procedures.

C. Tests Added to the Battery by Others

1. The Klove Grooved Pegboard Test: The subject must place pegs shaped like keys into a board containing recesses that are oriented in randomly varying directions. The test is administered twice, once with the right and once with the left hand. Scores are time to completion in seconds for each hand and errors for each hand defined as number of pegs dropped during performance of the task.

2. The Klove Roughness Discrimination Test: The subject must order four blocks covered with varying grades of sandpaper presented behind a blind with regard to degree of roughness. Time and error scores are record for each hand.

3. Visual Field Examination: Russell and colleagues (1970) include a formal visual field examination utilizing a perimeter as part of their assessment procedure.

It should be noted that many clinicians, including Reitan and his collarborators, frequently administer a number of additional tests mainly for purposes of assessing personality and level of academic achievement. The MMPI is the major personality assessment method used, and achievement may be assessed with such procedures as the Wide Range Achievement Test-R (Jastak & Wilkinson, 1984) or the Peabody Individual Achievement Test (Dunn & Markwardt, 1970). Some clinicians have also adopted the procedure of adding the Wechsler Memory Scale (WMS, WMS-R, WMS-III) to the battery, either in its original form (Wechsler, 1945, 1987, 1997) or the Russell modification (Russell, 1975a). Some form of lateral dominance examination is also generally administered, including tests for handedness, footedness, and eyedness.

D. The "Expanded Halstead-Reitan Battery"

The Heaton, Grant, & Matthews manual (1991) contains a number of additional tests that are not a part of the original HRB. They include the Wisconsin Card Sorting Test (Heaton, 1980), the Story Memory Test (Reitan, unpublished test), the Figure Memory Test (based on the Wechsler Memory Scale; Wechsler, 1945), the Seashore Tonal Memory Test (Seashore, Lewis, & Saetveit, 1960), the Digit Vigilance Test (Lewis & Rennick, 1979), the Peabody Individual Achievement Test (Dunn & Markwardt, 1970), and the Boston naming Test (Kaplan, Goodglass, & Weintraub, 1983).

Theoretical Foundations

There are really two theoretical bases for the HRB, one contained in *Brain and Intelligence* and related writings of Halstead, the other in numerous papers and chapters written by Reitan and various collaborators (e.g., Reitan, 1966; Reitan & Wolfson, 1993). They are quite different from each other in many ways, and the difference may be partly accounted for by the fact that Halstead was not primarily a practicing clinician and was not particularly interested in developing his tests as psychometric instruments to be used in clinical

assessment of patients. Indeed, he never published the tests. He was more interested in utilizing the tests to answer basic scientific questions in the area of brain-behavior relationships in general and frontal lobe function in particular. Reitan's program, on the other hand, can be conceptualized as an effort to demonstrate the usefulness and accuracy of Halstead's tests and related procedures in clinical assessment of brain-damaged patients. It is probably fair to say that Halstead's theory of biological intelligence and its factor-analytically based four components (the central integrative field, abstraction, power, and the directional factor), as well as his empirical findings concerning human frontal lobe function have not become major forces in modern clinical neuropsychology. However, they have had, in my view, a more subtle influence on the field.

Halstead was really the first to establish a human neuropsychology laboratory in which patients were administered objective tests, some of which were semiautomated, utilized standard procedures and sets of instructions. His Chicago laboratory may have been the initial stimulus for the now common practice of trained technician administration of neuropsychological tests. Halstead was also the first to utilize sophisticated, multivariate statistics in the analysis of neuropsychological test data. Even though Reitan did not pursue that course to any great extent, other researchers with the HRB have done so (e.g., Goldstein & Shelly, 1971, 1972). Thus though the specifics of Halstead's theoretical work have not become well-known and widely applied, the concept of a standard neuropsychological battery administered under laboratory conditions and consisting of objective, quantifiable procedures has made a major impact on the field of clinical neuropsychology. The other, perhaps more philosophical contribution of Halstead was what might be described as his Darwinian approach to neuropsychology. He viewed his discriminating tests as measures of adaptive abilities, as skills that assured man's survival on the planet. Many neuropsychologists are now greatly concerned with the relevance of their test procedures to adaptation—the capacity to carry on functional activities of daily living and to live independently (Sbordone & Long, 1996). This general philosophy is somewhat different from the more traditional models emanating from behavioral neurology, in which there is a much greater emphasis on the more medical-pathological implications of behavioral-test findings.

Reitan, while always sympathetic with Halstead's thinking, never developed a theoretical system in the form of a brain model or a general theory of the biological intelligence type. One could say that Reitan's great concern has always been with the empirical validity of test procedures. Such validity can be established through the collection of large amounts of data obtained from patients with reasonably complete documentation of their medical/neurological conditions. Both presence and absence of brain damage had to be well documented, and if present, findings related to site and type of lesion had to be established. He has described his work informally as one large experiment, necessitating maximal consistency in the procedures used, and to some extent, the methods of analyzing the data. Reitan and his various collaborators represent the group that was primarily responsible for introduction of the standard battery approach to clinical neuropsychology. It is clear from reviewing the Reitan group's work that there is substantial emphasis on performing controlled studies with samples sufficiently large to allow for application of conventional statistical procedures. One also gets the impression of an ongoing program in which initial findings are qualified and refined through subsequent studies.

It would probably be fair to say that the major thrust of Reitan's research and writings has not been espousal of some particular theory of brain function, but rather an extended examination of the inferences that can be made from behavioral indices relative to the condition to the brain. There is a great emphasis on methods of drawing such inferences in the case of the individual patient. Thus, this group's work has always involved empirical research and clinical interpretation, with one feeding into the other. In this regard, there has been a formulation of inferential methods used in neuropsychology (Reitan & Wolfson, 1993) that provides a framework for clinical interpretation. Four methods are outlined: level of performance, pattern of performance, specific behavioral deficits (pathognomonic signs) and right-left comparisons. In other words, one examines for whether or not the patient's general level of adaptive function compares with that of normal individuals, whether there is some characteristic performance profile that suggests impairment even though the average score may be within normal limits, whether there are unequivocal individual signs of deficits, and whether there is a marked discrepancy in functioning between the two sides of the body. As indicated

above, these methods have recently been operationalized in the form of the General Neuropsychological Deficit Scale.

Reitan's theoretical framework is basically empirical, objective, and data-oriented. An extensive research program, by now of about 40 years' duration, has provided the information needed to make increasingly sophisticated inferences from neuropsychological tests. It thereby constitutes to a significant extent the basis for clinical interpretation. The part of the system that remains subjective is the interpretation itself, but in that regard Reitan (1964) has made the following remark: "Additional statistical methods may be appropriate for this problem but, in any case, progress is urgently needed to replace the subjective decision-making processes in individual interpretation that presently are necessary" (p. 46).

Standardization Research

The HRB, as a whole, meets rigorous validity requirements. Following Halstead's initial validation (1947) it was cross-validated by Reitan (1955) and in several other laboratories (Russell et al., 1970; Vega & Parsons, 1967). As indicated above, reviews of validity studies with the HRB have been written over the years by several authors. Validity, in this sense, means that all component tests of the battery that contribute to the impairment index discriminate at levels satisfactory for producing usable cutoff scores for distinguishing between brain-damaged and non-brain-damaged patients. The major exceptions, the Time Sense and Flicker Fusion tests, have been dropped from the battery by most of its users. In general, the validation criteria for these studies consisted of neurosurgical and other definitive neurological data. It may be mentioned, however, that most of these studies were accomplished before the advent of the CT and MRI scans, and it would probably now be possible to do more sophisticated validity studies, perhaps through correlating extent of impairment with quantitative measures of brain damage (e.g., CT-scan density measures). In addition to what was done with Halstead's tests, validity studies were accomplished with tests added to the battery such as the Wechsler scales, the Trail Making test, and the Reitan Aphasia Screening tests, with generally satisfactory results (Reitan, 1966).

By virture of the level of interferences made by clinician from HRB data, validity studies must obviously go beyond the question of presence or absence of brain damage. The first issue raised related to discriminative validity between patients with left hemisphere and right hemisphere brain damage. Such measures as Finger Tapping, the Tactual Performance test, the perceptual disorders subbattery, and the Reitan Aphasia Screening test all were reported as having adequate discriminative validity in this regard. There have been very few studies, however, that go further and provide validity data related to more specific criteria such as localization and type of lesion. It would appear from one impressive study (Reitan, 1964) that valid inferences concerning prediction at this level must be made clinically, and one cannot call upon the standard univariate statistical procedures to make the necessary discriminations. This study provided the major impetus for Russell and collaborators' (1970) neuropsychological key approach, which was in essence an attempt to objectify higher order inferences.

There is one general area in which the discriminative validity of the HRB was not thought in the past to be particularly robust. The battery does not have great capacity to discriminate between brain-damaged patients and patients with functional psychiatric disorders; notably chronic schizophrenia. There is an extensive literature concerning this matter, but it should be said that some of the research contained in this literature has significant methodological flaws, leaving the findings ambiguous. It may also be pointed out that the constructors of the HRB did not have the intention of developing a procedure to discriminate between brain-damaged and schizophrenic patients, and the assumption that it should be able to do so is somewhat gratuitous. Futhermore, Heaton and Crowley (1981) find that with the exception of the diagnosis of chronic schizophrenia, the HRB does a reasonably good job of differential diagnosis. They provided the following conclusion:

> The bulk of the evidence...suggests that for most psychiatric patient groups there is little or no relationship between the degree of emotional disturbance and level of performance on neuropsychological tests. However, significant correlations of this type are sometimes found with schizophrenic groups. (p. 492)

This matter remains controversial and has become exceedingly complex, particularly since the discovery of cerebral atrophy in a substantial portion of the schizophrenic population and the

development of hypotheses concerning left hemisphere dysfunction in schizophrenics (Gruzelier, 1991). The point to be made here is that the user of the HRB should exercise caution in interpretation when asked to use the battery in resolving questions related to differential diagnosis between brain damage and schizophrenia. Some writers have advised the addition of some measure of psychiatric disability, such as the MMPI, when doing such assessments (Russell, 1975b, 1977).

Even though there have been several studies of the predictive validity of neuropsychological tests with children (Fletcher & Satz, 1980; Lyon & Flynn, 1991) and other studies with adults that did not utilize the full HRB (Meier, 1974), I know of no major formal assessment of the predictive validity of the HRB accomplished with adults. Within neuropsychology, predictive validity has two aspects: predicting everyday academic, vocational, and social functioning and predicting course of illness. With regard to the former matter, Heaton and Pendleton (1981) document the lack of predictive validity studies using extensive batteries of the HRB type. However, they do report one study (Newman, Heaton, & Lehman, 1978) in which the HRB successfully predicted employment status on six-month follow-up. With regard to prediction of course of illness, there appears to be a good deal of clinical expertise in this regard, but no major formal studies in which the battery's capacity to predict whether the patient will get better, worse, or stay the same are evaluated. This matter is of particular significance in such conditions as head injury and stroke, since outcome tends to be quite variable in these conditions. The changes that occur during those stages are often the most significant ones related to prognosis (e.g., length of time unconscious).

In general, there has not been a great deal of emphasis on studies involving the reliability of the HRB, probably because of the nature of the tests themselves, particularly with regard to the practice-effect problem, and because of the changing nature of those patients for whom the battery was developed. Those reliability studies that were done produced satisfactory results, particularly with regard to the reliability to the impairment index (Boll, 1981). The Category test can have its reliability assessed through the split-half method. In a study accomplished by Shaw (1966), a .98 reliability coefficient was obtained.

Norms for the HRB are available in numerous places (Heaton, Grant, & Matthews, 1991; Reitan & Wolfson, 1993; Russell et al., 1970; Russell, 1993; Vega & Parson, 1967), but since the battery was never published as a single procedure, there is no published manual that one can refer to for definitive information. Schear (1984) has published a table of age norms for neuropsychiatric patients. Several laboratories have collected local norms. A great deal is known about the influence of age, education, and gender on the various tests in the HRB, and this information has only recently been consolidated in the Heaton, Grant, & Matthews (1991) manual. It is somewhat unusual for a procedure in as widespread use as the HRB not to have a commercially published manual. However, detailed descriptions of the procedures as well as instructions for administration and scoring are available in several sources including Reitan and Wolfson (1993), Jarvis and Barth (1984) and Swiercinsky (1978).

In summary, the validity of the HRB seems well-established by literally hundreds of studies, including several major cross-validations. These studies have implications for the concurrent, predictive, and construct validity of the battery. Reliability has not received nearly as much attention, but it seems apparent that the battery is sufficiently reliable to not compromise its validity. Recently, extensive age, gender, and education norms have become available (Heaton, Grant, & Matthews, 1991; Russell et al., 1970; Russell, 1993), but the relevance of such norms to neuropsychological assessment, particularly with regard to age, is a controversial and unsettled matter (Reitan & Wolfson, 1995). There is no commercially available manual for the battery, and so the usual kinds of information generally contained in a manual are not available to the test user in a single place. However, the relevant information is available in a number of separate sources.

Evaluation

The HRB is without doubt the most widely used standard neuropsychological battery, at least in North America and perhaps throughout the world. Aside from its widespread clinical application, it is used in many multidisciplinary research programs as the procedure of choice for neuropsychological assessment. It therefore has taken on something of a definitive status and is viewed by many experts in the field as the state-of-the-art instrument for comprehensive neuropsychological assessment.

Nevertheless, several criticisms of it have emerged over the years, and some of them will be reviewed here. Each major criticism will be itemized and discussed.

1. The HRB is too long and redundant. The implication of this criticism is that pertinent, clinically relevant neuropsychological assessment can be accomplished in substantially less time than the six to eight hours generally required to administer the full HRB. Other batteries are, in fact, substantially briefer than the HRB. Aside from simple giving fewer or briefer tests, another means suggested of shortening neuropsychological assessment is through a targeted, individualized approach rather than through routine administration of a complete battery. The difficulty with this latter alternative is that such an approach can generally only be conducted by an experienced clinician, and one sacrifices the clinician time and expense that can be saved through administration by trained technicians. The response to the criticism concerning length is generally that shortening of the battery correspondingly reduces its comprehensiveness, and one sacrifices examination of areas that may be of crucial significance in individual cases. Indeed, the battery approach was, in part, a reaction to the naiveté inherent in the use of single tests for "brain damage." The extent to which the clinician reverts to a single-test approach may reflect the extent to which there is a return to the simplistic thinking of the past. In general, the argument is that to adequately cover what must be covered in a standard, comprehensive assessment, the length of the procedure is a necessity. From the point of view of patient comfort and fatigue, the battery can be administered in several sessions over a period of days if necessary.

2. The tests in the HRB are insufficiently specific, both in regard to the functions they assess and the underlying cerebral correlates of those functions. Most of the tests in the battery are quite complex, and it is often difficult to isolate the source of impairment within the context of a single test. Even as apparently simple a procedure as the Speech Perception test requires not only the ability to discriminate sounds, but to read, make the appropriate written response, and attend to the task. Therefore, failure on the test cannot unequivocally point to a specific difficulty with auditory discrimination. Difficulties of this type are even more pronounced in such highly complex procedures as the Category and Tactual Performance tests. This criticism eventuates in the conclusion that it is difficult to say anything meaningful about the patient's brain or about treatment because one cannot isolate the specific deficit. In Luria's (1973) terminology one cannot isolate the functional system that is involved, no less the link in that system that is impaired. Failure to do so makes it difficult if not impossible to identify the structures in the brain that are involved in the patient's impairment as well as to formulate a rehabilitation program, since one doesn't really know in sufficiently specific terms what the patient can and cannot do.

This criticism ideally requires a very detailed response, since it implies a substantially different approach to neuropsychological assessment from the one adopted by developers of the HRB. Perhaps the response can be summarized in a few points. The HRB is founded on empirical rather than on content validity. Inferences are drawn on the basis of pertinent research findings and clinical observations rather than on the basis of what the tests appear to be measuring. The fact that one cannot partial out the various factors involved in successful or impaired performance on the Category test, for example, does not detract from the significant empirical findings related to this test based on studies of various clinical populations. In any event, Reitan, Hom, and Wolfson (1988) have shown that complex abilities, notably abstraction, are dependent upon the functioning of both cerebral hemispheres, and not on a localized unilateral system. The use of highly specific items in order to identify a specific system or system link is a procedure that is closely tied to the syndrome approach of behavioral neurology. Developers of the HRB typically do not employ a syndrome approach for several reasons. First, it depends almost exclusively on the pathognomonic-signs method of inference to the neglect of other inferential methods, and second, the grouping together of specific deficits into a syndrome is felt to be more often in the brain of the examiner than of the patient. The lack of empirical validity of the so-called Gerstmann Syndrome is an example of this deficiency in this particular approach (Benton, 1961). Another major point is that the HRB is a series of tests in which interpretation is based not on isolated consideration of each test taken one at a time, but on relationships among performances on all of the tests. Therefore, specific deficits can be isolated, in some cases at least, through intertest comparisons rather than through isolated examination of a single test.

Returning to our example, the hypothesis that there is impairment on the Speech Perception test because of failure to read the items accurately can be evaluated through looking at the results of the aphasia screening or reading-achievement test given. Finally, complex tests are likely to have more ecological validity than simple tests of isolated abilities. Thus, the Category test or Tactual Performance test results can tell the clinician more about real-world functioning than can the simpler tests. Simple tests were developed in the context of neurological diagnosis, while the tests in the HRB seem more oriented to assessing adaptive functioning in the environment.

3. The HRB is not sufficiently comprehensive, particularly in that it completely neglects the area of memory. The absence of formal memory testing in this battery has been noted by many observers and appears to be a valid criticism. On the face of it, it would appear that the battery would be incapable of identifying and providing meaningful assessments of patients with pure amnesic syndromes (e.g., patients with Korsakoff's syndrome). The absence of formal memory testing as part of the HRB is something of a puzzlement; although memory is involved in many of the tests, it is difficult to isolate the memory component as a source of impairment. Such isolation is readily achieved through such standard, commonly available procedures as list or paired-associate learning.

We know of no formal response to this criticism, but the point of view could be taken that pure amnesic syndromes are relatively rare,. and the HRB would probably not be the assessment method of choice for many of the rarely occurring specific syndromes. I would view this response as weak in view of the recently reported significance of memory defect in a number of disorders (Baddeley, Wilson, & Watts, 1995). Apparently, Halstead did not work with patients of those types, particularly patients with Alzheimer's and Huntington's disease, and so may have failed to note the significance of memory function in those disorders. However, this criticism is probably the one most easily resolved, since all that is required is addition of some formal memory testing to the battery. Such tests are included in the CNEHRB and the HRNES.

4. The HRB cannot discriminate between brain-damaged and schizophrenic patients. This matter has already been discussed, and most of the evidence (Heaton & Crowley, 1981) indicates that the performance of chronic schizophrenics on the HRB may be indistinguishable from that of the patient with generalized, structural brain damage. There are essentially two classes of response to this criticism. First, there is a disclaimer that the HRB was never designed for this kind of differential diagnosis, and so it is not surprising that it fails when it is inappropriately used for that purpose. Second, and perhaps much more significantly, is the finding that many schizophrenics have brain atrophy as assessed by CT and MRI scans, and tests of the HRB type can now be viewed as accurately identifying the behavioral correlates of that condition (Marsh, Lauriello, Sullivan, & Pfefferbaum, 1996). Futhermore, there are now several studies that indicate that schizophrenia is a neuropsychologically heterogeneous condition, and that there is a lack of relationship between neuropsychological test results and psychiatric diagnosis in the case of several psychiatric disorders (Goldstein, 1994; Townes et al., 1985).

5. Findings reported from Reitan's laboratory cannot be replicated in other settings. Here we have particular references to the early criticisms raised by Smith of Reitan's early Wechsler-Bellevue laterality studies. In a series of papers, Smith (1965, 1966a, 1966b) presented empirical and theoretical arguments against the reported finding that patients with left hemisphere lesions have lower verbal than performance IQs on the Wechsler-Bellevue, while the reverse was true for patients with right hemisphere brain damage. Smith was unable to replicate these findings in patients with lateralized brain damage that he had Wechsler-Bellevue data available on and also presented theoretical arguments against the diagnostic and conceptual significance of this finding. Klove (1974) analyzed the Smith versus Reitan findings in terms of possible age and neurological differences between the studies. Reviewing the research done to the time of writing, he also concluded that most of the research, with Smith as the only pronounced exception, essentially confirmed Reitan's original findings.

6. In recent years, it has been asserted informally that the HRB is "old-fashioned" and out of date because it has not kept up with developments in neuroscience, experimental neuropsychology, and psychometrics. So-called process and cognitive

approaches now reflect the state-of-the-art in neuropsychological assessment. This view is not supported for a number of reasons including continued widespread clinical use of all or some of the HRB, great popularity of continuing education concerning clinical use of the HRB, and a large number of publications in the current literature written by Reitan and collaborators themselves (e.g., Reitan & Wolfson, 1997; Reitan & Wolfson, 1995) or by others using HRB-oriented test batteries (e.g., Goldstein, Beers, & Shemansky, 1996; Goldstein & Shemansky, 1997; Palmer et al., 1997).

In summary, many criticisms have been raised of the HRB as a comprehensive, standard neuropsychological assessment system. While pertinent and reasonable responses have been made to most or all of these critiques, members of the profession have nevertheless sensed in recent years the desire to develop alternative procedures. Despite the pertinent replies to criticisms, there appear to be many clinicians who still feel that the HRB *is* too long, *does* neglect memory, and in many cases *is* insufficiently specific. Some holders of these views adopted an individualized approach, or modified the HRB, while others sought alternative standard batteries.

The Luria-Nebraska Neuropsychological Battery (LNNB)

History

This procedure, previously known as the Luria-South Dakota Neuropsychological Battery or as A Standard Version of Luria's Neuropsychological Tests, was first reported on in 1978 (Golden, Hammeke, & Purisch, 1978; Purisch, Golden, & Hammeke, 1978) in the form of two initial validity studies. One could provide a lengthy history of this procedure, going back to Luria's original writings, or a brief one only recording events that occurred since the time of preparation of the two publications cited above. We will take the latter alternative, for reasons that will become apparent. Prior to the past quarter of a century, Luria was a shadowy figure to most English-speaking neuropsychologists. It was known that he was an excellent clinician who had developed his own methods for evaluating patients as well as his own theory, but the specific contents were unknown until translations of some of his major works appeared in the 1960s (e.g., Luria, 1966). However, when these works were read by English-speaking professionals, it became apparent that Luria did not have a standard battery of the HRB type and did not even appear to use standardized tests. Thus, while his formulations and case presentations were stimulating and innovative, nobody quite knew what to do with these materials in terms of practical clinical application. One alternative, of course, was to go to the Soviet Union and study with Luria, and, in fact, Anne-Lise Christensen did just that and reported what she had learned in a book called *Luria's Neuropsychological Investigation* (Christensen, 1975a). The book was accompanied by a manual and a kit containing test materials used by Luria and his coworkers (Christensen, 1975b, 1975c). Even though some of Luria's procedures previously appeared in English in the *Higher Cortical Functions* (1966) and *Traumatic Aphasia* (1970), they were never presented in a manner that encouraged direct administration of the test items to patients. Thus, the English-speaking public had in hand a manual and related materials that could be used to administer some of Luria's tests. These materials did not contain information relevant to standardization of these items. There are no scoring systems, norms, data regarding validity and reliability, or review of research accomplished with the procedure as a standard battery. This work was taken on by a group of investigators under the leadership of Charles J. Golden and was initially reported on in the two 1978 papers cited above. Thus, in historical sequence, Luria adopted or developed these items over the course of many years, Christensen published them in English but without standardization data, and finally Golden and collaborators provided quantification and standardization. Since that time, Golden's group as well as other investigators have produced a massive amount of studies with what is now known as the Luria-Nebraska Neuropsychological Battery. The battery was originally published in 1980 by Western Psychological Services (Golden et al., 1980s) and is now extensively used in clinical and research applications. An alternate form of the battery is now available (Golden, Purisch, & Hammeke, 1985), as is a children's version (Golden, 1981).

Structure and Content

The Luria-Nebraska is an evolving procedure, and the details presented here will no doubt change over the years. However, the basic structure of the battery will probably remain essentially the same. The current version contains 269 items, each of which may be scored on a two- and three- point scale. A score of 0 indicates normal performances. Some items may receive a score of 1, indicating borderline performance. A score of 2 indicates clearly abnormal performance. The items are organized into the categories provided in the Christensen kit (Christensen, 1975c), but while Christensen organized the items primarily to suggest how they were used by Luria, in the Luria-Nebraska version the organization is presented as a set of quantitative scales. The raw score for each scale is the sum of the 0, 1, and 2 item scores. Thus, the higher the score, the poorer the performance. Since the scales contain varying numbers of items, raw scale scores are converted to T scores with a mean of 50 and a standard deviation of 10. These T scores are displayed as a profile on a form prepared for that purpose. The scores for the individual items may be based on speed, accuracy, or quality of response. In some cases, two scores may be assigned to the same task, one for speed and the other for accuracy. These two scores are counted as individual items. For example, one of the items is a block-counting task, with separate scores assigned for number of errors and time to completion of the task. In the case of time scores, blocks of seconds are associated with the 0, 1, and 2 scores. When quality of response is scored, the manual provides both rules for scoring, and, in the case of copying tasks, illustrations of figures representing 0, 1, and 2 scores.

The 269 items are divided into 11 content scales, each of which is individually administrable. These scales were originally called the Motor, Rhythm, Tactile, Visual, Receptive Speech, Expressive Speech, Writing, Reading, Arithmetic, Memory, and Intellectual Processes scales. in the new manual, the names of the content scales have been replaced by abbreviations. Thus, the clinical scales are referred to as the C1 to C11 scales. In addition to these 11 content scales, there are three derived scales that appear on the standard profile form: the Pathognomic, Left Hemisphere, and Right Hemisphere scales. The Patognomonic scale contains items from throughout the battery found to be particularly sensitive to presence or absence of brain damage. The Left and Right Hemisphere scales are derived from the Motor and Tactile scale items that involve comparisons between the left and right sides of the body. They therefore reflect sensory-motor asymmetries between the two sides of the body.

Several other scales have been developed by Golden and various collaborators, all of which are based on different ways of scoring the same 269 items. These special scales include empirically derived right and left hemisphere scales (McKay & Golden, 1979a), a series of localization scales (McKay & Golden, 1979b), a series of factor scales (McKay & Golden, 1981), and double discrimination scales (Golden, 1979). The empirical right and left hemisphere scales contain items from throughout the battery and are based on actual comparisons among patients with right and left hemisphere, and diffuse brain damage. The localization scales are also empirically derived (McKay & Golden, 1979b), being based on studies of patients with localized brain lesions. There are frontal, sensory-motor, temporal, and parieto-occipital scales for each hemisphere. The factor scales are based on extensive factor-analytic studies of the battery involving factor analyses of each of the major content scales (e.g., Golden & Berg, 1983). The empirical right and left hemisphere, localization, and factor scales may all be expressed in T scores with a mean of 50. The double discrimination scales which have been shown to be effective in diagnosis of multiple sclerosis (Golden, 1979), involve development of two scales: one contains items on which patients with a particular diagnosis do worse than the general neurological population, and the other contains items on which patients do better. Classification to the specific group is made when scores are in the appropriate range on both scales. There are also two scales that provide global indices of dysfunction, and are meant as equivalents to the Halstead Impairment Index. They are called the Profile Elevation and Impairment scales.

The Luria-Nebraska procedure involves an age and education correction. it is accomplished through computation of a cut-off score for abnormal performance based on an equation that takes into consideration both age and education. The computed score is called the critical level and is equal to .214 (Age) + 1.47 (Education) + 68.8 (Constant). Typically, a horizontal line is drawn across the profile at the computed critical level point. The test user has the option of considering

scores above the critical level, which may be higher or lower than 60, as abnormal.

As indicated above, extensive factor-analytic studies have been accomplished, and the factor structure of each of the major scales has been identified. These analyses were based on item intercorrelations, rather on correlations among the scales. It is important to note that most items on any particular scale correlate more highly with other items on that scale than they do with items on other scales (Golden, 1981). This finding lends credence to the view that the scales are at least somewhat homogeneous, and thus that the organization of the 269 items into those scales can be justified.

Theoretical Foundations

As in the case of the HRB, one could present two theoretical bases for the Luria-Nebraska, one revolving around the name of Luria and the other around the Nebraska group, Golden and his collaborators. This view is elaborated upon in Goldstein (1986). It is to be noted in this regard that Luria himself had nothing to do with the development of the Luria-Nebraska Battery, nor did any of his coworkers. The use of his name in the title of the battery is, in fact, somewhat controversial, and seems to have been essentially honorific in intent, recognizing his development of the items and the underlying theory for their application. Indeed, Luria died some time before publication of the battery but was involved in the preparation of the Christensen materials, which he endorsed. Futhermore, the method of testing employed by the Luria-Nebraska was not Luria's Method, and the research done to establish the validity, reliability, and clinical relevance of the Luria-Nebraska was not the kind of research done by Luria and his collaborators. Therefore, our discussion of the theory underlying the Luria-Nebraska Battery will be based on the assumption that the only connecting link between Luria and that procedure is the set of Christensen items. In doing so, it becomes clear that the basic theory underlying the development of Luria-Nebraska is based on a philosophy of science that stresses empirical validity, quantification, and application of established psychometric procedures. Indeed, as pointed out elsewhere (Goldstein, 1986), it is essentially the same epistemology that characterizes the work of the Reitan group.

The general course charted for establishment of quantitative, standard neuropsychological assessment batteries involves several steps: (a) determining whether the battery discriminates between brain-damaged patients in general and normal controls; (b) determining whether it discriminates between patients with structural brain damage and conditions that may be confused with structural brain damage, notably various functional psychiatric disorders; (c) determination of whether the procedure has the capacity to lateralize and regionally localize brain damage; and (d) determination of whether there are performance patterns specific to particular neurological disorders, such as alcoholic dementia or multiple sclerosis. In proceeding along this course, it is highly desirable to accomplish appropriate cross-validations and to determine reliability. This course was taken by Golden and his collaborators, in some cases with remarkable success. Since the relevant research was accomplished during recent years, it had the advantages of being able to benefit from the new brain-imaging technology, notably the CT scan, and the application of high-speed computer technologies, allowing for extensive use of powerful multivariate statistical methods. With regard to methods of clinical inference, the same methods suggested by Reitan—level of performance, pattern of performance, pathognomonic signs, and right-left comparisons—are the methods generally used with the Luria-Nebraska.

Adhering to our assumption that the Luria-Nebraska bears little resemblance to Luria's methods and theories, there seems little point in examining the theoretical basis for the substance of the Luria-Nebraska Battery. For example, it seems that there would be little point in examining the theory of language that underlies the Receptive Speech and Expressive Speech scales or the theory of memory that provides the basis for the Memory scale. An attempt to produce such an analysis was made some time ago by Spiers (1981), who examined the content of the Luria-Nebraska scales and evaluated it with reference not so much to Luria's theories, but to current concepts in clinical neuropsychology in general. However, despite the thoroughness of the Spiers review, it seems to miss the essential point that the Luria-Nebraska is a procedure based primarily on studies of empirical validity. One can fault it on the quality of it empirical validity, but not on the basis that it utilizes such an approach. It therefore appears that the Luria-Nebraska Battery does not constitute a means of

using Luria's theory and methods in English-speaking countries, but rather is a standardized psychometric instrument with established validity for certain purposes and reliability. The choice of using items selected by Christensen (1975b) to illustrate Luria's testing methods was, in retrospect, probably less crucial than the research methods chosen to investigate the capabilities of this item set. Indeed, it is somewhat misleading to characterize these items as "Luria's tests," since many of them are standard items used by neuropsychologists and neurologists throughout the world. Surely, one cannot describe asking a patient to interpret proverbs or determine two-point thresholds as being exclusively "Luria's test." They are, in fact, venerable, widely used procedures.

Standardization Research

Fortunately, there are published manuals for the Luria-Nebraska (Golden, Hammeke, & Purisch, 1980; Golden, Purisch, & Hammeke, 1985) that describe the battery in detail and provide pertinent information relative to validity, reliability, and norms. There are also several review articles (e.g., Golden, 1981; Moses & Purisch, 1997; Purisch & Sbordone, 1986) that comprehensively describe the research done with the battery. Very briefly reviewing this material, satisfactory discriminative validity has been reported in studies directed toward differentiating miscellaneous brain-damaged patients from normal controls and from chronic schizophrenics. Cross validations were generally successful, but Shelly and Goldstein (1983) could not fully replicate the studies involved with discriminating between brain-damaged and schizophrenic patients. Discriminative validity studies involving lateralization and localization achieved satisfactory results, but the localization studies were based on small samples. Quantitative indices from the Luria-Nebraska were found to correlate significantly with CT-scan quantitative indices in alcoholic (Golden, Graber, Blose, Berg, Coffmanm, & Block, 1981) and schizophrenic (Golden, Moses, Zelazowski, Graber, Zatz, Horvath, & Berger, 1980) samples. There have been several studies of specific neurological disorders including multiple sclerosis (Golden, 1979), alcoholism (Chmielewski & Golden, 1980), Huntington's disease (Moses, Golden, Berger, & Wisniewski, 1981) and learning-disabled adults, (McCue, Shelly, Goldstein, & Katz-Garris, 1984), all with satisfactory results in terms of discrimination.

The test manual reports reliability data. Test-retest reliabilities for the 13 major scales range from .78 to .96. The problem of interjudge reliability is generally not a major one for neuropsychological assessment, since most of the tests used are quite objective and have quantitative scoring systenms. However, there could be a problem with the Luria-Nebraska, since the assignment of 0, 1, and 2 scores sometimes requires a judgment by the examiner. During the preliminary screening stage in the development of the battery, items in the original pool that did not attain satisfactory interjudge reliability were dropped. A 95 percent interrater agreement level was reported by the test constructors for the 282 items used in an early version of the battery developed after the dropping of those items. The manual contains means and standard deviations for each item based on samples of control, neurologically impaired, and schizophrenic subjects. An alternate form of the battery is available. In recent years there have been a small number of successful predictive or ecological validity studies reviewed in Moses and Purisch (1997). It is unclear whether or not there have been studies addressed to the issue of construct validity. Stambrook (1983) suggested that studies involved with item-scale consistency, factor analysis, and correlation with other instruments are construct-validity studies, but it does not appear to us that they are directed toward validation of Luria's constructs. The attempt to apply Luria's constructs has not in fact involved the empirical testing of specific hypotheses derived from Luria's theory. Thus, we appear to have diagnostic or discriminative validity established by a large number of studies. There also seems to be content validity, since the items correlate most highly with the scale to which they are assigned, but the degree of construct validity remains unclear. For example, there have been no studies of Luria's important construct of the functional system or of his hypotheses concerning the role of frontal lobe function in the programming, regulation, and verification of activity (Luria, 1973).

With regard to Form II, it is important to note that Moses and Purisch (1997) have provided clear evidence that Forms I and II are not equivalent forms, and should not be used in longitudinal studies as alternate forms. Form II cannot be hand-scored and must be scored using a computer pro-

gram. It also includes a new clinical scale; Intermediate Memory (C12).

Evaluation

At this writing the early heated controversies concerning the Luria-Nebraska Battery appear to have diminished and we no longer see the highly critical reviews that appeared shortly after the procedure first appeared. At that time Adams (1980) criticized it primarily on methodological grounds, Spiers (1981) on the basis that is was greatly lacking in its capacity to provide a comprehensive neuropsychological assessment, Crosson and Warren (1982) because of its deficiencies with regard to assessment of aphasia and aphasic patients, and Stambrook (1983) on the basis of a number of methodological and theoretical considerations. Replies were written to several of these reviews (e.g., Golden, 1980), and a rather heated literature controversy eventuated. This literature was supplemented by several case studies (e.g., Delis & Kaplan, 1982) in which it was shown that the inferences that would be drawn from the Luria-Nebraska were incorrect with reference to documentation obtained for those cases.

These criticisms can be divided into general and specific ones. Basically, there are two general criticisms: (a) The Luria-Nebraska Battery does not reflect Luria's thinking in any sense, and his name should not be used in describing it; and (b) there are several relatively flagrant methodological difficulties involved in the standardization of the procedure. The major specific criticisms primarily involve the language related and memory scales. With regard to aphasia, there are essentially two points. First, there is no system provided, nor do the items provide sufficient data to classify the aphasias in terms of some contemporary system (e.g., Goodglass & Kaplan, 1983). Second, the battery is so language-oriented that patients with aphasia may fail many of the nonlanguage tasks because of failure to comprehend the test instructions or to make the appropriate verbal responses indicative of a correct answer. For example, on the Tactile scale, the patient must name objects placed in the hands. Patients with anomia or anomic aphasia will be unable to do that even though their tactile recognition skills may be perfectly normal. With regard to memory, the Memory scale is criticized because of its failure to provide a state-of-the-art comprehensive memory assessment (Russell, 1981). Golden has responded to this criticism through adding additional items involving delayed recall to the alternate form of the battery. Moses and Purisch (1997) have reviewed this critical material and concluded that for various reasons several of the criticisms were unfounded, the validity and reliability of the Luria-Nebraska has been demonstrated in a large number of studies, and that efforts were made in updated versions of the battery to correct for reasonable criticisms.

In providing an evaluation of the Luria-Nebraska, one can only voice an opinion, as others have, since its existence has stimulated a polarization into "those for it" and "those against it." I would concur with Stambrook's view (1983), which essentially is that it is premature to make an evaluation, and that major research programs must be accomplished before an informed opinion can be reached. I hold this opinion at the present writing, some 15 years after appearance of the Stambrook paper. However, the need for an expanded database expressed in the previous versions of this chapter has been largely fulfilled through the efforts of James Moses and collaborators (Moses & Purisch, 1997). There is still need for more evaluation of the actual constructs on which the procedure is based, and assessment of its clinical usefulness relative to other established procedures such as the HRB or individual approaches. The following remark by Stambrook (1983) continues to reflect a highly reasoned approach to this issue. "The clinical utility of the LNNB does not depend upon either the publisher's and test developer's claims, or on conceptual and methodological critiques, but upon carefully planned and well-executed research" (p. 266). Various opinions have also been raised with regard to whether it is proper to utilize the Luria-Nebraska in clinical situations. It continues to be my view of the matter that it may be so used as long as inferences made from it do not go beyond what can be based on the available research literature. In particular, the test consumer should not be led to believe that administration and interpretation of the Luria-Nebraska battery provide an assessment of the type that would have been conducted by Luria and his coworkers, or that one is providing an application of Luria's method. The procedure is specifically not Luria's method at all, and the view that is provides valid measures of Luria's constructs and theories has not been verified. Even going beyond that point, attempts to verify some of Luria's hypotheses (e.g., Drewe, 1975; Goldberg & Tucker, 1979) have not always

been completely successful. Therefore, clinical interpretations, even when they are based on Luria's actual method of investigation, may be inaccurate because of inaccuracies in the underlying theory.

SUMMARY AND CONCLUSIONS

In the first part of this chapter, general problems in the area of standardization of comprehensive neuropsychological test batteries were discussed, while the second part contained brief reviews of the two most widely used procedures, the HRB and the Luria-Nebraska. it was generally concluded that these batteries have their advantages and disadvantages. The HRB is well established and detailed but is lengthy, cumbersome, and neglects certain areas, notably memory. The Luria-Nebraska is also fairly comprehensive and briefer than the HRB but is currently quite controversial and is thought to have major deficiencies in standardization and rationale, at least by some observers. I have taken the view that all of these standard batteries are screening instruments, but not in the sense of screening for presence or absence of brain damage. Rather, they may be productively used to screen a number of functional areas such as memory, language, or visual-spatial skills, that may be affected by brain damage. With the development of the new imaging techniques in particular, it is important that the neuropsychologist not simply tell the referring agent what he or she already knows. The unique contribution of standard neuropsychological assessment is the ability to describe functioning in many crucial areas on a quantitative basis. The extent to which one procedure can perform this type of task more accurately and efficiently than other procedures will no doubt greatly influence the relative acceptability of these batteries by the professional community.

REFERENCES

Adams, K. M. (1980). In search of Luria's battery: A false start. *Journal of Consulting and Clinical Psychology, 48,* 511–516.

Adams, R. L., & Trenton, S. L. (1981). Development of a paper-and-pen form of the Halstead Category test. *Journal of Consulting and Clinical Psychology, 49,* 298–299.

Albert, M. L., Goodglass, H., Helm, N. A., Rubens, A. B., & Alexander, M. P. (1981). *Clinical aspects of dysphasia.* New York: Springer-Verlag/Wein.

Allen, D. N., Goldstein, G., & Seaton, B. E. (1997). Cognitive rehabilitation of chronic alcohol abusers. *Neuropsychology Review, 7,* 21–39.

Baddeley, A. D., Wilson, B. A., & Watts, F. N. (1995). *Handbook of memory disorders.* Chichester, UK: Wiley.

Bender, L. (1938). A visual motor gestalt test and its clinical use. *American Orthopsychiatric Association, Research Monographs,* No. 3.

Bender, M. B. (1952). *Disorders in perception.* Springfield, IL: Charles C. Thomas.

Benson, D. F., & Ardila A. (1996). *Aphasia: A clinical perspective.* New York: Oxford University Press.

Benton, A. L. (1961). The fiction of the Gerstmann Syndrome. *Journal of Neurology, Neurosurgery and Psychiatry, 24,* 176–181.

Benton, A. L. (1963). *The Revised Visual Retention Test.* New York: Psychological Corporation.

Benton, A. L., Sivan, A. B., Hamsher, K. deS., Varney, N. R., & Spreen, O. (1994). *Contributions to neuropsychological assessment* (2nd ed.). New York: Oxford University Press.

Ben-Yishay, Y., Diller, L., Gertsman, L., & Gordon, W. (1970). Relationship between initial competence and ability to profit from cues in brain-damaged individuals *Journal of Abnormal Psychology, 78,* 248–259.

Blumstein, S. E. (1981). Neurolinguistic disorders: Language-brain relationships. In S. B. Filskov and T. J. Boll (Eds.), *Handbook of clinical neuropsychology* (pp. 227–256). New York: Wiley-Interscience.

Boll, T. J. (1981). The Halstead-Reitan neuropsychology battery. In S. B. Filskov & T. J. Boll (Eds.), *Handbook of clinical neuropsychology* (pp. 577–607). New York: Wiley-Interscience.

Butters, N. (1983, August). *Clinical aspects of memory disorders: Contributions from experimental studies of amnesia and dementia.* Presented at American Psycholocial Association, Division 40 Presidential Address, Anaheim, CA.

Butters, N. M., & Cermak, L. S. (1980). *Alcoholic Korsakoff's syndrome.* New York: Academic Press.

Calsyn, D. A., O'Leary, M. R., & Chaney, E. F. (1980). Shortening the Category Test. *Journal of Consulting and Clinical Psychology, 48,* 788–789.

Canter, A. (1970). *The Canter Background Interference Procedure for the Bender-Gestalt Test: Manual for administration, scoring and interpretation.* Iowa City, IA: Iowa Psychopathic Hospital.

Chmielewski, C., & Golden, C. J. (1980). Alcoholism and brain damage; An investigation using

the Luria-Nebraska Neuropsychological Battery. *International Journal of Neuroscience, 10,* 99–105.

Christensen, A. L. (1975a). *Luria's neuropsychological investigation.* New York: Spectrum.

Christensen, A. L. (1975b). *Luria's neuropsychological investigation: Manual.* New York: Spectrum.

Christensen, A. L. (1975c). *Luria's neuropsychological investigation:* Test cards. New York: Spectrum.

Colsher, P. L., & Wallace, R. B. (1991). Longitudinal application of cognitive function measures in a defined population of community-dwelling elders. *Annals of Epidemiology, 1,* 215–230.

Crary, M. A., Voeller, K. K. S., & Haak, N. J. (1988). Questions of developmental neurolinguistic assessment. In M. G. Tramontana & S. R. Hooper (Eds.), *Assessment issues in child neuropsychology* (pp. 249–279). New York: Plenum Press.

Cronbach, L. J. (1960). *Essentials of psychological testing* (2nd ed.). New York: Harper & Brothers.

Crosson, B., & Warren, R. L. (1982). Use of the Luria-Nebraska Neuropsychological Battery in aphasia: A conceptual critique. *Journal of Consulting and Clinical Psychology, 50,* 22–31.

Cummings, J. L. (Ed.). (1990). *Subcortical dementia.* New York: Oxford University Press.

Davis, K. (1983, October). *Potential neurochemical and neuroendocrine validators of assessment instruments.* Paper presented at the conference on Clinical Memory Assessment of Older Adults, Wakefield, MA.

DeFillippis, N. A., McCampbell, E., & Rogers, P. (1979). Development of a booklet form of the Category Test: normative and validity data. *Journal of Clinical Psychology, 50,* 32–39.

Delis, D. C., & Kaplan, E. (1982). The assessment of aphasia with the Luria-Nebraska neuropsychological battery: A case critique. *Journal of Consulting and Clinical Psychology, 50,* 32-39.

Drewe, E. A. (1975). An experimental investigation of Luria's theory on the effects of frontal lobe lesions in man. *Neuropsychological, 13,* 421–429.

Dunn, L. M., & Markwardt, F. C. (1970). *Peabody individual Achievement Test Manual.* Circle Pine, MN: American Guidance Service.

Evans, D. A., Beckett, L. A., Albert, M. S., Herbert, L. E., Scherr, P. A., Funkenstein, H. H., & Taylor, J. O. (1993). Level of education and change in cognitive function in a community population of older persons. *Annals of Epidemiology, 3,* 71–77.

Filskov, S. B., & Boll, T. J. (1981). *Handbook of clinical neuropsychology.* New York: Wiley-Interscience.

Filskov, S. B., & Goldstein, S. G. (1974). Diagnostic validity of the Halstead-Reitan neuropsychological battery. *Journal of Consulting and clinical Psychology, 42,* 382–388.

Fitzhugh, K. B., Fitzhugh, L. C., & Reitan, R. M. (1961): Psychological deficits in relation to acuteness of brain dysfunction. *Journal of Consulting Psychology, 25,* 61–66.

Fitzhugh, K. B., Fitzhugh, L. C., & Reitan, R. M. (1962). Wechsler-Bellevue comparisons in groups of 'chronic' and 'current' lateralized and diffuse brain lesions. *Journal of Consulting Psychology, 26,* 306–310.

Fletcher, J. M., & Satz, P. (1980). Developmental changes in the neuropsychological correlates of reading achievement: A six-year longitudinal follow-up. *Journal of Clinical Neuropsychology, 2,* 23–37.

Freedman, M. (1990). Parkinson's disease. In J. L. Cummings (Ed.), *Subcortical dementia* (pp. 108–122). New York: Oxford University Press.

Goldberg, E., & Tucker, D. (1979). Motor preservation and long-term memory for visual forms. *Journal of Clinical Neuropsychology, 1,* 273–288.

Golden, C. J. (1978). *Diagnosis and rehabilitation in clinical neuropsychology.* Springfield, IL: C. C. Thomas.

Golden, C. J. (1979). Identification of specific neurological disorders using double discrimination scales derived from the standardized Luria neuropsychological battery. *International Journal of Neuroscience, 10,* 51–56.

Golden, C. J. (1980). In reply to Adams' "In search of Luria's Battery: A false start." *Journal of Consulting and Clinical Psychology, 48,* 517–521.

Golden, C. J. (1981). A standardized version of Luria's neuropsychological tests: A quantitative and qualitative approach to neuropsychological evaluation. In S. B. Filskov & T. J. Boll (Eds.), *Handbook of clinical neuropsychology* (pp. 608–642). New York: Wiley-Interscience.

Golden, C. J., & Berg, R. A. (1983). Interpretation of the Luria-Nebraska Neuropsychological Battery by item intercorrelation: The memory scale. *Clinical Neuropsychology, 5,* 55–59.

Golden, C. J., Graber, B., Blose, I., Berg, R., Coffman, J., & Block, S. (1981). Differences in brain densities between chronic alcoholic and normal control patients. *Science, 211,* 508–510.

Golden, C. J., Hammeke, T., & Purisch, A. (1978). Diagnostic validity of the Luria neuropsycho-

logical battery. *Journal of Consulting and Clinical Psychology, 46,* 1258–1265.

Golden, C. J. Hammeke, T., & Purisch, A. (1980). *The Luria-Nebraska battery manual.* Los Angeles: Western Psychological Services.

Golden, C. J., Moses, J. A., Zelazowski, R., Graber, B., Zatz, L. M., Horvath, T. B., & Berger, P. A. (1980). Cerebral vertricular size and neuropsychological impairment in young chronic schizophrenics. *Archives of General Psychiatry, 37,* 619–623.

Golden, C. J., Purisch, A., & Hammeke, T. (1985). *Luria-Nebraska Neuropsychological Battery Manual-Forms I and II.* Los Angeles: Western Psychological Services.

Goldstein, G. (1978). Cognitive and perceptual differences between schizophrenics and organics. *Schizophrenia bullentin, 4,* 160–185.

Goldstein, G. (1986). The neuropsychology of schizophrenia. In I. Grant and K. M. Adams (Eds.), *Neuropsychological assessment of neuropsychiatric disorders* (pp. 147–171). New York: Oxford University Press.

Goldstein, G. (1986). An overview of similarities and differences between the Halstead-Reitan and Luria-Nebraska batteries. In T. Incagnoli, G. Goldstein, & C. J. Golden (Eds.), *Clinical application of neuropsychological test batteries* (pp. 235–275). New York: Plenum Press.

Goldstein, G. (1991). Comprehensive neuropsychological test batteries and research in schizophrenia. In S. R. Steinhauer, J. H. Gruzelier, & J. Zubin (Eds.), *Handbook of schizophrenia: Volume 5, Neuropsychology, psychophysiology, and information processing* (pp. 525–551). Amsterdam: Elsevier.

Goldstein, G. (1994). Cognitive heterogeneity in psychopathology: The case of schizophrenia. In P. Vernon (Ed.). *The neuropsychology of individual differences* (pp. 209-233). New York: Academic Press.

Goldstein, G., Beers, S. R., & Shemansky, W. J. (1996). Neuropsychological differences between schizophrenic patients with heterogeneous Wisconsin Card Sorting Test performance. *Schizophrenia Research, 21,* 1–18.

Goldstein, G., Nussbaum, P. D., & Beers, S. R. (1998). *Human brain function: Assessment and rehabilitation: Neuropsychology.* New York: Plenum Press.

Goldstein, G., & Beers, S. R. (1998). *Human brain function: Assessment and rehabilitation: Rehabilitation.* New York: Plenum Press.

Goldstein, G., & Ruthven, L. (1983). *Rehabilitation of the brain damaged adult.* New York: Plenum.

Goldstein, G., & Shelly, C. (1971). Field dependence and cognitive, perceptual and motor skills in alcoholics: A factor analytic study. *Quarterly Journal of Studies on Alcohol, 32,* 29–40.

Goldstein, G., & Shelly, C. (1972). Statistical and normative studies of the Halstead Neuropsychological Test Battery relevant to a neuropsychiatric hospital setting. *Perceptual and Motor Skills, 34,* 603–620.

Goldstein, G., & Shelly, C. H. (1975). Similarities and differences between psychological deficit in aging and brain damage. *Journal of Gerontology, 30,* 448–455.

Goldstein, G., & Shelly, C. (1982). A further attempt to cross-valiate the Russell, Neuringer, and Goldstein neuropsychological keys. *Journal of Consulting and Clinical Psychology, 50,* 721–726.

Goldstein, G., & Shemansky, W. J. (1997). Patterns of performance by neuropsychiatric patients on the Halstead Category Test: Evidence for conceptual learning in schizophrenic patients. *Archives of Clinical Neuropsychology, 12,* 251-255.

Goldstein, G., & Watson, J. R. (1989). Test-retest reliability of the Halstead-Reitan battery and the WAIS in a neuropsychiatric population. *The Clinical Neuropsychologist, 3,* 265-273.

Goldstein, K., & Scheerer, M. (1941). Abstract and concrete behavior: An experimental study with special tests. *Psychological Monographs, 63,* (Whole No. 239).

Goodglass, H. (1983, August). Aphasiology in the United States. In G. Goldstein (Chair), *Symposium: History of Human Neuropsychology in the United States.* Ninety-first annual convention of the American Psychological Association, Anaheim, CA.

Goodglass, H., & Kaplan, E. (1983). *The assessment of aphasia and related disorders* (2nd ed.). Philadelphia, PA: Lea & Febiger.

Gruzelier, J. H. (1991). Hemispheric imbalance: Syndromes of schizophrenia, premorbid personality, and neurodevelopmental influences. In S. R. Steinhauer, J. H. Gruzelier, & J. Zubin (Eds.), *Handbook of schizophrenia: Volume 5, Neuropsychology, physchophysiology, and information processing* (pp. 599–650. Amsterdam: Elsevier.

Guilmette, T. J., & Kastner, M. P. (1996). The prediction of vocational function from neuropsychological data. In R. J. Sbordone & C. J. Long (Eds.), *Ecological validity of neuropsychological testing* (pp. 387–411). Delray Beach, FL: GR Press/St. Lucie Press.

Halstead, W. C. (1947). *Brain and intelligence: A quantitative study of the frontal lobes.* Chicago: The University of Chicago Press.

Heaton, R. K. (1980). *A manual for the Wisconsin Card Sorting Testing.* Odessa, FL: Psychological Assessment Resources, Inc.

Heaton, R. K., Baade, L. E., & Johnson, K. L. (1978). Neuropsychological test results associated with psychiatric disorders in adults. *Psychological Bulletin, 85* 141–162.

Heaton, R. K., & Crowley, T. (1981). Effects of psychiatric disorders and their somatic treatment on neuropsychological test results. In S. B. Filskov & T. J. Boll (Eds.), *Handbook of clinical neuropsychology.* New York: Wiley-Interscience.

Heaton, R. K., Grant, I., & Matthews, C. G. (1991). *Comprehensive norms for an expanded Halstead-Reitan Battery.* Odessa, FL: Psychological Assessment Resources.

Heaton, R. K., & Pendleton, M. G. (1981). Use of neuropsychological tests to predict adult patients' everyday functioning *Journal of Consulting and Clinical Psychology, 49,* 807–821.

Henn, F. A., & Nasrallah, H. A. (1982). *Schizophrenia as a brain disease.* New York: Oxford University Press.

Jarvis, P. E., & Barth, J. T. (1984). *Halstead-Reitan Test Battery: An interpretive guide.* Odessa, FL: Psychological Assessment Resources.

Jastak, S., & Wilkinson, G. S. (1984). *The Wide Range Achievement Test-Revised.* Wilmington, DE: Jastak Associates, Inc.

Jeannerod, M. (Ed.). (1987). *Neurophysiological and neuropsychological aspects of spatial neglect.* Amsterdam: North Holland.

Jones, B. P., & Butters, N. (1983). Neuropsychological assessment. In M. Hersen, A. S. Bellack, and A. E. Kazdin (Eds.), *The clinical psychology handbook* (pp. 377–396). New York: Pergamon Press.

Kaplan, E. (1979). Presidential address. Presented at the International Neuropsychological Society, Noordwijkerhout, Holland.

Kaplan, E. H., Goodglass, H., & Weintraub, S. (1983). *The Boston Naming Test.* Philadelphia: Lea & Fibiger.

Kertesz, A. (1979). *Aphasia and associated disorders: Taxonomy, localization and recovery.* New York: Grune & Stratton.

Kimura, D. (1961). Some effects of temporal lobe damage on auditory perception. *Canadian Journal of Psychology, 15,* 156–165.

Kinsbourne, M. (1980). Attentional dysfunctions and the elderly: Theoretical models and research perspectives. In L. W. Poon, J. L. Fozard, L. S. Cermak, D. Arenberg, & L. W. Thompson (Eds.), *New directions in memory and aging* (pp.l 113–129). Hillsdale, NJ: Erlbaum.

Klove, H. (1974). Validation studies in adult clinical neuropsychology. In R. M. Reitan & L. H. Davison (Eds.), *Clinical neuropsychology: Current status and applications* (pp. 211–235). Washington, DC: V. H. Winston & Sons.

Levin, H. S., Benton, A. L., & Grossman, R. G. (1982). *Neurobehavioral consequences of closed head injury.* New York: Oxford University Press.

Lewis, R. F., & Rennick, P. M. (1979). *Manual for the Repeatable Cognitive-Perceptual-Motor Battery.* Grosse Pointe Park, MI: Axon Publishing Co.

Lezak, M. (1976). *Neuropsychological Assessment* (1st ed.). New York: Oxford University Press.

Lezak, M. (1995). *Neuropsychological Assessment* (3rd ed.). New York: Oxford University Press.

Lyon, G. R., & Flynn, J. F. (1991). Educational validation studies with subtypes of learning-disabled readers. In B. P. Rourke (Ed.), *Neuropsychological validation of learning disability subtypes* (pp. 233–242). New York: The Guilford Press.

Luria, A. R. (1966). *Higher cortical functions in man.* New York: Basic Books.

Luria, A. R. (1970). *Traumatic aphasia.* The Hague: Mouton and Co. Printers.

Luria, A. R. (1973). *The working brain.* New York: Basic Books.

Malec, J. (1978). Neuropsychological assessment of schizophrenia vs. brain damage: A review. *Journal of Nervous and Mental Disease, 166,* 507–516.

Marsh, L., Lauriello, J., Sullivan, E. V., & Pfefferbaum, A. (1996). Neuroimaging in psychiatric disorders. In E. Bigler (Ed.), *Neuroimaging II: Clinical applications* (pp. 73–125). New York: Plenum Press.

Matthews, C. G. (1981). Neuropsychology practice in a hospital setting. In S. B. Filskov & T. J. Boll (Eds.), *Handbook of clinical neuropsychology.* New York: Wiley-Interscience.

McCue, M. (1997). The relationship between neuropsychology and functional assessment in the elderly. In P. D. Nussbaum (Ed.), *Handbook of neuropsychology and aging* (pp. 394–408). New York: Plenum Press.

McCue, M., Shelly, C., Goldstein, G., & Katz-Garris, L. (1984). Neuropsychological apsects of learning disability in young adults. *Clinical Neuropsychology, 6,* 229–233.

McKay, S., & Golden, C. J. (1979a). Empirical derivation of experimental scales for the lateralization of brain damage using the Luria-Nebraska neuropsychological Battery. *Clinical Neuropsychology, 1,* 1–5.

McKay, S., & Golden, C. J. (1979b). Empirical derivation of experimental scales for localizing

brain lesions using the Luria-Nebraska Neuropsychological Battery. *Clinical Neuropsychology, 1,* 19–23.

McKay, S. E., & Golden, C. J. (1981). The assessment of specific neuropsychological skills using scales derived from factor analysis of the Luria-Nebraska Neuropsychological Battery. *International Journal of Neuroscience, 14,* 189–204.

Meier, M. J. (1974). Some challenges for clinical neuropsychology. In R. M. Reitan & L. A. Davison (Eds.), *Clinical Neuropsychology: Current status and applications* (pp. 289–323). Washington, DC: V. H. Winston and Sons.

Meier, M. J., Benton, A. L., & Diller, L. (1987). *Neuropsychological Rehabilitation.* Edinburgh: Churchill Livingstone.

Minshew, N. J., Goldstein, G., Dombrowski, S. N., Panchaligam, K., & Pettegrew, J. W. (1993). A preliminary 31 p-NMR study of autism: Evidence for under synthesis and increased degradation of brain membranes. *Biological Psychiatry, 33,* 762–773.

Mirsky, A. F., Anthony, B. J., Duncan, C. C., Ahearn, M. B., & Kellam, S. G. (1991). Analysis of the elements of attention: A neuropsychological approach. *Neuropsychology Review, 2,* 109–145.

Mooney, C. M. (1957). Age in the development of closure ability in children. *Canadian Journal of Psychology, 2,* 219–226.

Moses, J. A., Golden, C. J., Berger, P. A., & Wisniewski, A. M. (1981). Neuropsychological deficits in early, middle, and late stage Huntings's disease as measured by the Luria-Nebraska Neuropsychological Battery. *International Journal of Neuroscience, 14,* 95–100.

Moses, J. A. Jr., & Purisch, A. D. (1997). The evolution of the Luria-Nebraska Battery. In G. Goldstein & T. M. Incagnoli (Eds.), *Contemporary approaches to neuropsychological assessment* (pp. 131–170). New York: Plenum Press.

Newcombe, F. (1969). *Missile words of the brain: A study of psychological deficits.* Oxford: The Clarendon Press.

Newman, O. S., Heaton, R. K., & Lehman, R. A. W. (1978). Neuropsychological and MMPI correlates of patients' future employment characteristics. *Perceptual and Motor Skills, 46,* 635–642.

Nussbaum, P. D. (Ed.). (1997). *Handbook of neuropsychology and aging.* New York: Plenum Press.

Palmer, B. W., Heaton, R. K., Paulsen, J. S., Kuck, J., Braff, D., Harris, M. J., Zisook, S., & Jeste, D. V. (1997). Is it possible to be schizophrenic yet neuropsychologically normal? *Neuropsychology, 11* 437–446.

Purisch, A. D., Golden, C. J., & Hammeke, T. A. (1978). Discrimination of schizophrenic and brain-injured patients by a standardized version of Luria's neuropsychological tests. *Journal of Consulting and Clinical Psychology, 46,* 1266–1273.

Purisch, A. D., & Sbordone, R. J. (1986). The Luria-Nebraska Neuropsychological Battery. In G. Goldstein & R. E. Tarter (Eds.), *Advances in clinical neuropsychology,* (Vol. 3). New York: Plenum Press.

Reed, J. (1983, August). The Chicago-Indianapolis Group. In G. Goldstein (Chair). *Symposium: History of human neuropsychology in the United States.* Ninety-first annual convention of the American Psychological Association, Anaheim, CA.

Reed, J. C., & Reed, H. B. C. (1997). The Halstead-Reitan Neuropsychological Battery. In G. Goldstein & T. M. Incagnoli (Eds.), *Contemporary approaches to neuropsychological assessment* (pp. 93–129). New York: Plenum Press.

Reitan, R. M. (1955). An investigation of the validity of Halstead's measures of biological intelligence. *Archives of Neurology and Psychiatry, 73* 28–35.

Reitan, R. M. (1958). Qualitative versus quantitative mental changes following brain damage. *The Journal of Psychology, 46,* 339–346.

Reitan, R. M. (1959). Correlations between the trail making test and the Wechsler-Bellevue scale. *Perceptual and Motor Skills, 9,* 127–130.

Reitan, R. M. (1964). Psychological deficits resulting from cerebral lesions in man. In J. M. Warren & K. Akert (Eds.), *The frontal granular cortex and behavior* (pp. 295–312). New York: McGraw-Hill.

Reitan, R. M. (1966). A research program on the psychological effects of brain lesions in human beings. In N. R. Ellis (Ed.), *International review of research in mental retardation* (pp. 153–218). New York: Academic Press.

Reitan, R. M. (1973, August). Behavioral manifestations of impaired brain functions in aging. In J. L. Fozard (Chair), *Similarities and differences of brain-behavior relationships in aging and cerebral pathology.* Symposium presented at the American Psychological Association, Montreal, Canada.

Reitan, R. M. (1987). *The Neuropsychological Deficit Scale for Adults. Computer program.* Tucson, AZ: Neuropsychology Press.

Reitan, R. M. (1991). *The Neuropsychological Deficit Scale for Adults. Users program.* Tucson, AZ: Neuropsychology Press.

Reitan, R. M., Hom, J., & Wolfson, D. (1988). Verbal processing by the brain. *Journal of Clinical and Experimental Neuropsychology, 10,* 400–408.

Reitan, R. M., and Wolfson, D. (1993). *The Halstead-Reitan Neuropsychological Test Battery: Theory and clinical interpretation.* (2nd ed.). Tucson: Neuropsychology Press.

Reitan, R. M., & Wolfson, D. (1995). Influence of age and education on neuropsychological test results. *The Clinical Neuropsychologist, 9,* 151–158.

Reitan, R. M., & Wolfson, D. (1997). Emotional disturbances and their interaction with neuropsychological deficits. *Neuropsychology Review, 7,* 3–19.

Rey, A. (1941). L'examinen psychologique dans les cas d'encephalopahtie traumatique. *Archives de Psychologie, 28,* 286–340.

Rourke, B. P. (Ed.). (1985). *Neuropsychology of learning disabilities: Essentials of subtype analysis.* New York: The Guilford Press.

Russell, E. W. (1975a). A multiple scoring method for the assessment of complex memory functions. *Journal of Consulting and Clinical psychology, 43, 800–809.*

Russell, E. W. (1975b). Validation of a brain damage versus schizophrenia MMPI. *Journal of Clinical Psychology, 33,* 190–193.

Russell, E. W. (1981). The pathology and clinical examination of memory. In S. B. Filskov & T. J. Boll (Eds.), *Handbook of clinical neuropsychology.* New York: Wiley-Interscience.

Russell, E. W. (1993). *Halstead-Russell Neuropsychological Evaluation System, norms and conversion tables.*

Russell, E. W. (1997). Developments in the psychometric foundations of neuropsychological assessment. In G. Goldstein & T. M. Incagnoli (Eds.), *Contemporary approaches to neuropsychological assessment* (pp. 15-65). New York: Plenum.

Russell, E. W., & Levy, M. (1987). Revision of the Halstead Category Test. *Journal of Consulting and Clinical Psychology, 55,* 898–901.

Russell, E. W., Neuringer, C., & Goldstein, G. (1970). *Assessment of brain damage: A neuropsychological key approach.* New York: Wiley-Interscience.

Ryan, C. M. (1998). Assessing medically ill patients: Diabetes mellitus as a model disease. In G. Goldstein, P. D. Nussbaum, & S. R. Beers (Eds.), *Human brain function: Assessment and rehabilitation: Neuropsychology* (pp. 227–245). New York: Plenum Press.

Satz, P., Taylor, H. G., Friel, J., & Fletcher, J. M. (1978). Some developments and predictive precursors of reading disability. In A. L. Benton & D. Pearl (Eds), *Dyslexia: An appraisal of current knowledge* (pp. 313–347). New York: Oxford University Press.

Sbordone, R. J., & Long C. J. (Eds.). (1996). *Ecological validity of neuropsychological testing.* Delray Beach, FL: GR Press/St. Lucie Press.

Schear, J. M. (1984). Neuropsychological assessment of the elderly in clinical practice. In P. E. Logue and J. M. Schear (Eds.), *Clinical neuropsychology: A multidisciplinary approach* (pp. 199–235). Springfield, IL: C. C. Thomas.

Schear, J. M. (1987). Utility of cluster analysis in classification of mixed neuropsychiatric patients. *Archives of Clinical Neuropsychology, 2,* 329–341.

Scoville, W. B., & Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery, and Psychiatry, 20,* 11–21.

Seashore, C. B., Lewis, C., & Saaeteweit, J. G. (1960). *Seashore measures of musical talent: Manual.* San Antonio, TX: Psychological Corporation.

Selz, M., & Reitan, R. M. (1979). Rules for neuropsychological diagnosis: Classification of brain function in older children. *Journal of Consulting and Clinical Psychology, 47,* 258–264.

Semmes, J., Weinstein, S., Ghent, L., & Teuber, H.-L. (1960). Somatosensory changes after penetrating brain wounds in man. Cambridge, MA: Harvard University.

Shaw, D. (1966). The reliability and validity of the Halstead Category Test. *Journal of Clinical Psychology, 22,* 176–180.

Shelly, C., & Goldstein, G. (1983). Discrimination of chronic schizophrenia and brain damage with the Luria-Nebraska battery: A partially successful replication. *Clinical Neuropsychology, 5,* 82–85.

Sherrill, R. E., Jr. (1987). Options for shortening Halstead's Category Test for adults. *Archives of Clinical Neuropsychology, 5,* 82–85.

Smith, A. (1965). Certain hypothesized hemispheric differences in language and visual functions in human adults. *Cortex, 2,* 109–126.

Smith, A. (1966a). Intellectual functions in patients with lateralized frontal tumors. *Journal of Neurology, Neurosurgery, and Psychiatry, 29,* 52–59.

Smith, A. (1966b). Verbal and nonverbal test performance of patients with 'acute' lateralized brain lesions (tumors). *Journal of Nervous and Mental Disease, 141,* 517–523.

Smith, A. (1975). Neuropsychological testing in neurological disorders. In W. J. Friedlander (Ed.), *Advances in neurology* (Vol. 7, pp. 49–110). New York: Raven Press.

Sperry, R. W., Gazzaniga, M. S., & Bogen, J. E. (1969). Interhemispheric relationships: The neocortical commisures; syndromes of hemisphere disconnection. In P. J. Vinkin & G. W. Bruyen (Eds.), *Handbook of clinical neurology.* Amsterdam: North Holland.

Spiers, P. A. (1981). Have they come to praise Luria or to bury him; The Luria-Nebraska battery controversy. *Journal of Consulting and Clinical Psychology, 49,* 331–341.

Spreen, O., & Strauss, E. (1988). *A compendium of neuropsychological tests* (2nd ed.). New York: Oxford University Press.

Squire, L. R., & Butters, N. (Eds.) (1984). *Neuropsychology of memory.* New York: The Guilford Press.

Stambrook, M. (1983). The Luria-Nebraska Neuropsychological Battery: A promise that may be partly fulfilled. *Journal of Clinical Neuropsychology, 5,* 247–269.

Stein, D. G. (1988). Contextual factors in recovery from brain damage. In A.-L. Christensen & B. P. Uzzell (Eds.), *Neuropsychological rehabilitation* (pp. 1–18). Boston: Kluwer Academic Press.

Steinhauer, S. R., Hill, S. Y., & Zubin, J. (1987). Event-related potentials in alcoholics and their first-degree relatives. *Alcohol, 4,* 307–314.

Swiercinsky, D. (1978). *Manual for the adult neuropsychological evaluation.* Springfield, IL: C. C. Thomas.

Teasdale, G., & Jennett, B. (1974). Assessment of coma and impaired consciousness: A practical scale. *Lancet, 2,* 81-84.

Teuber, H.-L. (1959). Some alterations in behavior after cerebral lesions in man. In A. D. Bass (Ed.), *Evolution of nervous control from primitive organisms to man* (pp. 157–194). Washington, DC: American Association for Advancement of Science.

Teuber, H.-L. (1964). The riddle of frontal lobe function in man. In J. M. Warren & K. Akert (Eds.), *The frontal granular cortex and behavior.* (pp. 410–441). New York: McGraw-Hill.

Teuber, H.-L., Battersby, W. S., & Bender, M. B. (1951). Performance of complex visual tasks after cerebral lesions. *The Journal of Nervous and Mental Disease, 114,* 413–429.

Teuber, H.-L., & Weinstein, S. (1954). Performance on a form-board task after penetrating brain injury. *Journal of Psychology, 38,* 177–190.

Townes, B. D., Martin, D. C., Nelson, D., Prosser, R., Pepping, M., Maxwell, J., Peel, J., & Preston, M. (1985). Neurobehavioral approach to classification of psychiatric patients using a competency model. *Journal of Consulting and Clinical Psychology, 53,* 33–42.

Vega, A., & Parsons, O. (1967). Cross-validation of the Halstead-Reitan tests for brain damage. *Journal of Consulting Psychology, 31,* 619–625.

Warrington, E. K., & James, M. (1991). *Visual Object and Space Perception Battery.* Bury St. Edmunds, Suffolk, England: Thames Valley Test Co.; Gaylord, MI: National Rehabilitation Services.

Wechsler, D. (1987). *Wechsler adult intelligence—Revised* San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1997). Wechsler Adult Intelligence Scale III. San Antionio, TX: The Psychological Corporation.

Wechsler, D. (1945). *Wechsler Memory Scale Manual.* New York: Psychological Corporation.

Wechsler, D. (1987). *Wechsler Memory Scale-Revised.* New York: Psychological Corporation.

Wechsler, D. (1997). *Wechsler Memory Scale III.* San Antonio, TX: The Psychological Corporation.

Werthheimer, N. (1923). Studies in the theory of gestalt psychology. *Psychologische Forschung, 4,* 301–350.

Yozawitz, A. (1986). Applied neuropsychology in a psychiatric center. In I. Grant and K. M. Adams (Eds.), *Neuropsychological assessment of neuropsychiatric disorders* (pp. 121–146). New York: Oxford University Press.

Zimmer, B., & Grossberg, G. (1997). Geriatric psychopharmacology: An update and review. In P. D. Nussbaum (Ed.), *Handbook of neuropsychology and aging* (pp. 483–507). New York: Plenum Press.

This Page Intentionally Left Blank

CHAPTER 11

"PEDIATRIC NEUROPSYCHOLOGICAL ASSESSMENT" EXAMINED

Jane Holmes Bernstein
Michael D. Weiler

INTRODUCTION

> "Any psychological theory that does not interface coherently with facts about the environment, evolution, and real experience will either be incomplete, wrong, or both..." (Hoffman & Deffenbacher, 1993, p. 336)
>
> "the most adequate description possible of a child's neuropsychological abilities and deficits does not lead...to an *understanding* of the child's neuropsychological status." (Rourke, Bakker, Fisk, & Strang, 1983, p. 113)
>
> "tests must be embedded within an overall assessment strategy that is organized according to current knowledge of brain-behavior relationships and the process by which these unfold over the course of development." (Tramontana & Hooper, 1988, p. 29)
>
> "even actuarial models can stand or fall as a function of the clinician collecting the data." (Willis, 1986, p. 247)
>
> "The analysis of research data, however, does not in and of itself provide the answers to research questions. Interpretation of the data is necessary. To interpret is to explain, to find meaning." (Kerlinger, 1986, p. 125)

In this chapter we address the topic of neuropsychological assessment of children. Our perspective is that of clinicians and of teachers training a new generation of professionals. We regularly read a variety of texts entitled "neuropsychological assessment of the child," "pediatric neuropsychological assessment," etc., and appreciate the often detailed analyses typically addressed in such texts, offering them as required, or recommended, reading for our students. We share the concerns, both theoretical and clinical, raised by our colleagues, and have learned from them and applied what we have learned to our own practice. We continue, however, to feel that something is missing from current discussions of the neuropsychological assessment of the child. Certainly, the (advanced) students we have the privilege to teach at the postdoctoral level do not appreciate all of the issues we wish to raise here. While they may be familiar with well-established basic principles of the measurement of human behavior (Anastasi, 1988), they have not yet acquired the skills to analyze how and when to apply them in the moment-to-moment interaction of the clinical assessment, at least not in the neuropsychological context. This analysis and application is, one can certainly argue, the stuff of clinical training at the postdoctoral level. However, lacking clinical experience and the intuitions derived from that experience, beginning clinicians rely heavily on what they read. Few texts address the nuts and bolts, the principles and strategies, of on-line clinical behavior; fewer still address these

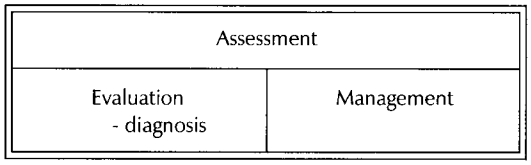


Figure 11.1. The Structure of Assessment

issues in the developmental neurobehavioral context.

We are also concerned that, even when students are equipped with the clinical skills needed to address the multiple components of a clinical assessment data set, in the face of the "uncertainty" inherent in a complex case, young professionals rely heavily on tests as uniquely valid and reliable tools in the description (diagnosis) of behavior. But the validity and reliability of the tests are by no means sufficient to do the job with which the clinician is charged. We agree strongly with Tramontana and Hooper (1988) that knowing the reliability and validity of individual tests does not mean that the validity and utility of a battery as a whole is known. We would go further, however, and argue that not only the validity and utility of batteries of tests should be examined for their appropriateness in assessing the child, but that the nature of the assessment process itself needs reexamining and updating in light of our expanding knowledge base in this area. It is this we wish to address here.

Different facets of the various issues we wish to discuss here have been examined in the context of clinical work with adults (Cimino, 1994; Lezak, 1995; Vanderploeg, 1994; Walsh, 1992). The principles they have highlighted are equally important when applied to clinical practice with children. They do, however, need to be scrutinized and reframed where necessary for the pediatric setting. Indeed, it is the challenge of neuropsychological analysis of behavior in the developmental context that has led us to explore the issues we present. We have two primary concerns: (1) that the assumptions underlying psychological assessment strategies in general are not necessarily appropriate for the *neuro*psychological assessment of children in particular; and (2) that neuropsychological (that is, biologically-referenced) assessment strategies for the child that do not integrate development—the cardinal feature of the child—at their core cannot be other than incomplete. These concerns seem to us to be particularly acute as the neuropsychological spotlight, initially focussed on the challenge of learning disabilities (Rourke, 1975), is now being trained with increasing intensity on the medical and neurological disorders of children (see Taylor & Fletcher, 1995). These disorders, long the province of pediatric psychologists and/or child neurologists, are now increasingly the topic of discussions of pediatric neuropsychology (Baron, Fennell, & Voeller, 1995; Batchelor & Dean, 1996; Hynd & Willis, 1988; Teeter & Semrud-Clikeman, 1997).

Two Problems

Our first problem in reviewing the literature on "neuropsychological assessment" in children is that there is little consensus regarding the application of the word "assessment" itself. "Assessment," "evaluation," and "testing" are used interchangeably by many writers (see, for example, Mattis, 1992) and the relationship of these to interviewing/record review, diagnosis, and/or management planning is not precisely specified. We agree with Matarazzo (1990) that (neuro)psychological assessment is not equivalent to (neuro)psychological testing and with Vanderploeg (1994) that it is a "complex clinical activity," one that involves the integration of information and data from multiple sources interpreted in light of a coherent conceptual model (Cimino, 1994; Lezak, 1995). Our use of the term is seen in Figure 11.1.

Assessment is the superordinate, theoretically-driven, clinical activity. It subsumes *evaluation* on the one hand and *management* on the other. The diagnostic strategy and formulation are components of the evaluation; history-gathering, observations and testing are components of the diagnostic strategy. The theory of the organism that guides the assessment must be the same for both the evaluative and management components: the brain that the child brings to the clinical session is the same brain that she or he takes into the world!

Our second problem in reviewing the literature is that assessment itself (that is, the assessment *process*) is not typically what is discussed in texts so titled. (Critical analyses of what students need to know about assessment are typically found under different labels: see, for example, Fennell & Bauer, 1997). Extensive discussions have laid the foundation for the emerging discipline by dissecting the assumptions of a neuropsychology of children, documenting the (many) hazards that clinicians face, reviewing (in increasingly fine

detail) the knowledge base of relevant medical, behavioral, and neuroscientific data, and/or providing detailed inventories of measurement tools and techniques (Baron, Fennell, & Voeller, 1995; Batchelor & Dean, 1996; Cohen, Branch, Willis, Weyandt, & Hynd, 1992; Dennis, 1983, 1988, 1989; Fennell, 1994; Fletcher & Taylor, 1984; Hartlage & Telzrow, 1986; Hynd & Willis, 1988; Pennington, 1991; Rourke, 1975, 1982, 1994; Rourke, Bakker, Fisk, & Strang, 1983; Tramontana & Hooper, 1988; Waber, 1989; Willis, 1986). However, it is not clear from these discussions *how* one should go about doing an assessment, how the various principles apply, or how the variously amassed data is used in the diagnostic process. What does not appear to have been questioned is the assumption that the clinician knows how to utilize the knowledge base and the measurement tools for the purpose of *neuro*psychological assessment of the child. The strategies and techniques of psychological assessment in general appear to be taken as some kind of fixed entity, independent of the population to be assessed; in the words of Slife and Williams (1997), "once techniques become established, they often have a life of their own, as though they exist apart from or are more important than the theories that spawned them." This is a charge that can all too easily be levelled at clinical neuropsychology today: too many would-be clinicians join a specific clinical team, learn their strategy/battery as part of their training, and then apply this in their own subsequent clinical practice. We challenge this. Professionals need to have a more active knowledge of their clinical behavior; they need to review what assessment is, what its limitations are, what its strengths can be; they need to view assessment, not as a rigidly fixed entity, but as a creative endeavor that can, and must, respond to new knowledge. It is a basic obligation of the training clinician to teach students these fundamental analytic principles, to demonstrate their application in the on-line assessment process, and to engage with them in ongoing dialogue about the value, precision and applicability of basic principles throughout their training experience.

Reexamining the Assessment Process

No meaningful assessment (measurement) of behavioral function can be undertaken independent of knowledge of the organism to be assessed. What one chooses to measure and how one decides to measure it is, importantly, a function of two things: the theory one holds about the way in which the organism "works" (typically derived from the knowledge base pertaining to the organism in question) and the available competencies of the organism at the point that its behavior is being measured. This being true, then any increase in understanding (modification of the theory) of the organism to be assessed mandates review of the *assessment process* (as distinct from the knowledge-base and the tools) in light of the new information. In this regard, pediatric neuropsychology is at a most exciting time in its development as a discipline: the increase in our knowledge of the developing nervous system and of the developing child is currently both rapid (see Dawson & Fischer, 1994; Krasnegor, Lyon, & Goldman-Rakic, 1997)—and ongoing. This increase from both psychology and the neurosciences in our knowledge of the organism under study thus mandates a reexamination of our methodology (and measurement tools) to do justice to the child who is the subject of our clinical analysis. This reexamination cannot simply entail updating of the tools, although we strongly endorse the call (Fennell, 1994; Rourke, 1994) for more appropriate normative data, and note the need to reexamine the tools in light of advances in measurement theory itself (Lowman, 1996). It is the nature of the assessment process itself that must be revisited.

The majority of the critical analyses of the assumptions underlying a neuropsychological assessment of the child have been framed in one of two ways: either in response to neuropsychological models derived from the study of adults, or in response to the widely used psychometric strategies that have been developed for the measurement of children's abilities within the child-psychology tradition. Neither of these frameworks can address the cardinal feature of the child, namely that it is a developing organism (Dennis, 1983; Fennell, 1994; Fletcher & Taylor, 1984). But, a true neuropsychology of the child, one whose goal is to understand the role of *brain* in the behavioral repertoire of the child, must grapple with the issue of dynamic change within and across both neural and behavioral domains as the child matures. And thus the reexamination of the assessment *process* (what to measure, how to measure) must be undertaken with the developmental character of the child as its focus. Our goal in this chapter is to examine various issues that bear on this more explicit integration of assessment and organism and to begin to

frame the discussion within which different solutions may emerge. The discussion has two main components: the first addresses issues relevant to the theoretical framework for clinical assessment in developmental neuropsychology; the second highlights methodological considerations and principles which follow from our initial discussion. In the final section, we provide the reader with the specific requirements of our approach to the assessment of the child.

ASSESSING THE CHILD: THE THEORETICAL FRAMEWORK

Neuropsychology and the Assessment of the Child

The neuropsychology of the child sits at the nexus of two traditions: those of adult neuropsychology and of the measurement of children's behavior (without reference to brain variables).

Neuropsychology, as a discipline, was initially defined in the context of adults with brain lesions. Localizationist models of behavioral function derived from these original observations have continued to exert considerable influence on subsequent model-building for both adult and child populations. This is in spite of the challenge of distributed functional models as postulated in the adult context (Damasio, 1989; Edelman, 1987; Mesulam, 1990) and of clear warnings as to the dangers of applying adult-derived localizationist models to children's performance (Fletcher & Taylor, 1984; Fletcher, Taylor, Levin, & Satz, 1995; Taylor, 1988). There are two primary ways in which neuropsychological principles can be applied: (1) neuropsychology can be used as a dissecting tool, to cleave the overall behavioral domain into its behavioral subcomponents; (2) neuropsychology can provide a framework for exploring theories of human brain-behavior relationships. In the first instance, it is not necessary to know what the neural substrate is (or might be) for the observed behavior; it suffices that, in the presence of brain insult or disease, behaviors are differentiated. Take the example of the patient who is asked to name a picture of a hammer. After repeated attempts and with examiner-cuing, he is unable to find the word "hammer." Asked what the object does, however, he eventually responds: "it's to hammer with." The observation is of itself evidence for differential processing (storage?) by the brain of nouns and verbs (it is also a particularly elegant one in that *hammer* is one of the relatively few noun-verb pairs whose surface forms are identical). See Caramazza & Hillis, 1991; Damasio, 1990; Warrington & Shallice, 1984 for other examples of category-specific retrieval failure.

The second instance is exemplified in an analysis conducted by Butters (1984): in a comparison of performance by patients with Huntington's disease and Korsakoff's syndrome, he highlighted the double dissociation of skill learning and verbal recognition. Here, not only do the data provide evidence for differential processing by the brain of memory for skill as contrasted with memory for knowledge, but also, given independent knowledge of the neuropathology of the two disorders, the data can be utilized in the generation of hypotheses relating to specific brain-behavior relationships. The goal of "double dissociation" is sought for precisely the purpose of identifying regular relationships between brain systems and behavior.

In the assessment of the child, the "dissecting-tool" role of neuropsychology has predominated over the last twenty years—in spite of the continuing need to assess children with biologically-based disorders. One might reasonably ask why. There are two strong reasons for this. One, the recognition that constructing a brain-behavior relationship model in a manner comparable to that used with adults is, at the very least, challenging and, indeed, may not be possible (see the problem of brain-behavior isomorphism: Fletcher & Taylor, 1984) appeared to divert attention away from brain-based modelling. And two, the ability to dissect the behavioral domain in increasing detail provided by (adult) neuropsychology proved to be an excellent (even seductive) match with the extensively employed methodology of the established tradition of measurement of children's abilities with its focus on cognitive abilities. Models of assessment with increasingly detailed characterizations of the child's cognitive ability structure at their core have been deployed for the clinical analysis of the child's behavior within a neuropsychological framework.

The measurement of children's abilities has been undertaken in the child-psychology tradition. Its goal has been to use psychological measures to rank children one against another and to characterize their cognitive profiles as the basis for educational placement, instruction, and—in the case of academic difficulty—remediation

strategies. The focus of this tradition, in the applied setting, has been measurement of ability and cognitive functioning. However, the analysis of children's behavior has also been the focus of another tradition in psychology, that of developmental psychology. This has focussed on the nature of children's abilities from the perspective of acquisition, integration, and consolidation of skills (cognitive, social, emotional, and regulatory). The emphasis has been not on *what* but on *how* and *when*. Although each tradition has influenced the other, they remain separate with different assumptions, biases, and goals. The strengths of child-psychology tradition lie in its psychometric rigor and its knowledge base. Its contribution to assessment lies in its potential for diagnostic classification. It is, however, limited by its focus on the here and now, the snapshot view, the "horizontal" analysis[1] and by the fact that interventions are content-based, not child-centered. The limitations of the developmental tradition include its lack of normative reference and the relatively limited applicability of its assessment tools. The strength of the developmental approach is, however, its "vertical" perspective, its focus on the processes of change that are integral to the developmental character of the organism and culminate in the competent child at each stage in its growth. Its contribution to assessment may be most powerful for management where the emphasis on the *how* of performance provides a basis for individualized, child-centered intervention strategies and where the longitudinal perspective provides for the prediction of risk.

A comprehensive assessment of the child must take advantage of the knowledge and techniques of both the child psychology and developmental psychology traditions, as well as that of neuropsychology and of the neurosciences. It cannot restrict itself to discrete cognitive skills, or brain systems, as its primary focus, but must take the child as the "unit of analysis" in order to address (1) the whole of the behavioral repertoire that the child's brain makes possible, and (2) the full range of contextual transactions (social and environmental) in the course of a child's life that elicit, facilitate, maintain, and/or modify the way in which the workings of neural mechanisms are manifest in behavior.

Basic Assumptions

In light of the rapid increase of knowledge relevant to the neuropsychological understanding of the child, we have reexamined our core assumptions. These are now as follows:

1. Assessment does not stop with evaluation and diagnosis; it necessarily includes management and intervention. The theory (of the child) that guides the clinician to a diagnostic statement is, optimally, the theory that permits the principles of management to be outlined.
2. Clinical assessment is not different from scientific enquiry (Fennell & Bauer, 1989, 1997; Pennington, 1991). The ability of the study (clinical assessment) to answer relevant questions depends on the research design (Kerlinger, 1986). Any update of the assessment process must start with the appropriacy of the design.
3. In the clinical assessment, the child, and not the cognitive ability structure or a brain system, is the unit of clinical analysis.
4. The assessment is guided by a conceptual model (Cimino, 1994; Lezak, 1995). The (minimum) components of this model are: the theory of the organism to be assessed (in this case, the child); the theory of the (relevant) disorders; the theory of the assessment process.
5. For *neuro*psychological analysis, the theory of the child must incorporate "*brain*" as a fundamental variable.
6. The brain neither develops nor is maintained in isolation. Both it and the behavior that it supports are context dependent.
7. The theory of a developing organism must incorporate principles of development. Thus, the assessment process itself must incorporate these principles. The observed behavior at any age is conceptualized as an outcome of the developmental course to date.
8. The behavior (and behavioral development) of children is shaped by their transactions with adults. The (adult) clinician is thus an integral component of the assessment and his or her behavior must thus be subjected to formal scrutiny.

Issues related to the above assumptions are reviewed in the remainder of this chapter.

Examining Assumptions

In attempting to formulate models of assessment that conform to these initial assumptions, several important issues must be addressed. The first is the "unit of analysis" at the core of the assessment strategy, that is, the frame of reference within which the clinical analysis is undertaken. The second relates to "brain": how is *brain* to be assessed/measured? The third derives from the need to appreciate that the brain belongs to a developing organism. The developmental course must be an integral component in the analysis of current behavior (Segalowitz & Hiscock, 1992)—with the important corollary that any perturbation (neurological or psychological) in development will also be incorporated into the developmental course and will potentially change the relationship between brain and behavior at subsequent points in time (Teuber, 1974).

The fourth issue concerns the "tools of the trade," psychological tests. There are three points to be made: (1) Many of the tests used by neuropsychologists were originally developed for use in psychological contexts for the measurement of children's abilities. As such, they carry with them enormous sociopolitical "baggage" (see Ceci, 1996). (2) The objectivity of tests as the "measure of man" (or child, male or female) is seriously overvalued (Matarazzo, 1990). (3) There is a range of behaviors that are "invisible" to current psychological tests, but are nonetheless critical to brain-referenced formulations of the theory of an individual child (Bernstein & Waber, 1997).

The fifth issue is that of the interaction of an adult with a child in both clinical and natural settings and the failure of many theoreticians of applied pediatric neuropsychology to analyze the role of the clinician in the observed behavior of the child. We note, as will be made clear below, that the role of the clinician as a diagnostic decision maker has been discussed in detail, and in the neuropsychological context. Our concern is with the integration of the fruits of this discussion into the formulation of assessment models matched explicitly to the organism under discussion.

The Core "Unit of Analysis"

Different approaches to neuropsychological assessment have, for both adults and children, been framed in terms of the data-collection technique. Thus, assessments are characterized as involving fixed or flexible batteries or patient-centered strategies (Bornstein, 1990), relying variously on quantitative or qualitative data (albeit, in most instances in pediatric practice, a combination of the two (Batchelor, 1996; Rourke, 1994). Not surprisingly, this reflects the centrality of data collection in the thinking of neuropsychologists.[2] Thus, in the most influential models of the analysis of children's behavior (Fletcher et al., 1995; Taylor & Fletcher, 1990; Rourke, 1994), the product of the data-collection techniques, that is, the cognitive ability structure (CAS) of the child, is the core of the diagnostic methodology. Indeed, variants of this strategy, albeit with some modifications, appear to be the most widespread approach to the neuropsychological assessment of the child today.

In the clinical context, however, we believe that this presents a significant problem. The formulation of clinical hypotheses does not typically await test- data collection—unless one rigorously employs the kind of data collection characterized by Rourke (1986) in which the interpreter of the psychological-test data does not collect it, and, initially, does not have access to historic and developmental data. The natural course of a clinical interaction (with or without a technician who collects data) involves meeting the patient and his or her family, knowing the index symptom that brings them to the clinical situation, interviewing, taking the history, reviewing medical records, and so forth. These sources all provide information on which hypotheses can be generated (Cimino, 1994). The clinician can all too easily have formed a hypothesis based on any or all of this information without realizing that she or he has done so. These hypotheses must, however, be scrutinized—as an integral part of the assessment process—not only for their diagnostic potential, but also for the possible bias that they may contribute. It is important to note that all assessment methodologies are vulnerable to bias. Bias inherent in the anchoring-and-adjustment heuristic (see below), for example, is a problem that must be addressed by all practitioners (we all have to start somewhere!). In the CAS-centered analysis, the data typically collected from interviewing, record review, and so on will have controls for bias imposed; this will presumably, however, be a separate step from the deployment of controls for the expectable biases that may influence the CAS data. These data sets will then need to be integrated separately at the interpretive level of the assessment. In the child-centered approach,

controls must be deployed on all relevant data, as it is collected, in an ongoing integrative fashion as the assessment proceeds. (Hypotheses will also be generated, and tested, in response to the theoretically-driven observation of behavior over the course of clinical data collection.) Thus, formally identifying the child as the unit of analysis requires the clinician to function in a rigorous, methodologically-sound fashion from the beginning, framing hypotheses and setting up initial tests thereof (where appropriate) in the "natural" sequence, and minimizing bias in a proactive fashion.

Placing the child at the center of the clinical analysis has other important implications. For us, the goal of the assessment is optimal adaptation of the child to the demands upon him or her (Bernstein & Waber, 1990, 1997). Such demands are never restricted to academic skills, although, to the extent that school is a child's "job," optimizing adaptation in the educational context must be a major goal in working with a child. In this regard the child is not wandering around with "holes" in his or her cognition; she or he functions as an integrated organism with different capacities for solving the varied challenges/demands of his or her particular environment. His or her neuropsychological "package" may limit the way in which these demands can be met, but may also permit compensation, either by use of alternative skills or by matching environmental demand to skills more effectively. Thus, an individual may use high motivation, good organizational capacities and well-developed conceptual skills to compensate for specific reading deficits (Fink, 1996), or may seek an alternative route to the same goal (keyboarding instead of [laborious] handwriting to complete writing assignments, more rule-bound and nonspoken Latin instead of French or German to fulfill the language requirement for college application).

With adaptation, rather than remediation, as the goal of the assessment, the output of the diagnostic process cannot be limited to identification of deficits, but—importantly—must include characterization of competencies. Identification of deficits may be necessary for diagnosing particular disorders (although, even here, the different patterns of strengths that co-occur in the diagnostic behavioral cluster will influence the interpretation of deficit performance), but knowing the child's strengths, that is, what she or he has to work with, is prerequisite to developing effective interventions.

This emphasis on competencies is fundamental to our approach. It is important not only clinically, that is, for intervention and management, but also theoretically. To the extent that the model driving our clinical behavior is couched in brain terms, it is important that we be able to generate and test brain referenced hypotheses of not only what the child cannot do, but also what she or he *can* do. The clinician must be able to account, in neuropsychological and/or psychological terms, for both aspects of performance. After all, given a problem sufficient to warrant a neuropsychological diagnosis, it is even more important to understand how successes are achieved and/or strengths are supported. Thus, to make the (strong) claim that, for example, left hemisphere mechanisms are implicated in (even the source of) deficits X and Y requires the clinician to provide the complementary explanation of how, given impaired left hemisphere mechanisms, the child is able to mobilize strengths A and B. Without the ability to explain both aspects of performance *in the same theoretical framework* the neuropsychologist cannot claim to have adequately characterized the child's brain function or neurobehavioral repertoire.

The child-centered analysis also influences the formulation of the initial questions that guide the assessment process. In line with the goal of the assessment, optimal adaptation, the initial questions are framed in "normal," not deficit, terms: "What is a child of this age able to do?" "What demands are expectable?" rather than "What is wrong with this child?" Identification of disorder is, of course, a necessary part of the assessment; it is important both for understanding the whole child as well as for obtaining appropriate treatment and services. In a whole child approach, however, it does not frame the analysis.

Incorporating "Brain"

The biggest single challenge to the formulation of a neuropsychological theory of the child is the question of the basis on which to establish relationships between brain and behavior. The problem is that of criterion-related validity: what external "brain" criterion can be used to establish such relationships in an organism in constant flux not only with respect to developing brain structures and emerging behavioral competencies but also in the relationship between them? (Chelune & Edwards, 1981; Emory, 1991; Emory, Savoie, Ballard, Eppler, & O'Dell, 1992; Pirozzolo & Bonnefil, 1996).

The criterion-referenced validity problem has been well recognized. Fletcher and Taylor (1984) have taken a strong position in this regard and argued that, since *brain* poses an as-yet-intractable problem in the pediatric context, analysis of behavioral function (albeit guided by neuropsychological principles) must be the basis for diagnosis and intervention. As noted above, this general position is currently widespread as the basis for the neuropsychological assessment of the child. While we appreciate the conceptual issues raised by Fletcher and Taylor and others, we do not agree that *brain* cannot be examined and strategies for incorporating it directly into assessment methods cannot be formulated, albeit at a beginning level. Indeed, we would argue that the criterion-related validity problem is inherent in the mismatch between "static" measurement strategies and the inherently dynamic nature of a developing child (a problem which is not going to go away any time soon), and thus that it behooves us, as *neuro*psychologists, to begin to devise assessment strategies that are more appropriate to the nature of the organism, and then to subject them to formal examination in the service of the child.

Perhaps the most important reason for taking this stand for us, as practicing clinicians, is the fact that the level at which a clinical analysis of behavior is made makes a difference for management planning and intervention. A diagnosis at the level of achieved skill or psychological function (or process) alone does not allow principled predictions about those nonachievement skills or behavioral functions that are not included in the diagnostic statement. We emphasize the word "principled:" it is our impression that clinicians can and do make effective recommendations for the child's general adjustment in the world, but are doing so based on their knowledge of general psychological principles, their "clinical" know-how, and/or plain (but unexamined) "common-sense" (which can be seriously flawed: Rabinowitz, 1994). They do not necessarily, however, integrate non-test elements of the child's overall behavioral repertoire into their neuropsychological analysis in principled fashion and thus may lose the opportunity to both predict a relevant range of outcomes and evaluate their success in so doing.

Take the example of a child who presents with reading difficulties. By means of formal analysis of cognitive abilities and academic skills, the neuropsychologist demonstrates that the problem is seen in the context of language problems. Given the available literature on this relationship, she or he may provide diagnoses of specific reading disability and language disorder. These diagnoses are at the manifest behavior (achievement) level and the psychological process level, respectively. Recommendations derived from them would, presumably, address the reading skill, using the nature of the language-processing deficit as a guide to promote more focussed intervention.

In a "whole child" model (Bernstein & Waber, 1990, 1997) that scrutinizes the behavior of the child within a *brain-context-development* matrix (Bernstein, 1999), however, the child's presentation would also be explicitly analyzed at the neuropsychological level. This entails invoking postulated brain substrates (both intact and atypical/dysfunctional) for the behavioral "package" observed. These postulated brain systems are derived from the knowledge base of neuropsychology in general (human, nonhuman, adults, children/young animals). In this model, the clinician seeks to identify clusters of behaviors whose congruence is determined by the theory of the organism guiding the assessment. (See Waber, Bernstein, Kammerer, Tarbell, & Sallan [1992] for a research application of this model.) Management and interventions are then guided by what is known about the function of the implicated brain system. Thus, viewed within the *brain-context-development* matrix as a "neuropsychological-psychological layercake" in Martha Denckla's colorful description (Denckla, personal communication, 1979), a child with language deficits implicating left hemisphere mechanisms (*brain*) would certainly be seen as being at risk with respect to academic skills presumed to be dependent on intact language processing capacities. He or she would also be considered at risk with respect to (1) the integrity of other functional skills thought to depend on left hemisphere mechanisms (such as graphomotor control, managing details in complex arrays, etc.); and (2) the continued development of skills in the face of increasing task demands (*context*) that require the integration of functional capacities subserved by the left hemisphere. In addition, the risk analysis would include consideration of (3) the skills needed in the wider *context* of childhood language use (listening skills in the classroom, following directions in all situations, participating in peer conversation); and (4) the availability of the linguistic foundation necessary to support the *development* of both higher-order linguistic capacities

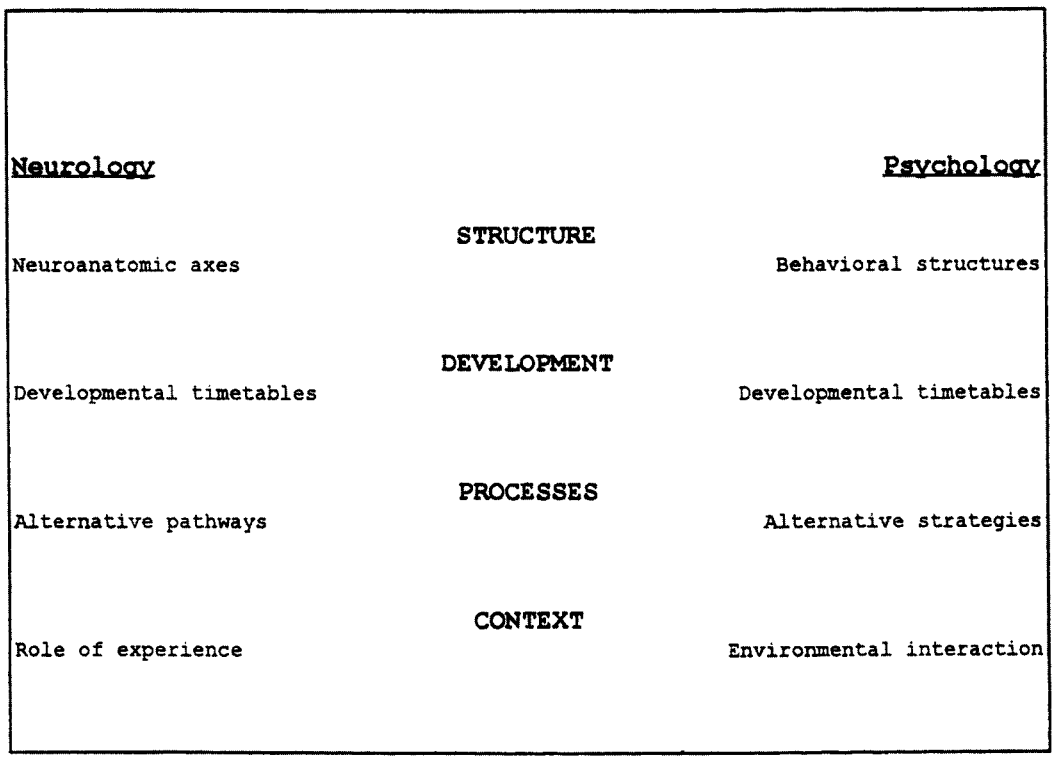


Figure 11.2. The Developmental Neuropsychological Model

Note: Adapted from Bernstein & Kaber, 1990.

(the metalinguistic/metacognitive skills required in the academic context) and psychosocial competencies (such as anger regulation in younger children or social interaction in latency and adolescent peer groups). Thus, even behaviors leading to a presenting complaint of depressed mood, for example, (not always, or even typically, viewed as "neuropsychological") would be scrutinized for their possible neuropsychological—as well as psychological—underpinnings: is the child withdrawn and isolated (as reported) because she or he has difficulty following the language of peer-group interaction (which now demands more active participation "on line"); or is she or he struggling with significant emotional issues that lead to impaired psychological functioning? These are, unfortunately, not independent: many children with neuropsychologically-based difficulties experience related distress that leads to emotional disorders.

In the "whole child" model, risks are both "horizontal" (to be seen in the child's current repertoire) and "vertical" (likely to have impact in the future). They may be manifest in either or both academic (work) and social (play) spheres, and in terms of contextual demands and/or content-referenced skills. They must be evaluated systematically for the child in question. Recommendations are then developed, in principled fashion, specifically to address each of the applicable risks, including *context* and *content* issues in both academic and non-academic/social settings of the child's life, now and in both the short and long term.

The Role of Development

A developmental perspective is both crucial to a brain-referenced neuropsychological model for the child (Baron et al., 1995; Tramontana & Hooper, 1988)—and not yet integrated into pediatric assessment strategies. As noted above, assessment approaches modelled on those of adults will not do: the parent discipline of a neuropsychology of the child cannot be that of adult neuropsychology extended downwards to chil-

dren of different ages; children are not small adults. Thus, a neuropsychology of the child cannot be derivative of that of its elders (relatively stable and modular), but requires its own formulation, matched to the child's (intrinsically dynamic and plastic, developing, not yet stable) "brain" parameters. The basic tenets of a developmental neuropsychology are (1) that it be developmental, and (2) that it incorporate in its analysis both the psychology and the neurology appropriate to itself (see Figure 11.2). The parent disciplines of this emerging discipline then are those of the developmental neurosciences—psychology, neurology, and neurobiology (Bernstein & Waber, 1990).

The challenge of incorporating development into our understanding of neurobehavioral competencies of the child is enormous. We do not underestimate it—but neither do we think it needs to be (or should be) deferred. Indeed, we believe that pediatric clinicians faced with patients with neurological, psychological, and behavioral disorders are already, as a daily consequence of trying to apply psychological and neuropsychological principles to their work, building models that attempt to incorporate developmental principles and manage the complexity that this entails. The assessment challenge is to specify the framework and relevant variables so as to be able to formulate testable hypotheses that can be shared with, tested by, and improved upon by, other clinicians. The theoretical challenge is that of modelling a dynamic system in which on-line behavior in response to shifting (and multi-level) demands in the environment is supported by changing neural structures interacting with emerging functional skills. Both brain systems and psychological/behavioral systems are in the process of maturation and both involve multiple subsystems with multiple components, each of which may be on a different maturational schedule. Different neural systems become available at different times (Conell, 1939-1963; Luria, 1973; Rodier, 1994; Spreen, Risser & Edgell, 1995; Thatcher, 1992; Yakovlev & LeCours, 1967). Behavioral systems also come on line at different times with different rates and timing of acquisition of given behavioral systems (Diamond, 1991; Dennis, 1988, 1989; Schneider & Pressley, 1990). (For example, neural systems required to "parse the grammar" of the visual-spatial world, that is, determine which perceptual stimuli group together as objects and which do not, are presumably mobilized prior to those involved in the acquisition and development of the language which permits the objects and their interactions to be talked about.) These various developmental and/or acquisition schedules may differ as a function of gender (Waber, 1976) or laterality (Carlson & Harris, 1985; Trevarthen, 1996). They will also be differentially derailed in the context of insult—as a function of the age at time of injury/diagnosis, location of injury, type of disorder (see Fennell, 1994), as well as type/duration of treatment (see Waber & Tarbell, 1997). The stage of acquisition of a given skill can also lead to differential response to insult (see Figure 11.3)—which may be further shaped by the natural history of the particular biological insult/disease process in question. The expression of both neural and behavioral structures is likely to be rendered even more complex by processes subserving plasticity in the CNS (Kolb, 1989, 1995). Neuropsychological models of the developing child must be cognizant of this dynamic complexity in order to be able to assess the outcome of the developmental process at any given point in time (Chelune & Edwards, 1981; Segalowitz & Hiscock, 1992). Clinicians will need to expand their knowledge base accordingly (Fennell, 1994).

We should note that the issues of brain and development are explicitly recognized in discussions of the neuropsychology of the child (Baron, 1995; Batchelor, 1996; Cohen, 1992; Fletcher & Taylor, 1984; Rourke, 1994; Hynd & Willis, 1988; Taylor & Fletcher, 1990; Teeter & Semrud-Clikeman, 1997). Indeed, Tramontana & Hooper (1988) have clearly articulated the need to incorporate developmental constructs into the next generation of assessment models. However, as yet, the relationship of brain and development to the assessment process has not been clearly delineated (if addressed), nor has the theoretical framework been specified within which testing, interviewing, the diagnostic process, and management and interventions are (presumably) related. These critical basic constructs must be incorporated as integral elements of assessment design in order to promote the ongoing development of the field.

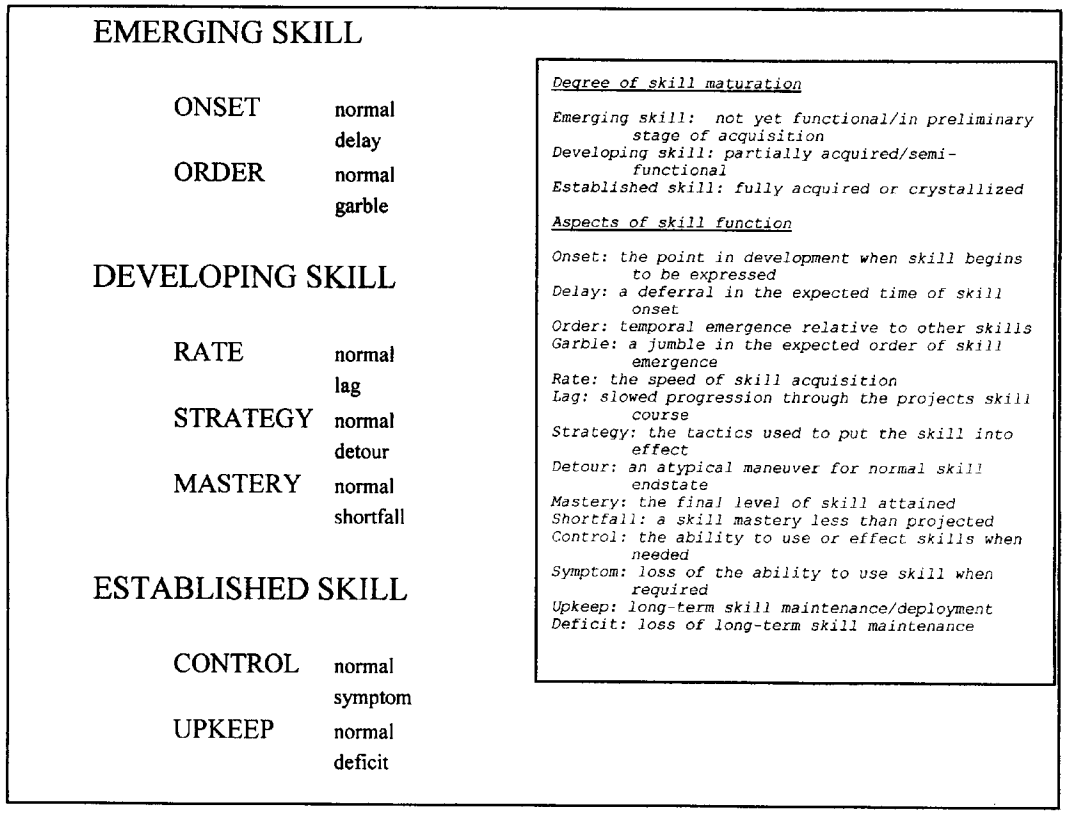


Figure 11.3. Developmental Course of Skill Acquistion
Note: Adapted from Dennis, M., 1988, 1999.

Psychological Tests

The Context

In pediatric neuropsychology, the use of psychological tests cannot be disembedded from the sociopolitical context within which the measurement of children's abilities takes place. The problem lies in the fact that, when utilized for neuropsychological analysis, tests bring along with them the particular biases, inherent errors and sociocultural ramifications of the psychometric, child-psychology tradition. Pediatric neuropsychologists are all too well aware of the constraints thus imposed: the practical problem of having, in an educational team meeting, to explain the neuropsychological analysis of a child to educators whose frame of reference for the use of the tests is very different can take every clinical "reframing" skill one has (see Bernstein, 1996a).

In the adult context, in contrast, an instrument such as the *WAIS-R* (Wechsler, 1981) is a member of the general psychometric armamentarium. It can thus be deployed simply as one measurement tool among others, a means of establishing the context of general ability within which the neuropsychological analysis of specific behaviors takes place. For the adult it measures where the individual is *now*. Note that, for the adult practitioner, neuropsychological assessment is typically requested when there is reason (often secondary to known insult) to question whether and to what degree the patient is functioning at a reduced level, and/or in a different manner, from that previously achieved. The clinician's role is to intervene to maximize current and ongoing adjustment in an adult whose "personhood" was established prior to the insult that brings him or her to clinical atten-

tion. (Although personality can be changed by brain injury, clinicians are not—not yet, at least—expected to be able to reverse the impact of insult and "re-change" personality.) This is not the case in the pediatric context. IQ tests used with children raise a daunting specter for the responsible clinician: the assignment of an individual to a particular category of IQ may affect not only educational decisions now, but also—and more importantly and possibly perniciously—the development of person and the achievement of particular life outcomes in the future. In the wider societal context this is confounded by the widespread assumptions that (1) what IQ tests measure is "biological intelligence" (a belief that, apparently, will not be laid to rest [Herrnstein & Murray, 1994; but see Ceci, 1996; Gould, 1996; Neisser, Boodoo, Bouchard et al., 1996]) and that (2) differences in this IQ-test-based biological intelligence may differentiate groups in society—usually to the disadvantage of one or another. In children, the results of psychological measurement sit not only in this wider sociopolitical context, but also have specific impact on educational policy and practice, both theoretical and fiscal, as manifest in federal and state educational regulations and funding. Because school budgets depend not only on state funds but also on tax levies at the community level, the distribution of monies to children with different educational classifications (i.e., gifted and talented, slow learner, learning disabled, mentally impaired, "average") can be an intensely political issue at the micro, community, level as well as at the macro, societal, level. It is the use of psychological tests in this context that presents a challenge to the neuropsychologist. The tests have a life of their own (to the extent that the tail of the measurement tools wags the dog of the trained clinician), are all too frequently thought to define behavior "biologically," and may generate intense emotion in parents, educators, policy makers, and theoreticians. Using them in an "objective," scientific fashion is challenging at best, fraught with pitfalls at worst. Nonetheless, the neuropsychologist has no real option but to use such instruments because so much of the educational system's assessment of children is essentially driven by them. Nor can the neuropsychologist realistically take a "purist" stance and maintain the argument that his or her use of psychological measurement instruments is untainted by the social, emotional, and political context of measurement of children's abilities. The tools cannot be used in a manner that is independent of the effects thereof; the clinician cannot fully control the way in which the results will be used in the wider society and the neuropsychologist must work with the consequences of this. The past practices governing the use of psychological tools within the child psychology tradition can thus, all too easily and unexamined, influence—rightly or wrongly—the practice of neuropsychology itself.

The Objectivity of Tests

How objective are psychological tests? How objective can they be? The authority conferred on psychological tests by the body of psychometric theory that supports them and by their (now extensive) history and wide application has led to significant over-confidence both in their objectivity in characterizing human behavior in general (see Matarazzo, 1990) and in their utility when applied to the neuropsychological setting. Tests may be objective, reliable, and valid in their own terms without their use by a clinician being so (Willis, 1986). Nor are they employed in a vacuum; they and their administration constitute a context within which the child and the examiner behave and interact—with the potential for expectable biases and influences one on the other (Banaji, 1996; Greenwald & Banaji, 1995; Sadker & Sadker, 1994).

Test-Invisible Behavior

For neuropsychological analysis, an over-reliance on psychological tests has a further problem: behaviors that may be critical to the elucidation of brain-based contributions to the clinical interpretation may be invisible to them (Bernstein & Waber, 1997). For example, no psychological test is currently available to measure a child's level of general arousal—but a clinician cued by the assessment strategy (Bernstein & Waber, 1990; Bernstein, Prather, & Rey-Casserly, 1995; see also Batchelor, 1996) to consciously observe arousal level has little difficulty in recognizing low arousal and slowed performance—and including hypotheses pertaining either to the potential contribution of brainstem or other subcortical involvement to the presenting complaint, or to the possibility of insult at a time when systems critical to the efficient main-

tenance of the behavior in question are developing—or both. (Note that, in principle, such a test could be devised with means and standard deviations, etc., but arousal is an "on-line" behavior that supports all others in ongoing fashion. It is manifest in the context of other behaviors (i.e., conversation, test taking, between-test interactions). While a formal test of arousal might provide data for a specific point in time, the use of changes in arousal as a diagnostic indicator requires a strategy that can characterize the behavior over a longer time course (Light, Satz, Asarnow, Lewis, Ribbler, & Neumann, 1996).

The Clinician (Adult)/Child Interaction

Models of developmental neuropsychological assessment must address the role of the clinician explicitly.[3] Why? First, as demonstrated in the context of physics as the so-called "uncertainty principle" of Heisenberg and Bohr, no observation is independent of the observing instrument (see Globus [1973] for an elegant interpretation of this principle in the behavioral context). In the clinical context, the "observing instruments" are not only the tests but also the clinician. Second, our goal is (ultimately) a *neuro*psychological understanding of children's behavior. Brain is necessarily a critical element in the undertaking. But brains neither develop nor function independently of the environment in which their owners find themselves; the matrix in which they are embedded is that of *brain-context-development.* Understanding context (and change over time) is thus crucial to understanding brain. Adults (including clinicians) are elements in the child's context; their role must be examined in light of the theoretical approach to the assessment process. (As noted previously[3], the types of test-cognition models that are the core of behavioral measurement in the positivist tradition do not [and cannot] acknowledge the necessity of addressing the historical and contextual requirements of a brain-based analysis and thus need not consider such contextual variables as the role of the clinician.) Third, over development, the adults in the child's context interact with the child in a manner that both promotes and constrains the way in which behavior is learned and is manifest. The evolutionary importance of this ongoing transaction between child and adult, not only for the development of the child but also for the benefit of society as a whole, argues for powerful "child-supporting" behaviors in adults. It is our position that this behavior—in the adequately socialized adult—is (inherited and learned) automatic and is thus undertaken "unconsciously" in all interactions with children. It is thus a critical feature of the context in which the brain functions at any age (the child's brain is supported (shaped) by the adult's brain). This inevitable component of the interaction in the psychological-testing situation cannot be eliminated without changing the very nature of the behavior under observation.[4] To the extent that it cannot be eliminated without severe disruption of a child's behavior (Draeger, Prior, & Sanson, 1986), the nature of the interaction between child and adult at different times and in different contexts must be subject to as much and as detailed analysis in the assessment of the behavior of the child as any other aspect of the clinical activity.

Models of Assessment in Children

With the developmental variable in mind, approaches to the neuropsychological assessment of children can be grouped into three general categories, broadly conceived. These derive from the fact that practitioners work with different populations and/or come from varying theoretical traditions. An initial distinction is one that has been made elsewhere in the context of epilepsy management in children (Bernstein et al., 1995): the distinction between presenting problems seen in the context of a grossly regular developmental course ("on developmental track") versus presenting problems seen in the context of noticeable deviation from expected developmental progress ("off developmental track"). Thus, educational and school psychologists, as well as clinical and neuropsychologists, who by and large deal with specific learning failures in school (learning disabilities) in students who are "on developmental track" are likely to approach neuropsychological analysis from a different perspective than clinical, pediatric, or neuro- psychologists who address the perturbations in development that result from direct disruption of brain growth and development secondary to insult in utero, infancy, childhood, etc. (structural anomalies, neurological, genetic and metabolic disease, severe prematurity/inter-uterine growth disorders, toxic exposures, etc.).[5] (The "deeper" the presumed source of the behavioral deficit(s) in terms of timing of influences on brain

formation and development, the more likely the child will be off developmental track.)

The second distinction derives from differences in theoretical perspective and the differences in focus of clinical analysis that ensue. Is the analysis "horizontal," focussing on the child's current behavioral repertoire, or does it incorporate developmental variables? If the latter, does it integrate appreciation for developmental *context* in the analysis of the current behavioral repertoire, or does it require that development be incorporated in the description as a "vertical" dimension and treat current behavioral functioning as an *outcome*? These three approaches we characterize as *cognitive ability structure, normative developmental,* and *systemic developmental models*, respectively.

Cognitive-Ability Structure Models

A "pure" cognitive- (or behavioral-) ability structure approach relies heavily on adult neuropsychological models, on psychological measures extended downwards to children of different ages, on interindividual ranking and on cognitive profiling. It treats behavioral functions as modular, make diagnoses in terms of specific cognitive deficits and/or perceptual modalities, and designs interventions in terms of remedial strategies and/or teaching to strengths defined in terms of discrete behavioral functions.[6]

Normative Developmental Models

These are widely practiced in pediatric neuropsychology. They require the administration of psychological tests to obtain a description of the child's cognitive ability structure. This description is then interpreted with reference to the context provided by the child's developmental, demographic and/or socioeconomic status. Developmental variables are clearly recognized—those that are derived from knowledge of the natural history of a given medical condition, and those that differentiate children of different ages, both of which operate as moderator variables. The biology and ecology of the behavior of eight-year-olds is recognized as different from that of five-year-olds, or 10-year-olds, or 15-year-olds, and is so used in clinical analysis. The interpretation of specific behaviors, the role of moderating variables, and the diagnostic inferences made, differ accordingly.

Systemic Developmental Models

These take the "whole child," rather than cognitive ability structures/profiles, as the focus of behavioral analysis. Their primary assumption is that the behavioral repertoire (at any point) is an *outcome* of the child's developmental course to date. The clinical analysis thus incorporates the developmental perspective as an integral component from the beginning: the neurobehavioral repertoire of the eight-year-old cannot be understood independently of his or her functioning at one, two, three, four, five, six, and seven years of age (to the extent that this can be ascertained or inferred). Rigorous history-taking by means of careful interview (with controls for bias in reporters), appropriately designed questionnaires and detailed record review (where indicated) are crucial to this approach, as is knowledge of the developmental course of possibly relevant medical, psychological and/or behavioral conditions—and are accorded detailed attention in the training of clinicians.

The models differ in important respects with regard to the goals and formulation of management and intervention. Approaches with the child's cognitive-ability structure at their core are likely to focus on specific skills and offer targeted remediation in an effort to approximate the skills of age peers. In contrast, the explicit focus of the systemic developmental model is on the "whole child" (Bernstein & Waber, 1990) and intervention is aimed at optimal *matching* of the child, as an integrated organism, with the ongoing challenges of childhood, adolescence and adulthood. The goal is expressly the "comfortable, competent 25 year old" (Bernstein, 1996a). The management of a given child in terms of medical/neurological referral, classroom placement and/or instructional programming may differ minimally, if at all, among experienced clinicians working in either of these models. Where indicated, "normative-developmental" practitioners are likely to go beyond the strict cognitive-ability-structure format to make broader, psychologically- and/or developmentally-relevant recommendations. These are not necessarily, however, intrinsic to, and formulated within, the *neuropsychological* analysis of the child's situation, but are likely to reflect an amalgam of neuropsychological *testing* and clinical psychological management strategies. The different approaches are likely to differ, for example, with respect to (1) the goals and orchestration of the feedback or informing session and (2) the nature of the predic-

tions. In the systemic model, with its emphasis on the whole child, the central educative function of the feedback session will address the given child's well-being and adaptation and thus will be framed within the *brain-context-development* matrix appropriate to the organism. Predictions will then be derived from consideration of this same matrix, that is, as time passes, contextual demands will change (e.g., the child in relation to family, peer group, grade expectations, curriculum requirements), new skills will be expected to emerge (e.g., maturation of executive control processes), and the child's brain will change in response. From this model one can anticipate the new challenges that will be presented to the child (e.g., increased expectations for independence in social and academic functioning, greater organizational demands in school, increased language demands in peer relationships) and anticipate how she or he will fare. Without consideration of these factors prediction can only address cognitive ability structures (i.e., the child who has difficulty with reading will continue to have difficulty with reading and written language-based tasks).

This type of formulation (and feedback session) contrasts with one centered on the cognitive ability structure which provides a detailed discussion of the child's skill profile with recommendations for specific interventions in relevant settings and predictions focussed on primarily academic intervention needs.

Model Differences

To highlight the differences between what we have called *normative developmental* and *systemic developmental* models-assessment approaches, we have analyzed the models of a selected group of our colleagues in terms of the assumptions highlighted above.

Since his formulation of an initial research program in 1975, Rourke and his many colleagues have had a major influence on the development of a neuropsychology of the child (Rourke, 1975, 1982, 1994; Rourke et al., 1983; Rourke, Fisk, & Strang, 1986). Examining their approach from our perspective, we note the following: the unit of analysis is the cognitive ability structure (CAS)—not the child. The research design is shaped by this. "Brain" is specifically included ("the basic aim of every neuropsychological assessment...is to provide a reliable and valid "picture" of the relationships between the brain and behavior" [Rourke et al., 1983, p. 112]). Neither context nor development are incorporated "up front" in the evaluative-diagnostic component of the assessment. The importance of contextual variables is clearly specified; their methodological role is, however, one of moderator variables influencing the interpretation of the cognitive ability structure. In the treatment-oriented model (Rourke et al., 1986) the role of development, at least in terms of the differential impact of developmentally referenced challenges, is clearly recognized in management planning and the formulation of recommendations. The introduction of the "non-verbal learning disability (NLD)" syndrome and its associated white-matter model (Rourke, 1989) leads Rourke to address the impact of developmental disorder on behavioral outcome within a framework of neural connectivity. Nonetheless, he continues to frame his behavioral analysis in terms of cognitive elements and relationships between them. The model is characterized as "dynamic" (Rourke, 1994), but the "dynamic" (among cognitive variables) is linear (essentially a flow chart), not systemic (with interacting variables), and the linear relationships posited between elements (primary, secondary, tertiary assets/deficits), although Lurian in flavor, are both not specified and can be challenged. (What, for example, is the rationale for positing that perception (primary asset/deficit) is the basis for attention (secondary) is the basis for memory (tertiary)—as suggested by his model? One can equally well argue that "matrix" attentional functions [Mesulam, 1985] are a necessary precursor of perception.) In Rourke's diagnostic methodology, the role of the clinician is deliberately restricted to interpretation and diagnostic decision making; the influence of the behavior of the tester on what the test elicits from the child is neither examined nor incorporated into the model that guides the analysis.

Few have analyzed the brain-behavior relationship problem so elegantly as Fletcher and Taylor (1984) in their four fallacies (differential-sensitivity, similar-skills, special-sign, brain-behavior-isomorphism). Their "function-based" approach to assessment also has a detailed analysis of the CAS of the child at the core of the evaluation of behavior (Fletcher et al., 1995). They situate their analysis, however, in a (horizontal) very rich "biobehavioral context," addressing in detail the limitations of psychometric approaches based on intelligence and achievement measures. They highlight the need to understand the normal devel-

opmental progress and argue strongly for a flexible, and integrative, approach to assessment that makes full use of the assessment tools, techniques, and strategies of relevant psychological, developmental, psychoeducational, and neurological disciplines. They do not specify the theoretical relationship of development to the biobehavioral context. "Brain" is in their assessment approach to the extent that the ability to make inferences about the CNS is a necessary element in the activity. Their inferences about CNS involvement, however, appear to be related to theories of potential disorders (i.e., neuro- or psychopathologies), rather than in terms of a theory of the child. They do insist on the importance of scrutinizing contextual (environmental, psychosocial, learning history) factors prior to making inferences about the CNS. They do not examine the role of the clinician (and/or other individuals) in interactions with the child as a critical part of the context in which the child manifests behavior.

Cohen and his colleagues (Cohen et al., 1992) frame their discussion of child neuropsychological assessment in a somewhat different manner. They set it in a Lurian context, emphasizing the functional-systems model of brain organization (Luria, 1973, 1980). They provide a detailed analysis of this and three additional theoretical issues: information processing, the role of the cerebral hemispheres and lateralization, and plasticity. In their consideration of the clinical implications of these, they outline the many constraints faced by the pediatric examiner in the neuropsychological assessment of the child. They also highlight "Developmental Issues" separately as a titled subsection of their paper. Nonetheless, they do not provide an explanation of the manner by which brain or development is integrated into their diagnostic methodology. Their application of functional-systems principles is not systemic (or vertical) in the sense that Luria himself—working in the Russian systemic psychology tradition of N. Bernstein (1967)—would presumably have understood it; it is horizontal: in spite of their wide-ranging grasp of developmental issues, their explicit goal, when they address the *assessment* of the child, still remains "to accurately describe the child's pattern of neuropsychological strengths and weaknesses and relate them to the specific learning disability or learning disabilities with which the child presents"—(Cohen et al., 1992, p. 71), making theirs, in our characterization, a normative-developmental model.

Contrasts in Clinical Analysis

What difference do these differences in stance make? Do they in fact make a difference? We believe that they can, and do. Examples of differences in clinical diagnostic analysis that follow from different diagnostic strategies may be helpful here.

Example 1: Differences in Hypothesis Generation

In the Rourkean analysis (as we read it in Rourke et al., 1986, p. 27), the interpreter of the test data initially deliberately lacks knowledge of historic and developmental variables and the focus of analysis is the cognitive-ability-structure profile of a given child of a given age, sex, and IQ. To the extent that the pattern of performance on a given variable (for example, a specific linguistic capacity) appears anomalous, the analysis would then (presumably) require scrutiny of historic data to ascertain whether there are moderating influences (such as late onset of language acquisition) that will change the interpretation of the cognitive ability profile. In the systemic developmental analysis, in contrast, the late acquisition of language is an intrinsic element of the developmental course which has yielded the youngster who is the object of the current examination. Here, in a hypothesis-generating/hypothesis-testing assessment model, the language delay is a crucial datum which sets up the hypothesis that left hemisphere mechanisms were not able to function as the primary substrate for language function at the expected time. In this analysis, the identification of deficits that are predicted by neuropsychological theory to be dependent on left hemisphere mechanisms will be necessary to test the hypothesis. The clinical examiner will actively look for disconfirmatory evidence of this hypothesis at both the neuropsychological and psychological levels of analysis.

Example 2: Horizontal versus Vertical Analysis

Difference in theoretical stance can also influence the interpretation of specific behaviors. Take the following scenario. Faced with item #10 of the WISC-III Block Design (Wechsler, 1991), a 10-year-old child begins to mutter to himself as he works. Although the examiner cannot make out the

words, the cadence of the utterances seems to match the movements of the child's hands as he tries one (unsuccessful) approach, regroups, tries another, and so on. Let us assume that each of two examiners appreciates the impact of increasing load on the child's behavior (Block Design #10 is the first of the 3 × 3 designs). The "horizontal analyst," focussing on the cognitive ability structure and taking into account the qualitative aspects of performance associated with this, might write in the report what he or she sees thus: "*In response to the more complex 9 block item, X mobilized a verbal mediation strategy.*" The vertical analyst, making observations from a developmental perspective (in the context of Vygotsky's (1986) characterization of the time course of the capacity for "inner speech"), might offer: "*In the face of the increased demand of the 9 block items X was unable to maintain age-appropriate inhibitory control of verbal output*". Both the inferences drawn, and the types of data sought to test the hypotheses, by the two examiners would be very different. The horizontal analyst, making use of a "complementary contribution" and process-based strategy (Kaplan, 1976, 1988), might test the hypothesis that the left hemisphere is being preferentially mobilized (as evidenced by the use of the verbal mediation strategy) to compensate for less effective right hemisphere-supported inputs, that is, the brain-referenced diagnostic inference would reflect the presumed neural substrate of the child's *response* (brain as independent variable).

In contrast, the vertical analyst would not (initially) assign diagnostic meaning to the verbal nature of the disinhibited behavior. She or he would interpret the loss of inhibitory control as reflecting the child's need to divert so many resources to address the Block Design challenge that brain energy is temporarily unavailable to maintain her or him on the expected rung of the "developmental ladder." To test this hypothesis this examiner would seek other evidence of loss of inhibitory control under specific processing demands. Here, the brain-referenced diagnostic inference would reflect the presumed neural substrate for the type of processing required by the *stimulus* (brain as dependent variable [Bakker, 1984]).

In this example, either examiner could be correct—depending on the other members of the diagnostic behavioral cluster (Bernstein & Waber, 1990) seen in the protocol. The important issue is that a pediatric neuropsychologist needs to have both the horizontal and the vertical perspectives available in the analysis of a given child's performance.

Example 3: Determining Clinical Questions

Differences in the theoretical context within which clinical questions are asked also sets up the questions in different forms. Consider the following example. A low performance on, say, Block Design and Object Assembly is characterized as resulting from "anxiety," that is, appeal is made to a moderator variable to explain the observed behavior. The clinician's question was, presumably, of the form: "*What could have undermined performance on Block Design and Object Assembly*"? (We have additional concerns about this tendency of (inexperienced) clinicians to essentially diagnose—here, anxiety—without realizing they have done so.) We would frame our question very differently—focussing on the behavioral observation and staying as diagnostically neutral as possible—in the form: "*What is it about the processing demands of Block Design and Object Assembly that leads to increased arousal in this child*"?—again seeking to understand the source of the stimulus that causes the brain to respond in this fashion and leaving the valence of the arousal to be judged separately.

Example 4: The Impact of Disorder

This example highlights the challenge for "horizontal" practitioners who focus on the cognitive ability profile which is, at a given point in time, a one-shot view of the child's adaptation. Here, variables associated with the disorder and its subsequent treatment have the potential for changing the child's neurodevelopmental course dramatically. The example is that of the myelodysplasias. (See Figure 11.4.)

Myelodysplasia at whatever level implies abnormal structural development. To the extent that the brain is atypically developed the child's behaviors are likely to be changed, and may be reflected in his or her cognitive ability structure—as indicated by the arrow (1) moving rightwards. But, the abnormal structure frequently results in a mechanical disturbance secondary to obstructed flow of

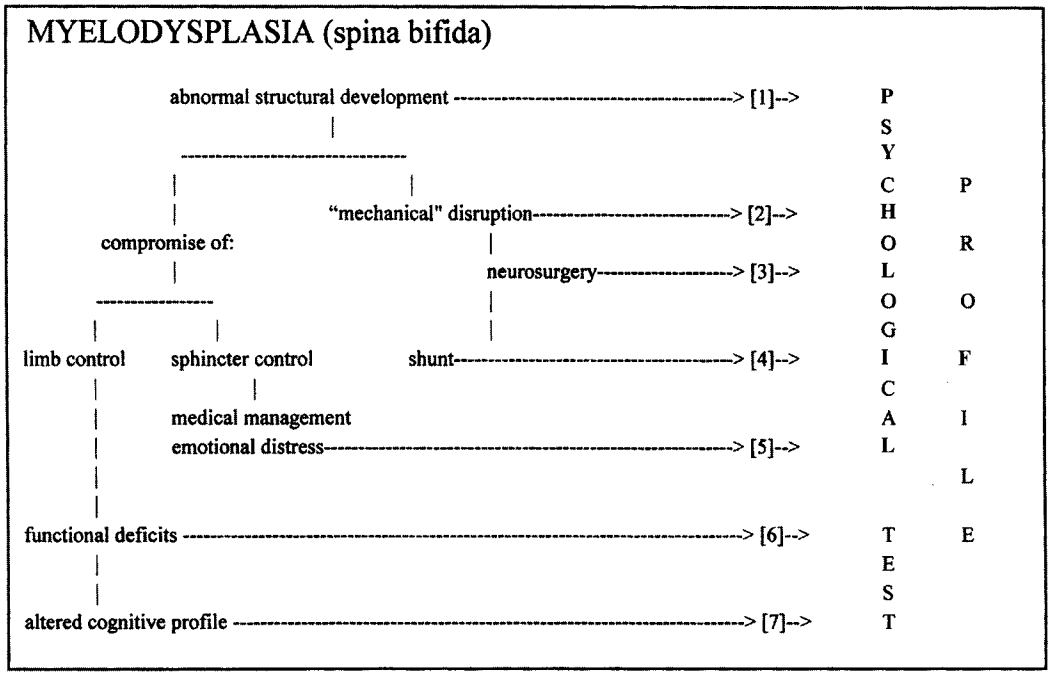


Figure 11.4. Myelodysplasia

cerebrospinal fluid (CSF). This creates conditions that result in disturbance to brain and thus affect behavior (arrow 2). The obstruction needs to be treated, requiring neurosurgical intervention. Surgery is a potentially traumatic event that may have a direct effect on psychological adjustment/development but, in so doing, also has potential for changing the child's interaction with the environment and thus, indirectly, the ongoing development of the brain—with, again, potential impact on the behavioral outcome (arrow 3). The surgical hardware that shunts excess CSF must be placed in the brain, typically going from the right parietal area to the left frontal horn: deficits associated with right parietal damage may be added to the behavioral outcome (arrow 4).

The abnormal development of the brain is likely, in this population, to lead to compromise of either limb or sphincter control, or both. This may necessitate not only additional surgical intervention with its potential for additional traumatic impact—which, again, may indirectly impact the brain and thus behavior—but also multiple medical treatments which can all too easily "medicalize" the child. This can interfere with the normal development of psychosocial competencies (particularly the motivational "value" assigned by the child to active problem-solving behaviors) and, as a result, potentially derail the opportunities for independent problem solving that we believe are critical for the development of normal brain-behavior relationships. Such youngsters are often relatively passive, requiring significant support, in the guise of energy input from clinicians in the assessment setting in the latency period, and being vulnerable to failure in school at or about sixth grade when contextual demands shift and expectations are for increasing independence on the part of students (arrow 5). Treatment for sphincter-control issues carries its own particular risks for breaching the developing sense of autonomy and self-control in a child. Problems with limb-control compromise locomotion—and thus the usual avenue for exploration of the environment. This puts limits on how the world and visual space are experienced—with potential impact on activities that rely on such processing (well represented in psychological tests batteries!) (arrow 6). But these children have not (typically) lost the urge (instinct?) to explore that is

natural to the young animal. If the locomotor domain is constrained, alternative strategies will be mobilized: language and social interaction by face and gesture are available and heavily deployed. Language development in particular is likely to proceed apace—but may be rather "adult" and "scripted" in character secondary to limited flexibility in its use, reduced opportunity for social interaction with peers, and difficulty in integrating language with other developing cognitive functions (the so-called "cocktail party" conversation style (Hadenius, Hagberg, Hyttnes-Bensch, & Sjogren, 1962). Performance on individual language skills will be well tapped by psychological tests (arrow 7). But, as these youngsters mature, later (meta-)language development will be hampered as language skills fail to be integrated with normally developing skills supported by other brain systems. This can be expected to influence the cognitive-ability structure profile derived from psychological test batteries (arrow 7).

This example is provided here to demonstrate, in a very general way, the complexity of the potential impact of the basic medical, functional, and treatment components of one type of disorder. In the case of this particular disorder, the complexity has been shown to be markedly greater once specific disease variables (which differ among individuals) have been taken into account with, additionally, differential impact of psychosocial variables (Holmbeck & Faier-Routman, 1995).

A pure cognitive-ability structure model cannot deal with these developmental medical variables. The normative and systemic developmental models have the problem of determining, given the contributing biological and treatment variables, what on the cognitive profile comes from what. The systemic model has the advantages of being able to make theoretically driven predictions of (many of) the elements of the cognitive profile ahead of time (based on the interplay of theory of child and theory of disorder) and to generate biologically and developmentally referenced hypotheses about the source of deficits on psychological measures in addition to cognitive/psychological process-based hypotheses. The normative developmental model must examine the child's test profile in light of the disorder "after the fact." Neither model can (optimally) relate the cognitive profile to the child's adaptation to the environment without reference to the developmental course of the child and of the disorder interacting with the skill profile. The normative developmental approach is, however, at greater risk of developing remedial plans that are referenced to the cognitive profile without accommodation to the fact that the ongoing "developmental difference" in this child can be expected to modify how she or he takes advantage of pedagogical strategies derived from cognitive ability structures defined in terms of the skills of children developing normally. This is the type of challenge for the practicing clinician that underscores the importance of Rourke's observation that moderating variables (in his model; independent variables in our approach) often (we would say almost always) have greater utility in planning intervention than can ever be provided by analysis of cognitive structures. (We would emphasize, however, that it is not a question of one being better than the other: both are necessary to a comprehensive assessment.)

THE METHODOLOGICAL FRAMEWORK FOR ASSESSMENT

The Theoretical Framework

Cimino's (1994) insistence on the importance of a conceptual framework for clinical assessment and Tramontana & Hooper's (1988) call for a revised conceptual framework for pediatric neuropsychological assessment are appropriate and timely. The conceptual framework must be able to incorporate:

1. gains in knowledge of the organism being studied;
2. revision of the assessment strategy to do justice to this new understanding;
3. review of the adequacy of currently available tests for the neuropsychological endeavor together with development of new instruments to respond to the greater understanding of the organism and its repertoire; and
4. increased knowledge of the pathophysiology and developmental trajectories of the disease processes to which the developing child is subject.

The need to incorporate new knowledge has significant implications for how assessment is done if we are to make real advances in understanding the child and its development from a neurobehavioral perspective. Given the paradigm shift (from phys-

ics to biology) in the way we view science and the science of behavior, it will not suffice to proceed by "accretion," simply adding new tests/measures to our clinical armamentarium and proceeding with the status quo. Our methods must be reexamined to address our emerging appreciation for the organism under study. This methodological review must start from the top, so to speak, with (re-)consideration of the kinds of questions that we wish to ask, followed by careful design of our assessment strategy to answer the questions posed. For us, the primary question is: "How do children 'work'; how does neurobehavioral development proceed"? Others may argue and feel that the primary question for the clinician is: "What is wrong with this child"? We do not believe, however, that this second question can meaningfully be tackled without some (beginning) idea of the answer to the first. Without some sense of what the organism can do—and when, how can we determine what to measure and how to measure it?

The theoretical framework for assessment will thus be tripartite: it will comprise (1) the *theory of the organism* to be assessed (that is, of the developing child in its social context); (2) the *theory of relevant disorders* (the pathophysiology and natural history thereof); and (3) the *theory of the assessment process* (research design and methodology). The latter, in turn, will require consideration of two crucial and complementary subtheories: the *theory of measurement tools* (tests, tasks, techniques) and the *theory of the clinician* (the tool user).

The Theory of the Organism

For a developing organism the theoretical framework must have developmental principles at its core. Models of development are by nature systemic, responding to the inherently dynamic transactions that an individual engages in over time. The systemic tradition is well established in modern developmental psychology, and, indeed, has been applied in neuropsychology through the contributions of Luria (1973, 1980). Though the details of his stage theory have more recently been subjected to intense scrutiny, Piaget (1936/1952) established the structural-systemic foundations of developmental psychology in his accommodation/assimilation principle and his "developmental stage" formulations. Recognition of the (necessarily) complex nature of the acquisition of the behavioral repertoire of the child over development—in critical interaction with experience—has led to increasingly dynamic formulations of the relationship of behavior and experience. Bronfenbrenner and Ceci have argued persuasively for the importance of a bioecological framework in understanding behavioral development (Bronfenbrenner, 1993; Bronfenbrenner & Ceci, 1994; Ceci, 1996). Ford (1987) highlights the transactions between organism and environment that lead to the human "self-constructed living system" over time. Gopnik & Meltzoff (1997), in a Chomskyan tradition, have recently posited the child as an active "theory builder" in these transactions. Models that capture the dynamic and emergent nature of complex behaviors, those involving the motor system in particular, have been explored in depth by Thelen and her colleagues (Smith & Thelen, 1993; Thelen & Smith, 1994). A call for the brain to be explicitly included in models of the developmental dynamic was made by Segalowitz and Rose-Krasnor (1992). Powerful formulations of the brain-behavior interface include the type of coordination-competition interaction framework outlined by Fischer and Rose (1994) using dynamic growth equations to model relationships among both behavioral skills and neural systems, and the central conceptual structure model of Case and his colleagues (Case & Okamoto, 1996), explicitly integrated with the development of neural structures via the work of Thatcher (Case, 1992). Thatcher (1989) has even tackled head-on the challenge inherent in the complexity of the task (of modelling neurobehavioral development) and has made the case for the application of dynamical-systems analyses of the brain-behavior developmental interface—see also Goldfield (1995). To date, such models have been largely the work of developmental scientists. Neuropsychologically-framed models have focussed on the relationship between the cerebral hemispheres in early brain development (Best, 1988; Kolb & Whishaw, 1985; see also Molfese & Segalowitz, 1988). In the assessment context appeal to more or less complex interactive models of development has largely been the province of infant researchers. They have always been, in a sense, "closer to the biology" of their chosen objects of study, both because of the nature of the early repertoire that is being

acquired and because their patients cannot yet do the sorts of psychological tests that dominate the educational lives of school-age children—and have thus shaped the form of assessment strategies for this age group!). Their goal has been to understand the course of development of children with neurodevelopmental disorders with the express intent of making developmentally sensitive interventions (see, for example, Als' (1982) "synactive" model for premature infant development). To do this, they have recognized the importance of characterizing the behavioral repertoire of the early development of the normally developing child in terms of its own unique ecology (Wolff, 1987) and maturational and experiential dynamics, and not through the lens of adult behavioral models. In an effort to extend these principles of dynamic interaction across development to the school-age population, Bernstein and Waber (1990) formulated a systemic model for the neuropsychological assessment of the child which has development at its core and with behavioral transactions framed—and analyzed—within a *brain-context-development* matrix. The development, analysis, and application of such models in the assessment setting will be increasingly important—to do justice to the expanding knowledge base of neurobehavioral development and to use it, via systematic hypothesis-generation and testing, in the service of the individual child, not just in infancy and the preschool period, but throughout childhood, adolescence, and young adulthood.

The Theory of the Disorders

The assessment process is guided not only by knowledge of the child, but also by knowledge of the potential pathological conditions (neurological and/or psychiatric) in which aberrant behavior may be situated. In considering the role/impact of a disorder, three sets of principles are critical: these derive from (1) the nature of the disorder (e.g., structural anomaly, focal brain insult, genetic/chromosomal abnormality, etc.)—this affects the manner in which the symptoms will be expressed; (2) the timing, course and duration of the insult—these will interact with the maturational schedule of the child to change the course of future development (Dennis, 1983, 1988; Fennell, 1994); and (3) development subsequent to both condition and treatment—both the condition and its necessary treatment (e.g., radiation/chemotherapy, surgical intervention, medication, pedagogical interventions) have the potential for affecting future development beyond the effects of the condition itself.

The clinician's understanding of the disorder(s) being assessed (i.e., the hypotheses being tested) *guides* the assessment design by ensuring that the critical pieces of disconfirming evidence are collected. As we have argued for the assessment process itself, gains in the knowledge base pertaining to a given disorder (e.g., attention deficit disorder [ADD], attention deficit hyperactivity disorder [ADHD]) will also require updating of the type of data necessary to test any given hypothesis (i.e., disconfirm the presence of the disorder).

Take the example of a child who presents with reading difficulty as the index symptom. If one's theoretical perspective holds that developmental dyslexia is the result of problems with phonological processing (Shankweiler, Liberman, Mark, Fowler, & Fischer, 1979), one would expect that dyslexia will emerge early in the child's reading education, will result in difficulties learning symbol/sound relationships, will persist over time, and (even in "compensated" dyslexics) continue to be present as difficulty with nonword reading and/or slowed reading rate. If, however, the reading problem was first noticed in third or fourth grade and manifested as reading comprehension difficulties in the context of secure sound/symbol associations and age-appropriate reading speed, the hypothesis that the child's reading problems were the result of dyslexia (as so defined) would be unproven.

However, a plausible rival hypothesis (Campbell & Stanley, 1966) might be that the child's problems in the reading domain are the result of difficulties with executive control processes. In this case, one might anticipate that the initial signs of the disorder would be impulsivity and overactivity during kindergarten and first grade. However, it might also first present in fourth grade as the average child faces increasing expectations to mobilize executive skills on an independent basis and thus to organize his or her work with less direct guidance. In this case, disconfirming evidence may take the form of adequate organizational, self regulatory, and reasoning skills during formal testing.

As an alternative example, consider the case of a youngster with normal language and motor development, who begins exhibiting at age four years

seizure activity emanating from left cortex; structural abnormalities are not observed on anatomical scanning. This youngster may very well demonstrate the same type of reading delays and retrieval difficulties as the child with dyslexia. In addition, attentional inefficiencies (relatively common among children with seizure disorders) may be observed across multiple settings/domains. In this case, appropriate diagnostic labels for the child's difficulties may very well be dyslexia and ADHD. However, from a child-centered, biologically-referenced perspective, the primary diagnostic formulation would be framed in terms of the child who in this case has a neurological condition (neurobehavioral disorder in the context of documented seizure disorder). Here, the seizure disorder has the potential for derailing the child's emergent development in general. All other things being equal, the "manifest disabilities" (Fletcher et al., 1995) of reading and attention would be considered secondary to the neurological condition—as a function of the presumed left hemisphere dysfunction in the one case and the more general, seizure-related undermining of attentional networks in the other. Strategies for subsequent intervention/management are likely to follow from this difference in framing the presenting problem. Note that other data, say, a family history of reading difficulty, would mandate close scrutiny of the left hemisphere disruption/reading deficit interpretation—although, even in such a case, the covariation of familial dyslexia (Pennington, 1991) is such that a family history of reading difficulty would not necessarily change the likelihood that the reading deficit results from disrupted left hemisphere function in the context of focal seizure disorder.

The Theory of the Assessment Process

The clinical assessment is a procedure for answering specific questions about behavior, in this case, that of a given individual. It is formally equivalent to the traditional research study (see also Fennell & Bauer, 1989, 1997; Pennington, 1993; Rabinowitz, 1994). Thus, it is our position, and the theoretical framework within which we train our students,[7] that the clinical assessment must be treated as an experiment with an *n* of 1. Both the experiment with the *n* of *N* (the *research experiment*) and the experiment with the *n* of 1 (the *clinical experiment*) seek to answer research (diagnostic) questions by confirming or disconfirming hypotheses involving observed phenomena. The answers to research questions are not derived directly from the statistical analysis of the data; interpretation is necessary to understand the meaning of the data (Kerlinger, 1986). Similarly, test scores alone are not sufficient for diagnosis (let alone intervention). The research or diagnostic question is answered by *the design of the research study (assessment)* (Table 11.1). A carefully designed assessment enables conclusions to be inferred; without such a design, alternative hypotheses cannot be excluded. We believe that the extensive knowledge base available in research methodology and design can (and should) be applied explicitly to the challenge of assessment. It is through the application of *n* of *N* procedures to the *n* of 1 environment that best practices for pediatric neuropsychological assessment (and assessment in general) can be derived.

As Kerlinger (1986) states, the three criteria for judging the effectiveness of research designs are: adequate testing of hypotheses, control of variance, and generalizability. For the neuropsychological assessment, adequate testing of hypotheses refers to the accumulation of information that will allow relevant hypotheses to be either confirmed or disconfirmed. Threats to construct validity which operate in the research study are equally of concern in the *n* of 1 clinical experiment (see Table 11.2). In order for a design to be able to truly test a hypothesis, data must be collected in such a way that it can be falsified if it is in fact incorrect (Rosenthal & Rosnow, 1984). In the experimental design, variance control pertains to three separate types of variance. Systemic variance related to the experimental manipulation should be maximized. The effects of variance related to factors extraneous to the experimental manipulation should be controlled through the use of random assignment or incorporation of the factors in the research analysis. Error variance is minimized through standardized procedures and selection of reliable measurement tools.

In the neuropsychological assessment, maximization of systemic variance can be achieved by utilizing tools and techniques that are sufficiently sensitive to unambiguously identify individuals with the diagnoses of concern (see discussion of positive predictive power below). Control of variance related to factors extraneous to the diagnosis can be achieved through inclusion of factors (e.g., developmental history, academic exposure, emotional trauma) in the overall understanding of the

Table 11.1. Scientific Design

	EXPERIMENT WITH *N* OF N	EXPERIMENT WITH *N* OF 1
Purpose	Answer research question.	Answer intervention question.
Process	Observation→question→ hypothesis→prediction. Design experiment to test hypotheses.	Presenting problem→question→ hypothesis→prediction. Design assessment to test hypotheses.
Hypotheses	Single or multiple.	Multiple.
Design	Experimental/quasi-experimental (nonrandom assignment).	Quasi-experimental.
Control group	Matched control group.	Developmentally-referenced data sets. Age-referenced data sets. Standardization sample.
Variable definition	Determined by design/purpose.	Determined by design/purpose. Defined by theoretical orientation.

Table 11.2. Threats to Construct Validity

	RESEARCH PROBLEM	ASSESSMENT SOLUTION
Inadequate preoperative explication of constructs	Research measures do not adequately capture the construct.	Systems and disorders of interest should be clearly understood.
Mono-operation bias	Single measures under-represent constructs and contain irrelevant variance.	Use multiple measures to support/ disconfirm diagnoses.
Mono-method bias	If data is collected using only one method, factors unrelated to the construct cannot be controlled.	Use multiple assessment methods (questionnaires, analytic interviewing, observation, standardized test results, etc.).
Hypothesis guessing by participants	Participants guess hypothesis and provide responses to fit.	Approach assessment with multiple hypotheses. Use multidimensional questionnaires.
Evaluation apprehension	Participants may be unwilling to respond in ways that put them in a less-than-favorable light.	Cross validate interview information, use scales with validity indices, establish rapport.
Experimenter expectancies	Knowledge of the experiment may lead to treatments being given in such a way as to confirm expectations.	Use structured assessment designs and decision-making de-biasing techniques.

Note: After Cook & Campbell, 1979.

child (i.e., the whole child in his or her system) and reduction of other variables (e.g., hunger, fatigue, boredom, fear, etc.) that can invalidate the testing results (See Table 11.3). As in research designs, error variance is minimized through the use and standardized application of reliable measurement tools and consistent (replicable) data integration procedures.

In research designs, the generalizability standard (one of the various forms of validity) requires that the results of the study be applicable to other groups in other situations. In the neuropsychological assessment, generalizability refers to the likelihood of the child's performance during the evaluation being relevant to the child's behavior in his or her real-life environment (ecological validity). See Tables 11.4 and 11.5.

Research Design and Methodology

Research design constitutes a multiply redundant system in which formal procedures are applied at different levels. These include: the original theory (theories); the generation of hypotheses; the experimental design; group

Table 11.3. Variance Control

	EXPERIMENT WITH *N* OF *N*	EXPERIMENT WITH *N* OF 1
Maximization of relevant systemic variance	E: Design tasks to maximally separate groups QE: Select IVs to separate groups	Use tools that are maximally sensitive to the hypotheses to be tested.
Control for extraneous variance	Employ random assignment matching. Include as IV.	Integrate "extraneous" systemic variance as IV (age, history of language delay, etc.) Modify environment/assessment process to minimize fatigue, boredom, etc.
Minimization of error variance	Use reliable tools. Use standardized measurement practices.	Use reliable tools. Use standardized procedures. Control the environment. Replicate observation sets. Reduce reliance on memory. Use stop watch. Follow test instructions. Systematically review administration procedures on regular basis.

Note: E = experimental; QE = quasi-experimental; IV = independent variable

selection; task design; task administration (formal procedures); and reliability in administration, in scoring, and in the collection of normative data.

The clinical "experiment" entails different responses to the requirements of research design and method than that of the research investigation (see Table 11.6). For example, the clinical experiment cannot conform to the assumptions of parametric statistics, nor can it eliminate the preconceptions and already acquired knowledge of the clinician—making a Bayesian analysis necessary (Murphy, 1979). Where formal strategies are not available, as in the case of Bayesian methodology, however, the same goal of experimental control may require alternative applications of methodological tools/techniques. For example, in the *n*-experimental setting the influence of the observer can be subjected to the same control as any other variable, i.e., it is "averaged" across multiple observations—of the same behavior by multiple persons. The impact of more than one observer can, if needed, be formally tested post-hoc by directly comparing the distributions of observations made by one observer with those of another. In the clinical-experimental setting, this is not possible: clinical assessment is done with one person at a time. Alternative methodological strategies must thus be employed to obtain the same goal of experimental control. Multiple observations from multiple informants (i.e., a multi-method design: Campbell & Fiske, 1959) are still necessary. These are, however, necessarily different. To be of value in diagnosis, they must be congruent in their implications: they must both converge on, and discriminate between, coherent diagnostic entities. It is the theory of the organism, not the theory of assessment or of the tools, that determines the congruence; the critical validities for diagnosis are convergent and discriminant (see also Pennington, 1991). Methodologically, two conditions must be met: (1) converging data must be derived from multiple domains (a cross-domain analysis); and (2) behaviors that are not predicted by the theory for the neural substrate in question should not be present. Note that the latter must be actively and systematically sought to counter the risk of confirmatory bias.

The Theory of the Tools

Psychology has long been the leading discipline in the measurement of behavior. The theoretical principles underlying the design, construction, and use of relevant measurement tools has been extensively addressed in a myriad of books, journals, and test manuals. Neuropsychologists have contributed to this process by submitting, at least some, "neuropsychological" tests to the same standardization procedures as are used in the larger field of psychology (Heaton, Grant, & Matthews, 1991; Korkman, Kirk, & Kemp, 1997) and have

Table 11.4. Validity

	EXPERIMENT WITH *N* OF *N*	EXPERIMENT WITH *N* OF 1
Reliability	Experimenter training. Tools. Standardized administration.	Clinical training of examiner. Tools. Standardized administration.
Internal	Control of variance. Blind testing. Research design. Double dissociation.	Use of tools with construct validity. (Use of psychometrist.) Tool selection to provide convergent/divergent validity.
External	Generalization.	Ecological.
Interpretation of results	Guided by logic of design and statistics employed.	Guided by design logic, diagnostic method, and test findings.

Table 11.5. Threats to Internal Validity

	EXPERIMENT WITH *n* OF *N*	EXPERIMENT WITH *n* OF 1
Maturation	Biological or physiological change not relevant to hypothesis that affects status of Ss on DV.	Consider maturational variables. Need to employ measures sensitive to maturational change with developmental norms.
History	An extraneous event outside or within the experimental situation can affect the status of Ss on the DVs.	Review historical factors: psychosocial experience, insult/trauma, sensory disruption.
Testing	Taking a test can alter performance at a second testing.	Has child been tested before—with current or different measures?
Instrumentation	Changes in measuring devices/procedures during the course of the study.	Comparing across different versions of tests (e.g., WISC-R vs. WISC-III) across tests within same battery with different normative groups.
Statistical regression	Selection of children with the poorest scores for the study group. Tendency for extreme scores to regress towards the mean on re-administration.	Referral itself increases the likelihood of receiving a diagnosis. Extreme scores are likely to regress towards the mean on retesting.
Mortality	Ss who drop out of one group may differ in some important way from Ss who drop out of other groups.	Unable to complete parts of test battery. Loss of data due to mistakes. Failure to receive all relevant materials (e.g., behavioral checklists).

Note: DV = dependent variable.

begun examining principles of validity (Franzen, 1989) and reliability (Goldstein & Shelly, 1984). To the extent that this is already a very well-examined area of psychology, we will not address it further. Suffice it to say that, for neuropsychological assessment as for behavioral measurement for other purposes, instruments that are well-designed, well-normed (ideally on the same populations) and appropriate for the use to which they are put are critical to the endeavor. From our diagnostic perspective, not all of these need be built on population-based (either general or local) normative data but certainly the core of any assessment protocol should include such instruments. Supplementary research-based instruments should be used with due care.

The Theory of the Clinician

It is our position that a comprehensive description of human neurobehavioral development

Table 11.6. Statistics

	EXPERIMENT WITH *N* OF *N*	EXPERIMENT WITH *N* OF 1
Type	Parametric/non-parametric. Descriptive. Inferential.	Bayesian. Inferential.
Significance levels	Alpha set at .05.	Clinical significance set at 2 SD~.02 (one tailed; 1.65~.05 (two tailed).
Power	Choose design, manipulation, and groups (size or types) to maximize chance of finding difference.	To test hypotheses, select tools sufficiently powerful to show weakness when it occurs.

(assuming such is possible) cannot be achieved via the study of groups alone, but must incorporate a detailed understanding of the behavioral repertoire of *individuals* as complex, integrated, organisms intrinsically linked to their environment/ecology with both a history and a future. This stance is, of course, precisely that of the clinician of behavior whose mandate is to maximize the experience/adjustment of the *individual* not only by utilizing past experience and current functioning to generate a diagnostic description now, but also by providing some prediction for the future. In this regard, clinicians—and the single case analysis—are an integral part of the larger neurobehavioral research endeavor (a fact recognized by the recent establishment of a new journal *Neurocase).*

We are equally convinced, however, that neurobehavioral clinicians are not yet altogether equal to the role we have assigned them. This is because the clinician's behavior in the process of assessing an individual—in the neurobehavioral context—has not yet been fully analyzed as both an indispensable source of relevant data and an intrinsic element in to the formulation of the diagnosis. Nor has it been subject to the degree of examination and operationalization as has been afforded the measurement instruments. This situation must be redressed if we are not to overlook diagnostic information crucial to the understanding of neurobehavioral development in children.

As clinicians ourselves and as trainers of the next generation of pediatric clinical neuropsychologists, we have two primary concerns. First, there are many behaviors that are invisible to psychological tests that are *both* critical to a complete description of a child's brain function *and* observable by clinicians. These observations are not necessarily "subjective;" they can be operationalized and subjected to formal tests of reliability (we are currently addressing this process in our own laboratory: Bernstein, 1996b). (The need to do so is long overdue, given the centrality of making and utilizing behavioral observations in the supervisory process.)

Second, clinicians are at serious risk of—unknowingly—generating diagnostic hypotheses based on non-psychometric data and then seeking to validate these—unexamined—hypotheses by selecting psychometric data to fit. The clinician, in any given instance, may be correct in his or her diagnosis. The problem is not only that she or he will not be correct in all of them, but also that she or he will not know in which instances she or he is correct and in which she or he is not. An additional concern is the fact that the clinician relies on the "objectivity" of the psychological tests, believing them to be a rigorous standard against which to establish diagnostic validity. He or she, however, may fail to appreciate that his or her *use of the tests* is by no means objective. This is no less of a potential and very worrisome problem in actuarially based assessment strategies as in more flexible and/or qualitative approaches (Willis, 1986).

The clinician's contribution to the diagnostic process must first be recognized in its many facets. It must then be subject to increased experimental structure. However, it cannot be so structured independently of the larger research design. The research design and the clinical decision-making function are separable, but intrinsically interrelated, elements of the assessment. The ethical clinician must thus function within a properly formulated—and explicit—investigative design with appropriate methodological controls for bias and error.[8] The investigative design and the controls employed *must be formulated expressly to address the particular assessment challenges of the developing child within the neurobehavioral context.* Within this larger research design the increased structuring of the clinician addresses well-known, but often imperfectly appreciated,

sources of bias and error that interfere with accurate diagnostic decision making. This structuring will include controls that are specific to the clinician and his or her behavior—and to the clinical goals of different aspects of the behavior. Strategies that serve to control variance in tests are not sufficient to "control" the examiner.

The clinician both uses tools and is a tool. She or he has four important roles in assessment: (1) administrator of tests; (2) data-collection technique; (3) analyzer of behavioral information; and (4) diagnostic decision maker. Each particular role must be scrutinized for the type of control that is most appropriate.

1. Test Administrator

The traditional role of the clinician is that of administrator, scorer, and interpreter of tests. This requires the clinician to learn to administer tests accurately and reliably, to respond appropriately in the interpersonal interaction with the child, to employ systematic limit-testing techniques in relevant instances and at appropriate times, and to maintain accuracy in scoring according to standard guidelines.

2. Data Collection Technique

The clinician him- or herself is, however, also a data collection "technique". She or he is a critical element in the clinician-patient system (Henderson, 1935) and thus is an integral part of the data to be derived from the transaction between adult and child in the assessment setting. She or he is also critical to the collection of ecologically-important data from the nonclinical environment via the clinical interview.

Adult-Child System/Transaction. The behaviors of adult and child in the clinical setting are reciprocal. The adult naturally supports the transaction by supplying what is needed to facilitate optimal communication in the dyad. This requires the clinician to be aware of his or her own behavioral baseline, to monitor any change from baseline that this particular child under this particular demand elicits, and to actively test the hypotheses that such behavioral change sets up. Thus, observing that one is slowing, simplifying, repeating, and/or rephrasing one's utterances in the course of ongoing conversation sets up a hypothesis of potential language impairment and requires that the examiner examine in detail the child's language processing skills, both in linguistic interactions and on specific tests of language capacities—as well as other, not overtly related, skills that may also depend on the integrity of left hemisphere brain mechanisms. (These must be derived from both language and nonlanguage behavioral domains. Deficits in language alone would not be a sufficient test of the *neuro*psychological hypothesis, that is, one specified in terms of a neural substrate: such would only provide information at the psychological level of analysis.) Such a hypothesis also, however, requires that the examiner actively look for, and evaluate the impact of, other reasons for slowed output or need for repetition, such as a general rate of processing deficit, attentional instability, or hearing impairment. These would then be seen in the context of a different diagnostic behavioral cluster. Note that the *change* in the examiner's behavior elicited during the interaction with the child will be a member of the diagnostic behavioral cluster, equivalent in this respect to test scores, quality of performance, historic variables, and so on.

The analytic interview. Interviewing technique, the ability to elicit information from caretakers, teachers, and so on that is as free from bias as possible, is crucial to any psychological assessment approach. Good interviewing technique is thus a *sine qua non* of the clinician's armamentarium and should be undertaken in systematic fashion (Maloney & Ward, 1976). The interview is an intrinsic part of the neuropsychological *assessment* (as opposed to testing), and not separate from it. It is thus governed by the research design and theoretical principles of the assessment. Given this, interviewing strategies need to be extended and tailored to the neuropsychological context specifically. Interviewing is an active process in which no observation is taken "cold," all observations are analyzed in light of their potential neuropsychological source or implications. Interviewees are thus queried to elucidate the actual behavior (rather than an interpreted version thereof) that they are describing. Strategies include: query providing a *targeted contrast* of a descriptive label (e.g., a child's response of "This is boring" elicits "Is it boring-easy or boring-hard?"); *clinical analysis* of a descriptive label (a parent or teacher description of *anxiety* cues the skilled examiner to consider the actual behaviors that would lead the layperson to use the label "anxiety"—such as press of speech or motor

activity—and to actively query the quality of speech and/or motor patterns with a view to evaluating the possibility of neuropsychological, rather than emotional, factors contributing to the observed behavior); and elicitation of *relevant anecdotes* (a complaint of *memory problems* in a child leads the clinician to ask for a specific example of the kind of situation in which the problem occurs—so that he or she can consider it from a broader neuropsychological perspective that may well include language processing or attentional issues, for example). The data from this analytic interview technique is cross-checked (where possible) against reports from other individuals/sources, and/or the neuropsychological hypotheses to which they give rise are tested against other types of assessment information (i.e., multi-method, multi-trait analysis).

3. Analyzer of Behavioral Information

A clinician, especially when working with children, is always a participatory observer who acts or does not act within a transaction with the express intent of eliciting/modifying behavior in the child. Awareness of this participatory role, as well as of the theoretical framework within which the assessment is undertaken, must be made explicit: the framework in which a clinician practices influences the way in which she or he interprets what she or he observes as she or he is observing it, and thus shapes the on-line formulation and testing of hypotheses. As noted above in the discussion of contrasting clinical-analysis strategies, different theoretical stances can lead to diametrically-opposed brain-referenced diagnostic formulations—or, at least, hypotheses that must be examined. Thus, an analysis of behavior in terms of the lateral neural axis (is it left or right hemisphere implicating?) that does not recognize the child's developmental status (can the child inhibit, on a developmental basis, the behavior in question under stress?) may well misrepresent the source of the behavior in question. Similarly, framing one's observations of behavior in their context (what is it about this situation that is eliciting this behavior from the brain?), as opposed to relying on a presumed brain-behavior relationship (this part of the brain "does" this), leads to very different hypotheses, to a search for very different supportive and disconfirmatory observations—and to (potentially) quite different intervention strategies.

4. Diagnostic Decision Making

The clinician is the primary analyzer of the behavioral information collected during the evaluation. In the clinical setting it is she or he who brings an appreciation of the human condition and its vagaries to this encounter with the patient, thus enriching in a uniquely human fashion the description of behavioral function provided by various measurement techniques. It is, however, the clinician's very humanness that makes him or her prone to error. The human mind is limited in its capacity to analyze information. Our attempts to circumvent this when confronted with complex cognitive tasks (as in the development of a diagnosis) lead to predictable types of decision-making biases. These are reviewed in Table 7.

To take just two of these sources of potential error: the *anchoring-and-adjustment* bias can have significant impact on the nature of the information that is collected and/or considered important in different assessment strategies. A neurodevelopmental assessment model "anchors" on a review of systems and frames the clinical analysis in "brain" terms. A flexible battery approach is likely to anchor on the interview and history as the basis for selecting measures. A fixed battery approach anchors on the tests that constitute the battery. None of these strategies are free of the potential for "adjusting" subsequent data to match the framework provided by the initial data set; the way in which they adjust—and the controls necessary to accommodate potential for error—are likely to be different.

Under-utilization of *base rate* information is a common source of clinician error. Consider a test with 90 percent sensitivity (i.e., the proportion of individuals with a disorder who exhibit a sign) and 90 percent specificity (i.e., the proportion of individuals without a disorder who do not exhibit the sign). The parameter that the practicing clinician is most interested in is the test's positive predictive power (PPP), or the probability that an individual who receives an abnormal test score actually has the disorder of interest (Ellwood, 1993). PPP is determined by the test's sensitivity and specificity in the context of the base rate of the condition. Even with 90 percent sensitivity and specificity, if the base rate of the condition is relatively rare, the majority of individuals who exhibit that sign will not have the condition.

The following is an illustration of this dilemma. In an unreferred population of 1,000 children, and

Table 11.7. Expectable Biases in Clinical Decision Making

DECISION-MAKING BIASES	NATURE OF ERROR
Limited capacity	Miller (1956) argued that incremental improvements in decision-making accuracy occur until approximately 7 (+/–2) pieces of information have been collected. Beyond this, the capacity of the system is overloaded. Provision of additional information (beyond a few of the most valid indicators) is unlikely to increase predictive accuracy (Oskamp, 1965) and may actually lead to a decrement in decision-making accuracy (Golden, 1964; Wedding, 1983a; Wedding, 1983b).
Simplification	When confronted with complex cognitive tasks, we typically resort to simplification of the information involved and are usually unaware of the manner in which these changes operate. Although we may believe that the exercise of complex pattern integration is what guides decision making, it is the linear combination of data that has been shown to account for a range of diagnostic decisions made by, for example, radiologists (Hoffman, Slovic, & Rorer, 1968), psychiatrists (Rorer, Hoffman, Dickman, & Slovic, 1967), and psychologists (Goldberg, 1968; Wiggins & Hoffman, 1968). Indeed, Fisch, Hammond, & Joyce (1982) demonstrated that psychiatrists' diagnoses of depression were fully accounted for by only one or two pieces of information despite their conviction that they were integrating many more data points into their decisions. Moreover, clinicians may believe that they rely upon a certain piece of information in making their decision, when analysis of their performance reveals that other information actually swayed their opinions (Gauron & Dickinson, 1966, 1969; Nisbett & Wilson, 1977).
Use of Heuristics:	These simplifications can take the form of "rules of thumb" or intuitive heuristics that may be useful in certain applications, but lead to predictable types of decision-making biases.
Representative heuristic	When using this, people assess the probability of occurrence (i.e., graduate school major) by the degree to which the information is consistent with their pre-conceptions of the category (Kahneman & Tversky, 1972).
Availability heuristic	This is applied when people assess the probability of occurrence by the ease with which that occurrence can be remembered (Tversky & Kahneman, 1974). For example, they are influenced by factors such as recency (recent observance of a similar case) or the degree to which an outcome is memorable (unusual outcomes are salient simply because they are unexpected).
Anchoring and adjustment heuristic	These may shape probability judgments: the person starts from an initial value (anchor) which is usually the first piece of information examined and adjust their interpretation of additional data to conform.
Confirmatory bias	This leads people to emphasize information that is consistent with their hypotheses and to ignore information that would be contradictory to the hypothesis (Nisbett & Ross, 1980; Ross, Lepper, Strack, & Steinmetz, 1977).
Covariation estimation:	As a result of these limitations on decision making, certain types of cognitive operations are difficult, if not impossible, to accomplish. For example, in order to determine whether the presence of a sign has a valid predictive relationship with an outcome (i.e., covariation), one must know: (a) the proportion of time that the sign is present when the condition is present and (b) the proportion of time that the sign is present when the condition is absent. Additionally, in order to know whether a sign has a *useful diagnostic* relationship to a given condition, one must know the base rate of the condition in the population of interest. This relationship is, however, extremely difficult to determine unless data are formally collected (Arkes, 1981). Informal estimations of sign-outcome relationships are usually unreliable because preconceived notions can bias judgments of how variables covary (e.g., large eyes on human-figure drawings and suspicious personality types; Chapman & Chapman, 1967).
Base rates:	A relationship between a sign and a condition may be valid and is certainly necessary to demonstrate the clinical utility of the sign; it is not, however, sufficient (Faust & Nurcombe, 1989). Information regarding the condition's base rate, that is, the prevalence of the condition within the population being examined (Meehl & Rosen, 1955), must also be considered. This is, however, not typically the case in clinical practice (Kennedy, Willis, & Faust, 1997).

a 4 percent base rate for ADHD, 40 children are expected to have ADHD. However, using a test with 90 percent sensitivity and 90 percent specificity, only 27 percent of the children who receive an abnormal score on the test can be expected to actually have ADHD (see Figure 11.5.).

It is, however, important to note that these factors must be considered in the context of the implications of making a false diagnosis. When risks to the individual of overlooking the diagnosis are severe (e.g., not diagnosing a brain tumor), the inclusion in the assessment of a "screening" sign with high sensitivity and moderate specificity may be indicated because of the test's capacity to reduce false negatives.

To address the above, the clinician must acquire a variety of analytic "thinking tools," the foremost of which is the appreciation of the clinical assessment as a formal investigative procedure. The role of the clinician in that formal procedure should be fully understood, both for its value in assisting in diagnosis of a child's neuropsychological competencies and its limitations in terms of its own neuropsychologically mediated biases. Such thinking tools should include the application of "corrective procedures" (Wedding & Faust, 1989): the clinician should not only know the literature on neuropsychology, but also be well versed in that on human judgment; should not depend on insight alone; should start with the most valid information and "think Bayesian" (in terms of base rates); should collect appropriate age-, sex-, and education-adjusted norms; should avoid over-reliance on highly intercorrelated measures; should avoid premature abandonment of useful decision rules, regress extreme estimates and confidence in relation to level of uncertainty, and not become overly focused on the esoteric; should list alternative diagnoses/options and seek evidence for each, systematically list disconfirmatory information, and make a deliberate effort to obtain feedback. (The clinician would also do well to heed the potential abuses outlined by Prigatano and Redner, 1993).

PRINCIPLES OF THE SYSTEMIC ASSESSMENT APPROACH

Our approach to the neuropsychological assessment of the developing child has not only been shaped by our own clinical practice and experience and what we have learned from our colleagues and the literature, but has also been honed by teaching several "generations" of pre- and postdoctoral fellows. We have learned both what students do not know and how difficult it is for them to integrate what they do know. We have learned that we cannot assume that students have absorbed the principles underlying clinical assessment from their reading of the literature, but need direct instruction and ongoing discussion in the supervisory interaction to identify and integrate them into their clinical behavior. The general principles governing our approach are as follows. Students are taught to:

1. understand that the clinical goal is a portrait of the "whole child" (Bernstein & Waber, 1990; Matarazzo, 1990).
2. view the assessment as an experiment with an *n* of 1. This experimental stance requires a theory to guide hypotheses and explicit attention to the formal requirements of investigative design and methodology.
3. frame the theoretical context as a tripartite structure that has, at its core, the theory of the organism (in this case, the child to be assessed), complemented by a theory of the possible disorders, and a theory of the assessment process. This entails recognition that the theoretical context must be the same for the evaluative phase of assessment as for management.
4. view themselves as an integral part of the research design and method. This necessitates learning to increase the structure of their contribution as a complement to the contribution of the tests, by identifying and using relevant "thinking tools" and by examining their own behavior, responses and reactions, neuropsychological and emotional, in the course of the interaction with both the child and parents as the assessment proceeds.
5. scrutinize all observations as a function of the *brain-context-development* matrix critical to a developmental neuropsychology.
6. approach the child's behavioral presentation as an outcome of his or her experience to date; consider both horizontal and vertical perspectives when analyzing observations.
7. actively mobilize the entirety of their psychological knowledge base to the service of the assessment. Thus, the principles of developmental, cognitive, and perceptual psychologies, of information-processing theory, of

	ADHD	NO ADHD	
Abnormal Score	36	96	**PPP = 36/(36+96) = 27%**
Normal Score	4	864	**NPP = 864/(864+4) = 99%**
	Sensitivity = 36/(36+4) = 90%	**Specificity = 864/(864+96) = 90%**	

Figure 11.5. Effect of Low Base Rates on Positive Predictive Power

behavioral medicine, of child psychopathology are just as crucial to the clinical examination and management of the individual as are neuropsychological principles—and those involving assessment techniques, psychometric theory, and test construction.

8. utilize the three major neuroanatomic axes as a primary organizing heuristic; understand the value and limitations of heuristics.
9. approach the diagnostic activity not as one of "ruling in" a list of signs and symptoms, but as one of *ruling out* all (possible) non-brain (contextual, psychosocial, environmental) variables that could explain the observed behavior.
10. explain both the "ups" and "downs" of performance. The child, however impaired his or her functioning, still functions as an integrated person. The description must characterize the "whole child." Deficits may be crucial for diagnosis, but competencies drive intervention.
11. apply a diagnostic method systematically:

a. use a review-of-(neurobehavioral)-systems strategy as an organizing structure to review important variables that are potentially invisible to tests;

b. evaluate the impact of systemic influences (from brain to society) on behaviors observed; do not assume a "brain" effect/influence when context can account for the observations;

c. seek to identify theoretically-coherent convergent and discriminant observations from multiple domains based on multiple assessment methods (the diagnostic behavioral cluster). The clinician must be able to account, in neuropsychological and/or psychological terms, for the child's ability to do what she or he *can* do, as well as what she or he cannot do. (Given a problem sufficient to warrant a neuropsychological assessment/diagnosis, it is even more important to understand how successes are achieved and/or strengths are supported; without such understanding the neuropsychologist cannot claim to have adequately characterized the child's brain function or neurobehavioral repertoire.)

12. subject history (developmental) and observational (contextual) variables to equal scrutiny as given to test performance—with the understanding of the limitations of retrospective reports and the influence of assessment ecologies (contexts);
13. administer tests reliably;
14. develop clinical limit testing skills; know how, and when, to use them; use them rigorously;
15. examine all behaviors from the perspective of both brain competencies and task demands;
16. formulate the diagnosis in neuropsychological terms; relate this as specifically as possible to the presenting complaint; determine the appropriate level of diagnostic category for the report and feedback;
17. determine the particular goals of the feedback for each family/parent (the overall goal of the assessment is to communicate the findings, but the individual clinical goals may also need to address psychological factors [such as denial, narcissistic injury, unrealistic expectations, etc.] in parents or other adults to maximize the needs/adaptation of the child);
18. provide a properly orchestrated feedback session that (a) empowers parents by educating them about neurobehavioral development in general (as framed in the *brain-context-devel-*

opment matrix), (b) situates their child's profile of skills in the larger context of neurobehavioral development, thus "normalizing" it to the extent possible, (c) addresses specific complaints, and (d) reframes the understanding of the child (where indicated);

19. provide a systematic analysis of short-term and longer-term risks that follow from the brain-context-development analysis for the given child;
20. relate the recommendations systematically to the risks identified;
21. prepare a detailed report of the assessment that (a) documents what was done, (b) presents sufficient data (both "negative," implying deficit, and "positive," documenting strength) to support the (c) diagnostic formulation (hypothesis), (d) outlines potential risks, both short- and longer-term, and (e) delineates a comprehensive management plan with (f) detailed recommendations.

Our overall approach is essentially a combined behavioral neurology/neuropsychological testing strategy (see also Batchelor, 1996). The core of our diagnostic method is the review of neurobehavioral systems, applied within the conceptual framework of the *brain-context-development* matrix. Our training program emphasizes rigor in observation and data collection from multiple sources, settings, and conditions. Our base clinical protocol includes direct analytic interview, standardized questionnaires, rating scales, and a set of psychological tests, all of which are used in the context of the review of neurobehavioral systems. The psychological tests are selected to provide normative information to "anchor" the clinician in developmental space, cover a relatively broad range of basic functional domains, and are matched to the expected competencies of children of different ages. The review of neurobehavioral systems plus the base protocol may provide sufficient information to answer the clinical question—or it may be used as the basis for directing more detailed analysis of specific domains.

SUMMARY

The goal of assessment, as we see it, is not a diagnosis, but a "portrait" (Matarazzo, 1990) of the whole child. Assessment thus needs to respond to the dynamic complexity that is the cardinal feature of the developing child. Models of assessment need to be scrutinized and (re-)formulated to incorporate the rapid advances in our understanding of the developing nervous system and the behavior that it both permits and constrains. This ongoing challenge of integrating knowledge with clinical practice and of passing on what we have learned to a new generation of practitioners is one that, for both students and teachers, can be expected to be difficult, but will certainly be exciting.

NOTES

1. In this discussion, "horizontal" refers to events, observations that are happening, and are analyzed, at the same stage in development. "Vertical" refers to those processes that have impact on behavior over time.

2. We note that this thinking has itself been shaped by the fact that neuropsychology originated as an "adult" discipline, developed in the context of an organism whose behavioral competencies are relatively "modular" (Fodor, 1983) and thus can be well characterized in terms of cognitive architecture. This is not, however, true for children. The thinking nonetheless has been downwardly extended to children!

3. Integrating the role of the clinician into research design bothers many scientists as introducing "subjective" (and difficult-to-control) variance. Objectivity is, after all, the gold standard for the generalizability of science in the positivist tradition. But, the notion that truth (about human behavior) can be expressed in causal relationships that are independent of time and place has been challenged, especially in the social sciences, as ahistorical and acontextual. In interactions between people, subject and object are humanly linked, social knowledge always interpreted within historical contexts (see Westcott, 1979). Development, a process that unfolds over time, cannot be other than historical with children necessarily depending on (adult) others in their environment to promote their development over time. From this perspective, the assumptions of scientific positivism are in direct conflict with a developmental analysis. In such an analysis, the end, to which "objectivity" is one means, will need to be achieved via alternative strategies.

4. Given this, we are entertained by the tension, in the professional psychological-test manu-

als, between the exhortations to test givers to establish *rapport*, on the one hand, and the "rules" of test administration with their all-too-often wooden delivery styles! How one establishes rapport is rarely discussed in detail; how one does it with the language provided by the manuals is not addressed (see also Vanderploeg, 1994).

5. Indeed, in our practice, this is becoming a serious practical problem. So many children with neurodevelopmental disorders, who were once condemned to retardation at best, are now so well managed medically and surgically in their early years that they may achieve normal intellectual function and thus participate in mainstream education. They are, however, in our "off-developmental-track" group; thus their learning disorders are not typically those of the educationally defined "learning-disabled child," where the emphasis is on academic skills, but may more importantly reflect deficits in basic information-processing functions, behavioral regulation, and executive function. Teachers and school psychologists all too frequently attempt, however, to fit them into the (cognitive; learning disability) theoretical framework of the educational setting with, often, painful results (see Bernstein, 1996a).

6. This approach to neuropsychological analysis, although very narrow in view to most neuropsychologists, may seem very compatible with the practice of educational and school psychologists in that many of the psychological tests used for neuropsychological analysis are already the (well-established) tools of their trade. Although a thorough grounding in neuropsychology is increasingly offered as part of the training of school/educational psychologists, psychologists currently practicing in the educational setting may well have added "neuropsychological tests" to their armamentarium by way of workshops or presentations without having the opportunity to develop the knowledge base and the assessment methodology that support the use of the tests. The fact that, in too many school systems, psychologists are constrained in the use of their professional skills to the role of (essentially) a psychometric technician (as contrasted with a clinician capable of integrating complex and wide-ranging behavioral information) may further insulate them from the wider developmental context of the individual child, making it difficult to keep in view the "whole child" perspective which we believe is crucial to the formulation of a developmental neuropsychology.

Models of credentialling in neuropsychology currently being developed will, we assume, render "pure" cognitive-ability structure models obsolete. Such credentialling models require coursework in the neurosciences, in behavioral neurology, in assessment, in normal and atypical development, in education, and so on, as well as applied clinical practice at pre- and postdoctoral levels. Thus, in due course, approaches dealing only with the child"s current behavioral repertoire could no longer be characterized as "neuropsychological".

7. In the training setting the research design itself and its role in answering the clinical investigative question(s) must be made explicit and directly taught. It is not adequate to simply expose students to already established assessment approaches which they must learn; they must analyze and understand the rationale for the protocol they are using in any given setting.

8. Note that a major goal of (clinical) research design is to reduce the occurrence of clinician error. All assessment approaches must attempt to do this. None of them succeed fully. All assessment approaches have their own particular "blind spots" and all must incorporate controls not only for the more general biases to which all clinicians are prone, but also for those that are particular to their strategy.

REFERENCES

Als, H. (1982). Towards a synactive theory of development: Promise for the assessment of infant individuality. *Infant Mental Health Journal, 3*, 229–243.

Anastasi, A. (1988). *Psychological Testing* (6th ed.). New York: MacMillan. 1988.

Arkes, H. (1981). Impediments to accurate clinical judgment and possible ways to minimize their impact. *Journal of Consulting and Clinical Psychology, 49*, 323-330.

Bakker, D. J. (1984). The brain as a dependent variable. *Journal of Clinical Neuropsychology, 6*, 1–16.

Banaji, M. (1996). Automatic stereotyping. *Psychological Science, 7*, 136–141.

Baron, I. S., Fennell, E., & Voeller, K. S. (1995). *Pediatric Neuropsychology in the Medical Setting*. New York: Oxford University Press.

Batchelor, E. S. (1996). Neuropsychological assessment of children. In E. S. Batchelor & R. S. Dean (Eds.), *Pediatric neuropsychology: Interfacing assessment and treatment for rehabilitation*. Boston: Allyn & Bacon.

Batchelor, E. S., & Dean, R. S. (Eds.). (1996). *Pediatric neuropsychology: Interfacing assessment and treatment for rehabilitation*. Boston: Allyn & Bacon.

Bernstein, J. H. (1996a, Fall). Issues in the psychoeducational management of children treated for malignant disease. *P.O.G.O. (Pediatric Oncology Group of Ontario) News*, 7(2), 9-11.

Bernstein, J. H. (1996b). Behavioral observation scale for clinical examiners (BOSCE). Working draft for research investigation. Unpublished manuscript.

Bernstein, J. H. (1999). Developmental neuropsychological assessment. In K. O. Yeates, D. Ris, & H. G. Taylor (Eds.), *Pediatric neuropsychology: Research, theory and practice*. New York: Guilford Press.

Bernstein, J. H., Prather, P. A., & Rey-Casserly, C. (1995). Neuropsychological assessment in pre and post-operative evaluation. *Neurosurgery Clinics, 6*, 443–454.

Bernstein, J. H., & Waber, D. P. (1990). Developmental neuropsychological assessment: The systemic approach. In: A. A. Boulton, G. B. Baker, & M. Hiscock (Eds.), *Neuromethods: Vol. 17, Neuropsychology*. Clifton, NJ: Humana Press.

Bernstein, J. H., & Waber, D. P. (1997). Pediatric neuropsychological assessment. In T. Feinberg & M. Farah. (Eds.), *Behavioral neurology and neuropsychology*. New York: McGraw-Hill.

Bernstein, N. (1967). *Coordination and regulation of movements*. New York: Pergamon Press

Best, C. T. (1988). The emergence of cerebral asymmetries in early human development: A review and a neuroembryological model. In D. L. Molfese & S.J. Segalowitz (Eds.), *Brain lateralization in children: Developmental implications*. New York: Guilford Press.

Bornstein, R. A. (1990). Neuropsychological test batteries in neuropsychological assessment. In A. A. Boulton, G. B. Baker, & M. Hiscock (Eds.), *Neuromethods: Vol. 17, Neuropsychology*. Clifton, NJ: Humana Press.

Bronfenbrenner, U. (1993). The ecology of cognitive development: research models and fugitive findings. In R. H. Wozniak & K. W. Fischer (Eds.), *Development in context: Acting and thinking in wpecific environments. The Jean Piaget Symposium Series*. Hillsdale, NJ:; Lawrence Erlbaum Associates.

Bronfenbrenner, U., & Ceci, S. (1994). Nature-nurture reconceptualized in developmental perspective: A bioecological model. *Psychological Review, 101*, 568–586..

Butters, N (1984). The clinical aspects of memory disorders: Contributions from experimental studies of amnesia and dementia. *Journal of Experimental Neuropsychology, 6*, 17–36.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105.

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

Caramazza, A. & H, A. E. (1991). Lexical organization of the nouns and verbs in the brain. *Nature, 349*, 788–790.

Carlson, D. F., & Harris, L. J. (1985). Development of the infant's hand preference for visually directed reaching: Preliminary report of a longitudinal study. *Infant Mental Health Journal, 6*, 158-174.

Case, R. (1992). The role of the frontal lobes in the regulation of cognitive development. *Brain and Cognition, 20*, 51–73.

Case, R., & Okamoto, Y. (1996). *The role of central conceptual structures in the development of children's thought. Monographs of the Society for Research in Child Development, 61*, (Serial Nos. 1–2).

Ceci, S. J. (1996). *On Intelligence: A bioecological treatise on intellectual development* (2nd ed.). Cambridge, MA: Harvard University Press.

Chapman, L. J., & Chapman, J. P. (1967). Genesis of popular but erroneous psychodiagnostic observations. *Journal of Abnormal Psychology, 72*, 193-204.

Chelune, G. J., & Edwards, P. (1981). Early brain lesions: Ontogenetic-environmental considerations. *Journal of Consulting and Clinical Psychology, 49*, 777–790.

Cimino, C. R. (1994). Principles of neuropsychological interpretation. In R. D. Vanderploeg (Ed.), *Clinician's guide to neuropsychological assessment*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, M. J., Branch, W. B., Willis, W. G., Weyandt, L. L., & Hynd, G. W. (1992). Childhood. In A. E. Puente & R. J. McCaffrey (Eds.), *Handbook of neuropsychological assessment*. New York: Plenum Press.

Conell, J. (1939–1963). *The postnatal development of the human cerebral cortex, Vols. 1–6*. Cambridge, MA: Harvard University Press.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston: Houghton Mifflin Company.

Damasio, A. R. (1989). Time-locked multiregional retroactivation: A systems-level proposal for the neural substrates of recall and recognition. *Cognition, 33*, 25–62.

Damasio, A. R. (1990). Category-related recognition defects as clues to the neural substrates of knowledge. *Trends in Neuroscience, 13*, 95–98.

Dawson, G., & Fischer, K. W. (1994). *Human behavior and the developing brain*. New York: Guilford Press.

Dennis, M (1983). The developmentally dyslexic brain and the written language skills of children with one hemisphere. In U. Kirk (Ed.), *Neurop-*

sychology of language, reading and spelling. New York: Academic Press.

Dennis, M. (1988). Language and the young damaged brain. In T. Boll & B. K. Bryant (Eds.), *Clinical neuropsychology and brain function.* Washington, DC: American Psychological Association.

Dennis, M. (1989). Assessing the neuropsychological abilities of children and adolescents for personal injury litigation. *The Clinical Neuropsychologist, 3,* 203–229.

Diamond, A. (1991). Neuropsychological insights into the meaning of object concept development. In S. Carey & R. Gelman (Eds.), *The epigenesis of mind. Essays on biology and cognition.* Hillsdale, NJ: Erlbaum.

Draeger, S., Prior, M., & Sanson, A. (1986). Visual and auditory attention performance in hyperative children: Competence or compliance? *Journal of Abnormal Child Psychology, 14,* 411-424.

Edelman, G. M. (1987). *Neural Darwinism.* New York: Basic Books.

Ellwood, R. W. (1993). Clinical discriminations and neuropsychological tests: An appeal to Bayes' theorem. *The Clinical Neuropsychologist, 7,* 224-233.

Emory, E. K (1991). A neuropsychological perspective on perinatal complications and the law. *The Clinical Neuropsychologist, 5,* 297–321.

Emory, E. K., Savoie, T. M., Ballard, J., Eppler, M., & O'Dell, C. (1992). Perinatal factors in clinical neuropsychological assessment. In A. E. Puente & R. J. McCaffrey (Eds.), *Handbook of Neuropsychological Assessment: A biopsychosocial perspective.* New York: Plenum Press. 1992.

Faust, D., & Nurcombe, B. (1989). Improving the accuracy of clinical judgment. *Psychiatry, 52,* 197-208.

Fennell, E. B. (1994). Issues in neuropsychological assessment. In R. D. Vanderploeg (Ed.), *Clinician's guide to neuropsychological assessment.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Fennell, E. B., & Bauer, R. M. (1997). Models of inference in evaluating brain-behavior relationships in children. In C. R. Reynolds, E. Fletcher-Jantzen (Eds.), *Handbook of clinical child neuropsychology* (Rev. ed.). New York: Plenum Press.

Fink, R. P. (1996). Successful dyslexics: A constructivist study of passionate interest in reading. *Journal of Adolescent and Adult Reading, 39,* 268–280.

Fisch, H. U., Hammond, K. R., & Joyce, C. R. B. (1982). On evaluating the severity of depression: An experimental study of psychiatrists. *British Journal of Psychiatry, 140,* 378-383

Fischer, K. W., & Rose, S. P. (1994). Dynamic development of coordination of components in brain and behavuir. In G. Dawson & K. W. Fischer (Eds.), *Human behavior and the developing brain.* New York: Guilford Press.

Fletcher, J. M., & Taylor, H. (1984). Neuropsychological approaches to children: Towards a developmental neuropsychology. *Journal of Clinical Neuropsychology, 6,* 39–56.

Fletcher, J. M., Taylor, H. G., Levin, H. S., & Satz, P. (1995). Neuropsychological and intellectual assessment of children. In H. I. Kaplan & B. Sadock (Eds.), *Comprehensive textbook of psychiatry.* Baltimore, MD: Williams & Wilkins.

Fodor, J. (1983). *Modularity of Mind.* Cambridge, MA: MIT Press.

Ford, D. H. (1987). *Humans as self-constructing living systems: A developmental perspective on behavior and personality.* Hillsdale, NJ: Erlbaum.

Franzen, M. D. (1989). *Reliability and validity in neuropsychological assessment.* New York: Plenum Press.

Gauron, E. F., & Dickinson, J. K. (1966). Diagnostic decision making in psychiatry. I. Information usage. II. Diagnostic styles. *Archives of General Psychiatry, 14,* 225-232, 233-277.

Gauron, E. F, & Dickinson, J. K. (1969). The influence of seeing the patient first on diagnostic decision-making in psychiatry. *American Journal of Psychiatry, 126,* 199–205.

Globus, G. G. (1973). Consciousness and brain, I. The identity thesis. *Archives of General Psychiatry, 29,* 153–160.

Goldberg, L. (1968). Simple models or simple processes? Some research on clinical judgments. *American Psychologist, 23,* 483-496.

Golden, M. (1964). Some effects of combining psychological tests on clinical inferences. *Journal of Consulting Psychology, 28,* 440-446.

Goldfield, E. C. (1995). *Emergent Forms: Origins and early development of human action and perception.* New York: Oxford University Press.

Goldstein, G., & Shelly, C. (1984). Discriminative validity of various intelligence and neuropsychological tests. *Journal of Clinical and Consulting Psychology, 52,* 383–389.

Gopnik, A., & Meltzoff, A. N. (1997). *Words, thoughts, and theories.* Cambridge, MA: MIT Press.

Gould, S. J. (1996). *The mismeasure of man,* (2nd ed). New York: W.H. Norton.

Greenwald, A., & Banaji, M. R. (1995). Implicit social cognition: Attitude, self-esteem and stereotypes. *Psychological Review, 102,* 4–27.

Hadenius, A. M., Hagberg, B., Hyttnes-Bensch, K., & Sjogren, I. (1962). The natural prognosis of

infantile hydrocephalus. *Acta Paediatrica (Uppsala), 51*, 117–121.

Hartlage, L. C., & Telzrow, C. F. (1986). *Neuropsychological assessment and intervention with children and adolescents*. Sarasota, FL: Professional Resources Exchange.

Heaton, R. K., Grant, I., & Matthews, C. G. (1991). *Comprehensive norms for an expanded Halstead-Reitan battery: Demographic corrections, research findings and clinical applications*. Odessa, FL: Psychological Assessment Resources.

Henderson, L. J. (1935). Physician and patient as a social system. *New England Journal of Medicine, 212*, 819–823.

Herrnstein, R. J. & Murray, C. (1994). *The bell curve*. New York: Free Press.

Hoffman, P. J., Slovic, P., & Rorer, L. G. (1968). An analysis-of-variance model for the assessment of configural cue utilization in human judgment. *Psychological Bulletin, 69*, 338-349.

Hoffman, R. R., & Deffenbacher, K. A. (1993). An analysis of the relations between basic and applied psychology. *Ecological Psychology, 5*, 315–352.

Holmbeck, G. N., & Faier-Routman, J. (1995). Spinal lesion level, shunt status, family relationships, and psychosocial adjustment in children and adolescents with spina bifida myelomeningocele. *Journal of Pediatric Psychology, 20*, 817–832.

Hynd, G. W., & Willis, W. G. (1988). *Pediatric neuropsychology*. Orlando, FL.: Grune & Stratton.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3*, 430-454.

Kaplan, E. (1976, August). The role of the non-compromised hemisphere in patients with local brain disease. In H.-L. Teuber (Chair), *Alterations in brain functioning and changes in cognition*. Symposium conducted at the Annual Meeting of the American Psychological Association.

Kaplan, E. (1988). A process approach to neuropsychological assessment. In T. Boll & B. K. Bryant (Eds.), *Clinical neuropsychology and brain function*. Washington, DC: American Psychological Association.

Kennedy, M. L., Willis, W. G., & Faust, D. (1997). The base rate fallacy in school psychology. *Journal of Psycho-educational Assessment, 15*, 292-307.

Kerlinger, F. N. (1986). *Foundations of behavioral research* (3rd ed). Fort Worth, TX: Holt, Rinehart & Winston.

Kolb, B. (1989). Brain development, plasticity and behavior. *American Psychologist, 44*, 1203–1212.

Kolb, B. (1995). *Brain plasticity and behavior*. Hillsdale, NJ: Erlbaum.

Kolb, B., & Whishaw, I. Q. (1985). *Fundamentals of human neuropsychology* (2nd ed.). New York: W.H. Freeman.

Korkman, M., Kirk, U., & Kemp, S. (1997). *NEPSY: Comprehensive assessment of neuropsychological development in children*. San Antonio, TX: Psychological Corporation.

Krasnegor, N. A., Lyon, G. R., & Goldman-Rakic, P. (1997). *Development of the prefrontal cortex*. Baltimore: Paul H. Brookes Publishing.

Lezak, M. (1995). *Neuropsychological Assessment* (3rd ed.). New York: Oxford University Press.

Light, R., Satz, P., Asarnow, R. F., Lewis, R., Ribbler, A., & Neumann, E. (1996). Disorders of attention. In E. S. Batchelor & R. S. Dean (Eds.), *Pediatric neuropsychology: Interfacing assessment and treatment for rehabilitation*. Boston: Allyn & Bacon.

Lowman, R. L. (1996). Introduction to the special series on what every psychologist should know about assessment. *Psychological Assessment, 8*, 339-340.

Luria, A. K. (1973). *The working brain*. New York: Basic Books.

Luria, A. K. (1980). *Higher cortical functions in man* (2nd ed.). New York: Basic Books.

Maloney, M. P., & Ward, M. P. (1976). *Psychological assessment: A conceptual approach*. New York: Oxford University Press.

Matarazzo, J. D. (1990). Psychological assessment versus psychological testing. *American Psychologist, 45*, 999–1017.

Mattis, S. (1992). Neuropsychological assessment of school-aged children. In I. Rapin & S. J. Segalowitz (Eds.), *Handbook of neuropsychology, Volume 6: Child neuropsychology*. Amsterdam: Elsevier.

Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin, 52*, 194-216.

Mesulam, M.-M. (1981). A cortical network for directed attention and unilateral neglect. *Annals of Neurology, 10*, 309–325.

Mesulam, M.-M. (1990). Large scale neurocognitive networks and distributed processing for attention, language and memory. *Annals of Neurology, 28*, 597–613.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*, 81-97.

Molfese, D. L., & Segalowitz, S. J. (1988). *Brain lateralization in children: Developmental implications*. New York: Guilford Press.

Murphy, E. A. (1979). *Probability in Medicine*. Baltimore, MD: Johns Hopkins Press.

Neisser, U., Boodoo, G., Bouchard, T. J., Boykin, A. W., Brody, N., Ceci, S. J., Halpern, D. F., Loehlin, J. C., Perloff, R., Sternberg, R. J., & Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist, 51,* 77-101.

Nisbett, R. E., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment.* Englewood Cliffs, NJ: Prentice-Hall.

Nisbett, R. E., & Wilson, T. D. (1977). The halo effect: Evidence for unconscious alteration of judgments. *Journal of Personality and Social Psychology, 35,* 250-256.

Oskamp, S (1965). Overconfidence in case-study judgments. *Journal of Consulting Psychology, 29,* 261-265.

Pennington, B. F. (1991). *Diagnosing learning disorders.* New York: Guilford Press.

Piaget, J. (1952). *The origins of intelligence in children* (M. Cook, Trans.). New York: International Universities Press. (Original work published 1936)

Pirozzolo, F. J., & Bonnefil, V. (1996). Disorders appearing in the perinatal and neonatal period. In E. S. Batchelor & R. S. Dean (Eds.), *Pediatric neuropsychology: Interfacing assessment and treatment for rehabilitation.* Boston: Allyn & Bacon. 1996.

Prigatano, G. P., & Redner, J. E (1993). Uses and abuses of neuropsychological testing in behavioral neurology. *BNI Quarterly, 9,* 22–29.

Rabinowitz, J. (1994). Guide to identifying and correcting decision-making errors in mental disability practice. *Bulletin of the American Academy of Psychiatry and Law, 22,* 561–575.

Rodier, P. M. (1994). Vulnerable periods and processes during central nervous system development. *Environmental Health Perspectives, 102* (Suppl.2), 121–124.

Rorer, L. G., Hoffman, H. D., Dickman, H. D., & Slovic, P. (1967). Configural judgments revealed (summary). *Proceedings of the 75th Annual Convention of the American Psychological Association, 2,* 195-196.

Rosenthal, R., & Rosnow, R. L. (1984). *Essentials of behavioral research: Methods and data analysis.* New York: McGraw-Hill.

Ross, L. D., Lepper, M. R., Strack, F., & Steinmetz, J. (1977). Social explanation and social expectation: Effects of real and hypothetical explanations on subjective likelihood. *Journal of Personality & Social Psychology, 35,* 817-829.

Rourke, B. P. (1975). Brain-behavior relationships in children with learning disabilities: A research program. *American Psychologist, 30,* 911–920.

Rourke, B. P. (1982). Central processing deficiencies in children: Toward a developmental neuropsychological model. *Journal of Clinical Neuropsychology, 4,* 1–18.

Rourke, B. P. (1989). *Nonverbal learning disabilities: The syndrome and the model.* New York: Guilford Press.

Rourke, B. P. (1994). Neuropsychological assessment of children with learning disabilities. In G.R. Lyon (Ed.), *Frames of reference for the assessment of learning disabilities.* Baltimore, MD: Paul H. Brookes Publishing.

Rourke, B. P., Bakker, D. J., Fisk, J. L., & Strang, J. D. (1983). *Child neuropsychology: An introduction to theory, research and clinical practice.* New York: Guilford Press.

Rourke, B. P., Fisk, J. L., & Strang, J. D. (1986). *Neuropsychological assessment of children.* New York: Guilford Press.

Sadker, M., & Sadker, D. (1994). *Failing at fairness: How our schools cheat girls.* New York: MacMillan.

Schneider, W., & Pressley, M. (1990). *Memory development between 2 and 20 years.* New York: Springer-Verlag.

Segalowitz, S. J., & Hiscock, M. (1992). The emergence of a neuropsychology of normal development: Rapprochement between neuroscience and developmental neuropsychology. In I. Rapin & S. J. Segalowitz (Eds.), *Handbook of neuropsychology, Volume 6: Child neuropsychology.* Amsterdam: Elsevier.

Segalowitz, S. J., & Rose-Krasnor, L. (1992). The construct of brain maturation in theories of child development. *Brain and Cognition, 20, 1-7.*

Shaheen, S. J. (1984). Neuromaturation and behavior development: The case of childhood lead poisoning. *Developmental Psychology, 20,* 542-550.

Shankweiler, D., Liberman, I. Y., Mark, L. S., Fowler, C. A., & Fischer, F. W. (1979). The speech code and learning to read. *Journal of Experimental Psychology: Human Learning and Memory, 5,* 531-545.

Slife, B. D., & Williams, R. N. (1997). Toward a theoretical psychology: Should a subdiscipline be formally recognized? *American Psychologist, 52,* 117–129.

Smith, L. B., & Thelen, E. (Eds.). (1993). *A dynamic systems approach to development: Applications.* Cambridge, MA: MIT Press.

Spreen O., Risser, A. T., & Edgell, D. (1995). *Developmental neuropsychology* (2nd ed.). New York: Oxford University Press.

Taylor, H. G (1988). Neuropsychological testing: Relevance of assessing children's learning disabilities. *Journal of Consulting and Clinical Psychology, 56,* 795–800.

Taylor, H. G., & Fletcher, J. M. (1990). Neuropsychological assessment of children. In G. Goldstein & M. Hersen (Eds.), *Handbook of Psychological Assessment* (2nd ed.). New York: Pergamon.

Taylor, H. G., & Fletcher, J. M. (1995). Editorial: Progress in pediatric neuropsychology. *Journal of Pediatric Psychology, 20,* 695–701.

Teeter, P. A., & Semrud-Clikeman, M. (1997). *Child neuropsychology: Assessment and interventions for neurodevelopmental disorders.* Boston: Allyn & Bacon.

Teuber, H. L. (1974). Why two brains? In F. O. Schmitt & F. G. Worden (Eds.), *The neurosciences: Third study program.* Cambridge, MA: MIT Press.

Thatcher, R. W. (1989). *Nonlinear dynamics of human cerebral development.* Paper presented at the First International Conference, Mechanisms of Mind, Havana, Cuba.

Thatcher, R. W. (1992). Cyclic cortical reorganization during early childhood development. *Brain and Cognition, 20,* 24–50.

Thelen, E., & Smith, L. B. (1994). *A dynamic systems approach to the development of cognition and action.* Cambridge, MA: MIT Press.

Tramontana, M. G., & Hooper S. R. (1988). Child neuropsychological assessment: Overview of current status. In M. G. Tramontana & S. R. Hooper (Eds.), *Assessment issues in Child neuropsychology.* New York: Plenum.

Trevarthen, C. (1996). Lateral asymmetries in infancy: Implications for the development of the hemispheres. *Neuroscience and Behavioral Reviews, 20,* 571-586.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and Biases. *Science, 183,* 1124-1131.

Vanderploeg, R. D. (1994). Interview and testing: The data collection phase of neuropsychological evaluations. In R. D. Vanderploeg (Ed.), *Clinician's guide to neuropsychological assessment.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Vygotsky, L. (1986). *Thought and language.* Cambridge, MA: MIT Press.

Waber, D. P. (1976). Sex differences in cognition: A function of maturational rate. *Science, 192,* 572-573.

Waber, D. P. (1989). Rate and state: A critique of models guiding the assessment of learning disordered children. In P. R. Zelazo & R. G. Barr (Eds.), *Challenges to developmental paradigms: Implications for theory, assessment and treatment.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Waber, D. P., Bernstein, J. H., Kammerer, B. L., Tarbell, N. J., & Sallan, S. E. (1992). Neuropsychological diagnostic profiles of children who received CNS treatment for acute lymphoblastic leukemia: The systemic approach to assessment. *Developmental Neuropsychology, 8,* 1–28.

Waber, D. P., & Tarbell, N. J. (1997). Toxicity of CNS prophylaxis in childhood leukemia. *Oncology, 11,* 259–263.

Waber, D. P., Urion, D. K., Tarbell, N. J., Niemeyer, C., Gelber, R., & Sallan, S. E. (1990). Late effects of central nervous system treatment of acute lymphoblastic leukemia are sex dependent. *Developmental Medicine and Child Neurology, 32,* 238–248.

Walsh, K. W. (1992). Some gnomes worth knowing. *The Clinical Neuropsychologist, 6,* 119–133.

Warrington, E. K., & Shallice, T. (1984). Category specific semantic impairments. *Brain, 107,* 829–854.

Wechsler, D. (1981). *Wechsler Adult Intelligence Scale-Revised (WAIS-R).* New York: Psychological Corporation.

Wechsler, D. (1991). *Wechsler Intelligence Scale for Children, 3rd edition (WISC-III).* San Antonio, TX: Psychological Corporation.

Wedding, D. (1983a). Clinical and statistical prediction in neuropsychology. *Clinical Neuropsychology, 5,* 49-55.

Wedding, D. (1983b). Comparison of statistical and actuarial models for predicting lateralization of brain damage. *Clinical Neuropsychology, 5,* 15-20.

Wedding, D., & Faust, D. (1989). Clinical judgment and decision making in neuropsychology. *Archives of Clinical Neuropsychology, 4,* 233-265.

Westcott, M. (1979). Feminist criticism of the social sciences. *Harvard Educational Review, 49,* 422–430.

Wiggins, N., & Hoffman, P. J. (1968). Three models of clinical judgment. *Journal of Abnormal Psychology, 73,* 70-77.

Willis, W. G. (1986). Actuarial and clinical approaches to neuropsychological diagnosis: Applied considerations. In J. E. Obrzut & G. W. Hynd (Eds.), *Child neuropsychology, Volume 2: Clinical practice.* Orlando, FL: Academic Press.

Wolff, P. H. (1987). *The development of behavioral states and the expression of emotions in early infancy.* Chicago: University of Chicago Press.

Yakovlev, P. I., & LeCours, A.-R. (1967). The myelogenetic cycles of regional maturation of the brain. In A. Minkowsky (Ed.), *Regional development of the brain in early life.* Oxford, U.K.: Blackwell Scientific Publications.

CHAPTER 12

SPECIALIZED NEUROPSYCHOLOGICAL ASSESSMENT METHODS

Glenn J. Larrabee, Ph.D.

Clinical neuropsychological assessment is the measurement and analysis of the cognitive, behavioral, and emotional consequences of brain damage or dysfunction. Historically, neuropsychological assessment has had a variety of influences including behavioral neurology, psychometrics and test construction, and experimental psychology (chapter 10, this volume; Walsh, 1987).

Neuropsychological assessment has been characterized by two basic approaches: (1) the fixed battery approach exemplified by the Halstead-Reitan battery and the Luria-Nebraska neuropsychological battery (Reitan & Wolfson, 1993; Golden, Purisch, & Hammeke, 1985); and (2) the ability-focused, cognitive process, hypothesis-testing approach, exemplified by the Benton-Iowa Group (Benton, Sivan, Hamsher, Varney, & Spreen, 1994), Lezak (1995), and the Boston group (Milberg, Hebben, & Kaplan, 1996). It is an oversimplification to characterize these two general orientations as (a) a fixed battery, without modification, administered to all patients regardless of complaint or reason of referral, versus (b) flexible, but inconsistent across patients. In actual practice, the core Halstead-Reitan battery and the Luria-Nebraska are frequently administered in conjunction with measures of intelligence and memory such as the Wechsler Adult Intelligence Scale-Revised (WAIS-R) (Wechsler, 1981) or Wechsler Memory Scale-Revised (WMS-R) (Wechsler, 1987). Indeed, the recently developed comprehensive norms for the Halstead-Reitan Battery were co-normed with the WAIS and other measures of memory and language (Heaton, Grant, & Matthews, 1991). The process, ability-focused approaches typically include a standard core set of procedures, usually assessing memory and intelligence, which are augmented by additional flexible-adjustive exploration of cognitive deficits specific to the particular patient undergoing evaluation (Hamsher, 1990; Lezak, 1995; Milberg, Hebben, & Kaplan, 1996).

More recently, Bauer (1994) has discussed an approach which he characterizes as intermediate to the fixed and flexible battery approach: multiple fixed batteries. Bauer distinguishes three types of multiple fixed batteries, and provides several examples of each type. The first, a general "screening" battery, is comprised of items maximally sensitive to clinically significant abnormalities requiring more detailed exploration with additional testing. A second alternative is the "population specific" battery for evaluation of specific patient populations or disease entities (e.g., HIV seropositive status, cf. Butters, Grant, Haxby, Judd, Martin, McClelland, Pequegnat, Schacter, & Stover, 1990), wherein the goal is to provide a selective but standardized evaluation of the cognitive functions most relevant to diagnosis and treatment of individu-

als within the specific population. Finally, batteries can be "domain specific," providing a detailed evaluation of particular neurobehavioral functions such as language (e.g., Boston Diagnostic Aphasia Examination: Goodglass & Kaplan, 1983) or memory (WMS-R: Wechsler, 1987).

Various interpretive strategies have been developed for distinguishing between normal and abnormal neuropsychological test performance. One strategy, which developed out of the fixed battery approach, was the determination of an optimal "brain-damage cutting score" that maximally separated a brain-damaged sample from a normal sample (cf. Reitan & Wolfson, 1993; Golden et al., 1985). Certainly, one would not dispute the fact that a neuropsychological test should be sensitive to brain damage or dysfunction; however, the "brain-damage cutting score" approach is dependent on a variety of factors including demographic characteristics (age, education, gender) of the brain-damaged and control groups, the nature and severity of brain damage or dysfunction in the brain-damaged group, as well as on where the cutting score is set (Lezak, 1995). Again, it is an oversimplification to characterize battery approaches as based only on optimal "cutting scores;" for example, Reitan has supported a four-tiered method of analysis including level of performance, pattern of performance, comparison of lateralized sensorimotor processes, and analysis of pathognomonic signs (Reitan & Wolfson, 1993).

An alternative approach, associated with the ability-focused, hypothesis-testing orientation, is to evaluate each cognitive function relative to the range of performance in a representative normal sample, adjusting for age, gender, education, and other relevant demographic factors. This approach is analogous to the ranges of normality developed for laboratory values in clinical medicine. This approach, which references normality to a normal control sample, remains dependent on the representativeness of the normal sample, as well as on the level of performance at which the interpretation of abnormality is set (referred to as "abnormal-performance cutting score"). The advantage of this approach over the more traditional "brain-damage cutting score" approach, is that one is not dependent on the variability inherent in a mixed brain-damaged population, and the "abnormal performance cutting score" can be set at a value that keeps the false positive error rate constant. Benton, Sivan, and colleagues (1994) typically set the cutoff for performance abnormality to match the bottom 5 percent of control-subject test performance. Heaton and colleagues (1991) set this value at −1.1 SD (T score of 39 or less), which defines abnormality as performance lower than approximately 86 percent of normal control subjects.

A clinician utilizing the ability-focused, hypothesis-testing approach evaluates the patterns of cognitive strengths and weaknesses of a particular patient relative to one another, considers these patterns in light of the referral question and other clinical/historical data, and integrates these data to form a diagnostic impression and clinical recommendations (Lezak, 1995). As Walsh (1995) has noted, the hypothesis-testing approach is dependent on two major elements: (1) current familiarity with the body of knowledge of neuropsychological findings in relation to neurological disorders, and (2) personal experience of the clinician with as wide a range of neurologic disorders as possible, seen at various stages of evolution and resolution.

OVERVIEW OF SPECIALIZED NEUROPSYCHOLOGICAL PROCEDURES

Lezak (1995) has characterized neuropsychology as one of the most rapidly evolving fields in the clinical sciences. Despite the vast proliferation of specialized tests that have been developed since the first edition of Lezak's book (Lezak, 1976), several reviews of the field on neuropsychology (Lezak, 1995; Mapou & Spector, 1995; Spreen & Strauss, 1998), have identified seven major functional areas: (1) language and related verbal and communicative functions, (2) spatial/perceptual skills, (3) sensorimotor functions, (4) attention- and related information-processing tasks, (5) memory (verbal, visual, remote), (6) intellectual and problem-solving skills (including "executive" functions), and (7) emotional and adaptive functions. These rationally defined areas are supported by recent factor analyses of comprehensive test batteries conducted by Larrabee and Curtiss (1992) and Leonberger, Nicks, Larrabee, and Goldfader (1992). Table 12.1 displays the results of a factor analysis of the WAIS-R, the WMS-R, and the Halstead-Reitan Neuropsychological Battery (HRNB) (Leonberger et al., 1992). This analysis, employing the delayed-recall WMS-R sub-

Table 12.1. Factor Loadings of the Wechsler Memory Scale-Revised, Wechsler Adult Intelligence Scale-Revised, and the Halstead-Reitan Neuropsychological Test Battery: Analysis of Delayed Recall Scores

	FACTOR				
MEASURE	1	2	3	4	5
Wechsler Memory Scale-Revised					
Mental Control	.06	.36	.17	.46	.31
Figural Memory	.23	.04	.36	.04	.19
Digit Span	.09	.31	.11	.69	.06
Visual Memory Span	.50	.07	.12	.34	.28
Logical Memory II	.10	.31	.67	.06	.02
Visual Paired Associate II	.32	.10	.60	.09	.14
Verbal Paired Associates II	.09	.10	.76	.19	.14
Visual Reproduction II	.55	.04	.49	.15	.19
Wechsler Adult Intelligence Scale-Revised					
Information	.07	.82	.05	.15	.07
Vocabulary	.07	.88	.13	.16	.09
Arithmetic	.16	.56	−.01	.42	.09
Comprehension	.22	.76	.10	.16	.04
Similarities	.13	.74	.22	.14	.02
Picture Completion	.62	.22	.17	.17	.07
Picture Arrangement	.59	.22	.18	.18	.01
Block Design	.76	.18	.02	.18	.24
Object Assembly	.80	.13	.05	−.03	.15
Digit Symbol	.42	−.08	.40	.12	.50
Halstead-Reitan Neuropsychological Test Battery					
Category Test (VII)	−.51	−.11	−.35	−.18	−.05
Tactual Performance Test (location)	.54	−.06	.31	−.02	.11
Speech Sounds Perception Test	−.16	−.17	−.34	−.42	−.32
Rhythm Test	−.22	−.19	−.11	−.59	.06
Finger Tapping Test (dominant hand)	.11	.08	.08	.01	.37
Trail Making Test (Part B)	−.43	−.05	−.33	−.26	−.47

Note: $n = 237$, orthogonal rotation. From "Factor structure of the Wechsler Memory Scale-Revised within a comprehensive neuropsychological battery," by F.T. Leonberger, S.D. Nicks, G.J. Larrabee & P.R. Goldfader, 1992, *Neuropsychology, 6*, p. 245. Copyright 1992, Educational Publishing Foundation. Reprinted with permission of authors.

tests, and in an attempt to identify a memory component of the HRNB, Category Test subtest VII and TPT location, yielded five factors, identified by the authors as: (1) Nonverbal and Spatial Reasoning, (2) Verbal Comprehension and Expression, (3) Memory, (4) Attention and Concentration, (and 5) Psychomotor Speed. Table 12.2 displays the results of the factor analysis by Larrabee and Curtiss (1992), employing several of the Benton-Iowa tests (Benton, Sivan, et al., 1994), selected HRNB sensorimotor procedures (Heaton et al., 1991), specialized measures of attention and memory from the head trauma research literature (Levin, Benton, & Grossman, 1982), the WAIS-R, Wisconsin Card Sorting Test, and Wide Range Achievement Test-Revised (WRAT-R) (Jastak & Wilkinson, 1984). This factor analysis, employing delayed-recall memory tests, yielded six factors, identified by the authors as (1) General Verbal Ability and Problem Solving, (2) Visual/Nonverbal Problem Solving, (3) Memory, (4) Gross Motor Skills, (5) Attention/Information Processing, and (6) Finger Localization.

Tables 12.1 and 12.2 demonstrate two noteworthy findings. Concept-formation tasks considered to be related to frontal lobe functioning, such as the Categories Test, Wisconsin Card Sorting Test, and Trailmaking B are more closely associated with WAIS-R Per-

formance IQ subtests than with a separate "frontal cognitive" factor. Complex sensorimotor tasks such as the Purdue Pegboard, Grooved Pegboard, Benton-Iowa Tactile Form Perception, and HRNB Tactual Performance Test are more closely associated with WAIS-R Performance IQ subtests than with a separate sensorimotor factor.

GENERAL CONSIDERATIONS IN THE USE AND INTERPRETATION OF SPECIALIZED NEUROPSYCHOLOGICAL ASSESSMENT PROCEDURES

Selection of procedures for neuropsychological assessment should provide appropriate breadth and depth of evaluation. Some assessment should be

Table 12.2. Factor Structure of Neuropsychology Battery With Delayed Recall Memory Tests

	FACTORS					
VARIABLES	1	2	3	4	5	6
Visual Naming	.63					
Controlled Oral Word Assoc.	.35				.40	
Visual Form Discrimination			.34			
Judgment of Line Orientation	.31					
Facial Recognition		.41				
WMS Mental Control					.48	
Trailmaking B		−.74			−.31	
PASAT–Trial 4					.47	
Serial Digit Learning			.55			
Expanded Paired Assoc.–Delay			.54			
Selective Reminding–Delay			.81			
Visual Reproduction–Delay			.55			
Continuous Recog. Mem–Delay			.57			
Continous Visual Mem–Delay			.55			
Finger Tap–DOM				.73	.34	
Finger Tap–N DOM				.74	.35	
Grip–DOM				.89		
Grip–N DOM				.92		
Purdue Pegs–DOM		.39	.33		.42	
Purdue Pegs–N DOM		.69			.34	
Grooved Pegs–DOM		−.39	−.40			
Grooved Pegs–N DOM		−.84				
Tactile Form–DOM		.77				
Tactile Form–N DOM		.76				
Finger Localiz–DOM						.77
Finger Localize–N DOM						.71
WAIS-R Information	.84					
Digit Span	.38				.54	
Vocabulary	.92					
Similarities	.68					
Comprehension	.80					
Arithmetic	.64					
Picture Completion	.47					
Picture Arrangement	.49	.50				
Block Design		.55				
Object Assembly		.62				
Digit Symbol		.50	.31			
Wisconsin Persev. Errors		−.55				
WRAT-R Reading	.78				.35	
Spelling	.68				.43	
Arithmetic	.61				.36	

Note: n = 151, Oblique rotation. Range of factor intercorrelations is .02 (3 with 6) to −.45 (2 with 3). Loadings of .30 or higher are reported. From Larrabee and Curtiss, 1992.

made of each of the key neurobehavioral domains, including language, spatial processes, sensorimotor processes, attention, memory, intelligence/problem solving, and emotional/adaptive processes. Selection of tests should be based upon proven reliability and validity, utilizing procedures with sufficient normative data, which are corrected for demographic factors when these are empirically related to test performance in the normative sample. The present author follows the Benton-Iowa tradition of interpreting performance as "impaired" when exceeded by 95 percent of the normative sample and as "borderline" when falling between the 6th to 16th percentiles relative to normal controls.

The examination should start with an interview of the patient. This serves several purposes including establishing rapport, gathering relevant history regarding symptomatic complaints, and providing an initial assessment of the patient's degree of awareness of their problems. The complaints of the malingerer of massive cognitive impairment following whiplash without head trauma (with detailed examples given of past memory and cognitive failures) are just as important as the denial of deficit made by the patient with suspected Alzheimer-type dementia, when these deficits are only too apparent to the examiner and to the patient's spouse. Careful interviewing and observation can yield clinical data on language function, spatial abilities, motor function, attention, memory, intellectual function, and emotional status. These observations, symptomatic complaints, and clinical history provide the initial hypotheses regarding a patient's neurobehavioral status. These hypotheses can then be tested by formal psychometric procedures, sampling the seven neurobehavioral domains. Observations of normal language function during the clinical interview, followed by normal performance on sensitive measures of word-finding ability, with Verbal IQ within the range of premorbid estimation, would preclude the need for more detailed language evaluation. In the same patient, decreased Performance IQ in the context of normal language, preserved Verbal IQ, and normal verbal memory would indicate the need for assessment of more basic spatial perceptual and spatial constructional skills, as well as manual motor and manual tactile assessment, for evaluation of a potential focal non-dominant hemisphere problem.

Walsh's (1995) caveats regarding extent of supervised, didactic, and experiential knowledge in neuropsychology are critical for effective use of the hypothesis-testing approach to clinical neuropsychological assessment. Aphasia following stroke evolves over time; post-traumatic amnesia resolves over time; cognitive functions sensitive to Alzheimer's-type dementia follow a pattern of differential decline over time. Evaluation of the language-disordered patient following "dominant"-hemisphere cerebrovascular accident (CVA) requires particular skill and experience, because aphasics can fail so-called "non-verbal" tasks (Benton, Sivan, et al., 1994; Hamsher, 1991), including WAIS-R Performance IQ subtests (Larrabee, 1986), in spite of an intact "nondominant" hemisphere. Patients suffering right hemisphere cerebrovascular accident can display impaired verbal-memory test performance due to generalized attentional problems during the subacute stages of recovery, despite having an "intact" left hemisphere (Trahan, Larrabee, Quintana, Goethe, & Willingham, 1989). Additionally, patients with history of nondominant CVA may present with bilateral impairment on stereognostic tasks such as the Benton-Iowa Tactile Form Perception, despite a perfectly normal left parietal lobe (Semmes, 1965). Persons with premorbid history of learning disability can appear to have persistent focal left-hemisphere cognitive problems, including poor verbal learning and lower Verbal relative to Performance IQ, following the typical recovery period for minor closed head injury. An elderly patient with a focal vascular lesion in the area of the angular gyrus in the "dominant" (left) hemisphere can appear to have Alzheimer-type dementia (Cummings & Benson, 1992).

SPECIALIZED NEUROPSYCHOLOGICAL ASSESSMENT PROCEDURES

The ability-focused, flexible-adjustive examination can be considered in a more or less hierarchical fashion ranging from assessment of basic skills (language ability) to more complex skills (memory, intellectual and problem-solving ability). This hierarchy presumes a conscious and alert patient.

Language and Related Functions

In a patient with a history significant for aphasia, the neuropsychological evaluation should begin with a comprehensive examination of language

function. This can be conducted by a speech and language pathologist or can be conducted by the neuropsychologist if she or he has particular expertise and training in language assessment. Modern language-assessment batteries such as the Boston Diagnostic Aphasia Examination (Goodglass & Kaplan, 1983), Western Aphasia Battery (Kertesz, 1982), and Multilingual Aphasia Examination (Benton, Hamsher, & Sivan, 1994) typically include measures of word-finding skills (eg., confrontation naming of objects; fluency tasks requiring generation of words beginning with a certain letter; or generation of names in a semantic category such as animal names), repetition (of words, sentences, digit sequences), auditory comprehension (of serial commands; appropriate yes-no responses to brief questions; matching a picture to a word or phrase spoken by the examiner; following commands to manipulate objects), reading comprehension (of words, sentences or paragraphs), writing (to dictation or from copy), and ratings of the fluency and articulatory features of the patient's spontaneous speech. Modern language evaluation generally follows the classification scheme developed by clinicians at the Boston Veterans Administration (VA), utilizing analysis of the fluent/non-fluent aspects of speech and whether repetition is preserved or impaired to yield seven major types: Broca, Wernicke, Global, Anomic, Conduction, Transcortical Motor, and Transcortical Sensory (Benson, 1993). Additional types/features of aphasic disorders have been related to lesions of subcortical structures in the language dominant hemisphere (Crosson, 1992).

Language evaluation can contribute to differential diagnosis as well as provide information on prognosis following development of aphasia. In Alzheimer-type dementia, language deficits progress from anomic aphasia to transcortical sensory aphasia (impaired comprehension, fluent speech, preserved repetition) to Wernicke's aphasia (impaired comprehension, fluent speech, impaired repetition), to echolalia, palilalia (involuntary repetition during speech), dysarthria, and terminal mutism (Cummings & Benson, 1992).

Comparison of performance on phonemic (letter) versus semantic (category) fluency tasks may show different patterns for different dementing syndromes. Monsch and colleagues (1994) found that patients with Alzheimer's-type dementia were disproportionately impaired on category-fluency relative to letter-fluency tasks, whereas patients with Huntington's disease were equally impaired. Moreover, category-fluency tasks correctly classified more Alzheimer's and elderly control subjects than did letter-fluency tasks. Mickanin, Grossman, Onishi, Auriacombe, and Clark (1994) have also reported greater impairment in semantic relative to letter fluency in Alzheimer's disease.

In neuropsychological examination of the aphasic patient, the impact of the language disturbance on so-called "nonverbal" functions is important. As Benton, Sivan, and collaborators (1994) have reported, up to 44 percent of left-posterior aphasics with comprehension impairment fail the Facial Recognition Test, a "nonverbal" measure requiring discrimination and matching of shaded photographs of non-familiar persons (to be discussed further in the next section). Larrabee (1986) reported significant associations of global language impairment in left-hemisphere damaged patients with WAIS-R Performance IQ (r=.74) and with all the Performance subtests, (range: r=.72 with Object Assembly to r=.44 with Block Design). Hence, an aphasic with demonstrated auditory-comprehension impairment who passes the Facial Recognition Test and performs normally on WAIS-R Block Design might be expected to have a better prognosis than another aphasic with equal degree of comprehension impairment who fails both of these "nonverbal" tasks, presumably because the first patient is able to better monitor the disruptive effects of his disordered language system.

Given the ubiquitous nature of word-finding problems in all aphasic disorders, tests of word-finding are good screens for aphasia. The Multilingual Aphasia Examination (MAE) (Benton, Hamsher, et al., 1994) contains a Visual Naming subtest, requiring confrontation naming of pictures and parts of pictures (eg., elephant, ear, tusk). The MAE also contains a word-fluency task, Controlled Oral Word Association, which requires the subject to produce as many words as possible, with three different letters, 60 seconds per letter. This type of task is also sensitive to non-aphasic, frontal lobe dysfunction (Benton, 1968; Butler, Rorsman, Hill, & Tuma, 1993), is related to level of social skill following very severe closed head injury (Marsh & Knight, 1991), and is predictive of competency to consent to medical procedures in normal elderly patients and patients with Alzheimer's disease (Marson, Cody, Ingram, & Harrell, 1995). The MAE Controlled Oral Word Association Test is part of the three tests of the Iowa Battery for Mental Decline in the elderly (Eslinger, Damasio,

Benton, & VanAllen, 1985). Spreen and Strauss (1998) present an earlier variant (F,A,S) of the MAE Controlled Oral Word Association Test.

Kaplan, Goodglass, and Weintraub (1983) have published a widely used measure of visual- confrontation naming, the Boston Naming Test (BNT). The BNT is highly correlated with the MAE Visual Naming Test (r=.86) and both share significant variance with the WAIS-R Verbal Comprehension factor (Axelrod, Ricker, & Cherry, 1994). Table 12.2 shows a high loading for MAE Visual Naming on a factor defined by WAIS-R Verbal IQ subtests, consistent with the results of Axelrod and colleagues (1994). MAE Controlled Oral Word Association shows a complex pattern of loadings, sharing loadings with the General Verbal Ability factor and the Attention/ Information processing factor. This is not surprising given the timed component of this task.

Certain caveats are important in evaluating performance on word-finding tasks. Tests of visual-confrontation naming can be failed due to modality-specific impairments, which disconnect visual input from preserved language functions (Bauer, 1993; Larrabee, Levin, Huff, Kay, & Guinto, 1985). These non-aphasic patients suffering from visual-verbal disconnection, "optic aphasia," or associative visual agnosia are able to demonstrate normal word-finding skills on tasks not involving visual processing (eg., naming to verbal description). Also, non-aphasic patients with left frontal or bilateral frontal lobe disease or injury had reduced performance on letter fluency (Benton, 1968). Lastly, recent investigation by Jones and Benton (1994) raised questions about the superior sensitivity of word-finding tasks to aphasic disturbance. These authors contrasted the performance of 48 aphasics with 15 normal controls. In this sample, the Token Test (a variant of the procedure originally devised by DeRenzi and Vignolo (1962), requiring the subject to follow commands of increasing complexity to manipulate colored plastic tokens) was the most discriminating, followed by Sentence Repetition, Controlled Oral Word Association, and Visual Naming.

Visuoperceptual and Visuospatial Skills

Measures of visuoperceptual and spatial skills evaluate the patient's ability to visually analyze (e.g., match or discriminate) stimuli, make judgments about the spatial aspects of stimuli, and graphically or constructionally reproduce stimuli. Visuoperceptual or pattern analysis can be dissociated from spatial processing. The former undergoes end-stage processing in the inferior temporal lobe ("what" an object is) whereas the latter undergoes end-stage processing in the posterior parietal cortex ("where" an object is located in space) (Capruso, Hamsher & Benton, 1995; Mishkin, Ungerleider, & Macko, 1983). Additionally, perceptual and constructional skills can be dissociated. Perceptual and spatial tasks may involve no motor response, such as the Benton-Iowa Visual Form Discrimination, Facial Recognition or Judgement of Line Orientation tasks (Benton, Sivan, et al., 1994), or they may involve constructional skills such as two- or three-dimensional puzzle or object assembly (WAIS-R Block Design and Object Assembly subtests, Benton-Iowa Three Dimensional Block Construction; Benton, Sivan, et al., 1994), drawing from copy (Rey-Osterrieth Complex Figure copy administration; Lezak, 1995; Meyers & Meyers, 1995) or line bisection and line cancellation.

As Kane (1991) observed, patients or their families do not often spontaneously complain of impaired spatial skills. The exception, of course, is the patient with profound neglect, resulting in inattention, usually to the left hemi-space. Although patients with neglect do not frequently complain of this problem due to their anosognosia (denial or minimization of deficit), the neglect is readily apparent to family members and professional staff.

Perceptual-spatial tasks can vary from assessing the angular orientation between pairs of lines such as on the Benton-Iowa Judgment of Line Orientation task (Benton, Sivan, et al., 1994), to the complex problem-solving requirements of WAIS-R Block Design. Impaired spatial problem-solving is one of the impairments seen in the earlier stages of Alzheimer's-type dementia (Cummings & Benson, 1992; Ska, Poissant, & Joanette, 1990).

The Benton-Iowa Facial Recognition Test evaluates the patient's ability to discriminate and match shaded black and white photographs of unfamiliar persons (Benton, Sivan, et al., 1994). In non-aphasic patients, only those with disease of the right hemisphere show an excessively high number of impaired performances. Moreover, among patients with right hemisphere disease, there is a high failure rate for those with posterior lesions. Failure is independent of visual field impairment. In patients with left hemisphere disease, only those with impaired auditory comprehension had a high

rate of failure on the Facial Recognition Test. Hermann, Seidenberg, Wyler, and Haltiner (1993) found significant postoperative decline in Facial Recognition performances for both left and right temporal lobectomy patients, whereas their subjects improved on the Judgment-of-Line-Orientation performance. The authors explained this dissociation on the basis of Mishkin's two-component theory of visual processing contrasting "where an object is located" versus "what an object is" (Mishkin, et al., 1983). Benton, Sivan, and coworkers (1994) review a number of other investigations which have employed the Facial Recognition Test for analysis of perceptual abilities of a variety of patients. Table 12.2 demonstrates an association of Facial Recognition performance with the visual/non-verbal problem-solving factor.

The Visual Form Discrimination Test requires matching-to-sample of complex geometric patterns (Benton, Sivan, et al., 1994). This task requires both spatial perceptual skills as well as sustained attention. Patients who are impulsive will perform poorly on this task because the discriminations required are often subtle. Benton, Sivan, and colleagues (1994) report a high failure rate in both left- and right-hemisphere-lesioned patients, attesting to the multifactorial nature of this task.

The Judgment of Line Orientation Test requires the patient to visually judge the angle between two lines, which is compared to a multiple-choice display of 11 lines varying in their degree of angular orientation (Benton, Sivan, et al., 1994). This test is particularly sensitive to focal disease of the right posterior hemisphere. Moreover, the test is typically passed by patients who have left hemisphere disease, even by those who have auditory-comprehension impairment. Hence, the Judgment of Line Orientation Test can provide useful information on the differential diagnosis of an aphasic syndrome secondary to unilateral stroke, from patterns of impaired language and spatial/perceptual skills resulting from bilateral or diffuse brain disease due to multi-infarct or Alzheimer's-type dementia. In the evaluation of dementia, the Judgment of Line Orientation Test appears to be more sensitive than Facial Recognition, although performance dissociations can occur (Eslinger & Benton, 1983). Although Judgment of Line Orientation does not show any sizable loadings in Table 12.2, a prior factor analysis of neuropsychological tests in a normal elderly sample showed a high loading on a factor defined by WAIS-R Block Design and a measure of consistency of retrieval for spatial memory (Larrabee & Levin, 1984).

The Rey-Osterrieth Complex Figure Test can be utilized to evaluate constructional skills, organizational skills sensitive to frontal cognitive functions and memory (Lezak, 1995). Recently, Meyers and Meyers (1995) have provided extensive normative data for the Complex Figure Test.

Visuospatial or visuoperceptual neglect is frequently discussed in reviews of spatial and perceptual assessment (cf. Lezak, 1995) although it is more appropriately considered as a disorder of attention (Heilman, Watson, & Valenstein, 1993). Neglect or inattention to one hemi-space can be assessed via analysis of a patient's drawings, which may appear on one-half (usually the right-hand side) of the paper, by line-cancellation (patients are presented with a page covered by lines which they must cross out), or by line bisection (they must bisect horizontal lines of differing width) (Lezak, 1995; Heilman, Watson, et al., 1993). Neglect is most often the consequence of a right hemisphere lesion and is not due to primary sensory impairment (i.e., it is a cognitive deficit, not a sensory deficit secondary to hemianopsia). It is not uncommon for patients suffering neglect to be unaware of this problem (Heilman, Watson, et al., 1993).

Other less commonly employed measures of perception have also been utilized in evaluating neuropsychological functions in specialized populations. These include measures of stereopsis and measures of color discrimination. Hamsher (1991) explains stereopsis as the ability to ascertain that two objects lie at different distances from the observer, based on the fact that each eye receives slightly different retinal images of these objects. Global stereopsis is the ability to perceive, binocularly, in a stereoscope (which presents separate images to each eye), a form in space that cannot be seen by either eye, individually. Hamsher notes that performance on this type of task can be impaired for persons with right hemisphere lesions, but stereopsis is unimpaired in left-hemisphere-lesioned patients, including those with evidence of auditory comprehension impairment. Hence, performance on global stereopsis and on the Judgment of Line Orientation task may be useful in differentiating bilateral from unilateral dominant hemisphere dysfunction, particularly when dominant hemisphere dysfunction is accompanied by auditory comprehension impairment.

Color-discrimination and color-association tasks have also been evaluated in specialized neuropsychological populations. Alexia without agraphia (the unique presentation of a patient who can write but not read what they have written), resulting from posterior cerebral artery infarction of the left occipital lobe and splenium of the corpus callosum, is frequently associated with the inability to name colors, although color matching may be preserved (Benson, 1979). This is primarily a linguistic deficit (anomia) in contrast to the perceptual impairment reflected by color-matching deficits in persons with posterior right hemisphere disease (Hamsher, 1991).

Braun, Daigneault, and Gilbert (1989), reported sensitivity of color discrimination to solvent neurotoxicity. In an investigation of print- shop workers exposed to toxic solvents, these authors found that performance on the Lanthony D-15 desaturated panel test of chromatopsia significantly discriminated solvent-exposed workers from controls, and reflected a dose effect, with a significant association of impairment with greater solvent exposure. By contrast, performance on 20 neuropsychological tests, including the Wisconsin Card Sort, Rey Auditory Verbal Learning, Trail-making Test and Grooved Pegboard Test, did not discriminate the solvent-exposed workers from nonexposed controls.

Sensorimotor Function

The examination of sensorimotor functions has a long tradition in neurology and neuropsychology. The evaluation of motor and tactile functions of the hands is of particular importance given the known contralateral representation of motor and tactile areas in the cerebral hemispheres.

Hom and Reitan (1982) investigated the effects of left and right hemisphere lesions due to either head trauma, cerebrovascular event, or tumor, on sensorimotor measures from the HRNB (Finger-tapping; Grip Strength; Tactual Performance Test; Suppressions in tactile, auditory and visual modalities; Finger Agnosia; Finger Tip Number Writing; and Tactile Form Recognition). All three etiologies produced greater impairment for the hand contralateral to the lesion, with greatest effects for cerebrovascular, less for tumor, with trauma producing the least-pronounced effects. Right hemisphere lesions produced greater contra and ipsilateral impairment overall, than did left hemisphere lesions.

Sensorimotor examination can encompass motor and tactile ability, basic visual processes, olfaction, and audition. Schwartz and collaborators (1990) reported a dose-dependent decrement in olfactory discrimination in nonsmoking paint-manufacturing workers. Varney (1988) reported poor prognosis for patients developing posttraumatic anosmia due to personality changes secondary to damage to the orbital frontal cortex.

The Sensory Perceptual Examination, frequently performed as part of the Halstead-Reitan Battery, includes assessment of finger-tip number writing, finger localization, and single versus double simultaneous stimulation in the visual, tactile, and auditory modalities (Jarvis & Barth, 1994). As noted in the preceding section, unilateral (usually left-sided) extinction to double simultaneous stimulation, especially across several modalities, can be seen with focal hemispheric lesions contralateral to the neglected hemispace (Heilman, Watson, & Valenstein, 1993). As with any neuropsychological test, adequate normative data are important. Thompson, Heaton, Matthews, and Grant (1987) reported significant age, gender, and education effects for several HRNB sensorimotor tasks (e.g., Tapping, Tactual performance, Grip Strength, Grooved Pegboard) that varied in magnitude and direction, depending on the task, and upon the subjects' preferred hand. Considerable inter-manual variability was found in these normal subjects, suggesting caution in the interpretation of a lateralized lesion based on differences in right- and left-hand performance.

Because of the effects of motivation on manual motor tasks (Binder & Willis, 1991; Heaton, Smith, Lehman, & Vogt, 1978), assessment across a range of tasks is recommended for evaluation of consistency of performance. Heaton et al. (1991) provide age, gender, and education-adjusted normative data for the Halstead-Reitan Finger Tapping Test (requiring repetitive tapping of a telegraph-like key for 10 seconds), hand dynamometer, and the Grooved Pegboard Test (requiring the subject to rapidly place small grooved pegs in sequential rows). The Purdue Pegboard Test is also widely used. This requires the subject to place small metal pegs in a columnar array, by 30-second trials, first with the dominant, then nondominant hand, followed by a bimanual trial. Normative data are provided for younger adults by Yeudall, Fromm, Redden, and Stefanyk (1986) and for mid-

dle-aged and older adults by Agnew, Bolla-Wilson, Kawas, and Bleecker (1988).

Tactile functions can be evaluated with the Benton-Iowa Tactile Form Perception Test, which requires the subject to palpate 10 different sandpaper geometric forms, one at a time, with vision obscured, and match these forms to their visual referents on a multiple-choice card containing 12 different stimuli. Each hand is examined individually, using a different set of forms. This procedure is sensitive to unilateral as well as bilateral brain disease (Benton, Sivan, et al., 1994). Unilateral impairment is associated with a lesion in the contralateral hemisphere. Bilateral impairment can occur with bilateral lesions or with right (nondominant) hemisphere lesions (cf. Semmes, 1965).

Several procedures exist for evaluating finger localization skills. As noted, earlier, there is a finger localization task on the Halstead-Reitan battery. Benton, Sivan, and colleagues (1994) have also published a test of Finger Localization. This utilizes a model of the left and right hands which is placed on top of a screen, in free vision of the patient, with each finger identified by a number. By utilizing the model, demands on language are minimized. The dominant hand is examined first, beginning with touching individual fingers in free vision. This task is repeated for the nondominant hand. Then, alternating from dominant to nondominant, the examiner touches individual fingers with the hand hidden, followed by double simultaneous stimulation of two fingers with the hand hidden. Benton, Sivan, et al. (1994) provide data demonstrating the sensitivity of this task to bilateral and unilateral cerebral disease. Bilateral impairment in finger localization can be seen with either bilateral disease, or unilateral left cerebral disease. Finger localization skills have also been related to reading achievement in children (Satz, Taylor, Friel, & Fletcher, 1978).

Table12.2 presents data demonstrating a dissociation in factor loadings. The Purdue Pegboard and Grooved Pegboard show more complex loadings with the spatial/perceptual factor. The Finger Tapping and hand dynamometer load together on a relatively pure motor factor. Of particular interest, the Purdue Pegboard and Grooved Pegboard show an association of left-hand skill with the nonverbal/problem solving factor and also with dominant and nondominant Tactile Form Recognition scores.

ATTENTION, INFORMATION PROCESSING, AND IMMEDIATE MEMORY

Measures of attention, information processing, and immediate memory are grouped together due to factor-analytic evidence that these tasks are assessing a common underlying construct (Larrabee & Curtiss, 1995; Larrabee, Kane, & Schuck, 1983; Larrabee, & Curtiss, 1992; see Table 12.2). Although these measures are discussed separately from memory in this chapter, measures of attention are frequently included in memory batteries such as the WMS-R (Wechsler, 1987) and Larrabee and Crook (1995) have included this domain as part of a five-component model for assessment of memory including: (1) orientation, (2) attention/concentration information-processing and immediate memory, (3) verbal learning and memory, (4) visual learning and memory, and (5) recent and remote memory.

Tables 12.1 and 12.2, which include measures of attention as well as sensorimotor function, memory, and verbal and visual intellectual ability, also suggest a high degree of shared variance in attentional tasks. Other factor analyses which have analyzed attentional tasks in the absence of measures of memory, verbal, and visual intelligence, have yielded multiple dimensions of performance including visuo-motor scanning, sustained selective attention and visual/auditory spanning (Shum, McFarland, & Bain, 1990), visuo-motor scanning and visual/auditory spanning (Schmidt, Trueblood, Merwin, & Durham, 1994) and focus/execute, shift, sustain, and encode (Mirsky, Anthony, Duncan, Ahearn, & Kellam, 1991).

Various definitions of attention exist. Lezak (1995) has observed that a universally accepted definition of attention has yet to appear. She defines attention as several different capacities or processes by which the subject becomes receptive to stimuli and begins to process incoming or attended-to excitations. Attention is ascribed a certain limited capacity, and is related to sustained effort and shifting of focus. Cohen (1993) also highlights the multidimensional nature of attention. Attention is described as facilitating cognitive and behavioral performance by reducing or expanding the amount of information which is to receive further processing by the brain, and assessing the saliency of information. Cohen relates these processes, metaphorically, to the aperture and lens system of a camera. Cohen also describes

other features of attention, including evaluating the spatial and temporal characteristics of a particular context, analogous to a "spotlight." He synthesizes various theoretical conceptualizations of attention into 4 components: sensory selection, response selection, attentional capacity, and sustained performance. Mirsky and coworkers (1991) take a more psychometric approach, based on the factors identified in their factor analysis of purported measures of attention.

A variety of neuropsychological tests have been utilized as measures of the various aspects of attention and information processing, including the Arithmetic and Digit Symbol subtests of the WAIS-R, the Digit Span subtest of the WAIS-R and WMS-R, and Mental Control and Visual Memory Span measures of the WMS-R (Larrabee et al., 1983; Larrabee & Curtiss, 1992, 1995; Mirsky et al., 1991; Schmidt et al., 1994; Shum et al., 1990). Other procedures related to attentional processes include the Seashore Rhythm Test, Speech Sounds Perception Test, and Trailmaking Test from the Halstead-Reitan (Leonberger et al., 1992; Schmidt et al., 1994; see Table 12.1). The Stroop Test (Golden, 1978; Trenerry, Crosson, DeBoe, & Leber, 1989), measures of letter cancellation, serial subtraction, and the Knox Cube (Shum et al., 1990) have been utilized as measures of attention. Mirsky and colleagues (1991) include scores from the Wisconsin Card Sorting Test as an assessment of the "shift" aspect of attention, and the Continuous Performance Test as a measure of sustained attention.

Attentional measures from the WAIS-R and WMS-R will not be reviewed in detail. The respective test manuals provide adequate normative data for these measures, which can be extended into the upper age ranges with the Mayo Older Adult Normative Studies (Ivnik et al., 1992a, 1992b). These procedures can also provide important information on the motivation of the subject being evaluated, particularly if there is disproportionate impairment of attention relative to other memory and intellectual functions (Mittenberg, Azrin, Millsaps, & Heilbronner, 1993; Mittenberg, Theroux-Fichera, Zielinski, & Heilbronner, 1995). Forward digit span has a weaker association with age than reversed digit span (Craik, 1984; Wechsler, 1987). There is some evidence that reversed digit span may be related to visual scanning and visuospatial skill (Costa, 1975; Larrabee & Kane, 1986), although, this has not been demonstrated consistently (Wechsler, 1987). Leonberger and collaborators (1992; see Table 12.1) found a closer association of WMS-R Visual Memory span with a spatial cognitive factor than with an attentional factor. This raises questions regarding the interpretation of this particular test as a measure of attention on the WMS-R.

One of the most widely used measures of attentional tracking and sequencing is the Trailmaking Test, in particular, Trailmaking B, which requires the subject to perform a divided attention- task and alternately connect numbers and letters in increasing order of value (e.g., 1 to A to 2 to B, etc.). The factor analyses by Leonberger and colleagues (1992) (see Table 12.1) suggest that performance on this test is determined by both spatial cognitive as well as attention and psychomotor speed abilities. Normative data corrected for age, education, and gender, are provided by Heaton and coworkers (1991). Stuss, Stethem, and Poirer (1987) also provide normative data for Trailmaking, which was co-normed with the Paced Auditory Serial Addition Test (PASAT) (Gronwall, 1977) and Consonant Trigrams procedure (Brown, 1958; Peterson & Peterson, 1959). In a recent meta-analytic review of the sensitivity of neuropsychological tests to brain damage, Trailmaking B and WAIS-R Digit Symbol were among the most sensitive measures (Chouinard & Braun, 1993).

The PASAT was originally developed to investigate information-processing rate after closed head trauma (Gronwall, 1977). In this task, the subject has to perform rapid serial addition across four blocks of numbers, with the time between numbers decreasing from 2.4 seconds to 2.0, 1.6 and 1.2 seconds. Two versions of the test exist. The original version developed by Gronwall uses 61 numbers (Gronwall, 1977). Normative data for this version are provided by Stuss and colleagues (1987). A revised version has been developed utilizing computer-synthesized speech and 50 numbers per trial block (Levin, Mattis, et al., 1987). Normative data for this version of the PASAT are provided by Brittain, LaMarche, Reeder, Roth, and Boll (1991) and by Roman, Edwall, Buchanan, and Patterson (1991).

The PASAT is sensitive to the information- processing deficits seen in the early stages of recovery from mild closed head injury (Gronwall, 1977; Levin, Mattis, et al., 1987). This sensitivity is related to the speeded nature of the task as well as the demands the task places on working memory (i.e., in the sequence, "2, 8, 4, and 6," the subject must provide the response "10" to the numbers "2"

and "8," then following hearing the number "4," add it to the preceding number heard, "8," rather than the preceding sum, producing the response "12," followed by hearing "6," then adding it to the preceding number, "4," etc.).

Gronwall (1977) also recommended utilizing the pattern of responding to evaluate level of motivation of a particular patient. Gronwall described the case of a 14-year-old girl who, following moderate concussion, was making satisfactory recovery. On the 28th day post-trauma, the girl's PASAT results were completely inconsistent, with as many correct at the fast trial as at the slow trial. Gronwall noted that the girl had been a mediocre student and was reluctant to return to school full-time. The week following the girl's agreement that she had no choice but to return to school, her PASAT scores were consistent and normal.

Variations of the Consonant Trigrams Procedure have been utilized to evaluate sensitivity of short-term memory to interference in research on alcoholic Korsakoff syndrome, schizophrenics who had undergone frontal leukotomy, and patients who had sustained mild or severe closed head trauma (Butters & Cermak, 1980; Stuss, Kaplan, Benson, Weir, Chirilli, & Sarazin, 1982; Stuss, Stethem, Hugenholtz, & Richard, 1989). In this procedure, subjects are provided with three consonants, for example, C, F, L, then must engage in an interfering activity, counting backwards by threes for either 3, 9 or 18 seconds, following which they are asked to provide the letters. In their original research (Stuss et al., 1982), Consonant Trigrams was the only test out of several measures of learning and memory that was sensitive to residual effects of orbito frontal leucotomy.

In subsequent research utilizing longer delay periods of 9, 18 and 36 seconds, Stuss and collaborators (1989) found that the Trailmaking Test, PASAT, and Consonant Trigrams all discriminated control subjects from severe closed head trauma patients, whereas Consonant Trigrams alone discriminated patients with mild closed head trauma from controls. Normative data for the 9-, 18-, and 36-second version of Consonant Trigrams are provided by Stuss and colleagues (1987).

SPECIALIZED ASSESSMENT OF LEARNING AND MEMORY

There have been two basic approaches to the evaluation of learning and memory. The one most familiar to general clinicians is based on an omnibus battery, such as the Wechsler Memory Scale-Revised (Wechsler, 1987) or the Memory Assessment Scales (Williams, 1991). These omnibus batteries typically include a variety of measures of attention, verbal and visual learning, and memory.

The second approach is based on utilizing a selection of specialized, individually developed measures of various components of memory (Erickson & Scott, 1977; Larrabee & Crook, 1995). For a comprehensive assessment, Larrabee and Crook (1995) have recommended assessing five components: (1) Orientation; (2) Attention/Concentration, Information Processing, and Immediate Memory; (3) Verbal Learning and Memory; (4) Visual Learning and Memory; and (5) Recent and Remote Memory function. Larrabee and Crook also recommend analysis of forgetting scores, which can be particularly sensitive to amnestic and dementing conditions (Butters et al., 1988; Ivnik, Smith, Malec, Kokmen, & Tangalos, 1994; Larrabee, Youngjohn, Sudilovsky, & Crook, 1993; Martin, Loring, Meador, & Lee, 1988).

Although appropriate age-based normative data are important for any neuropsychological-test procedure, age-based norms are particularly critical in assessment of learning and memory. Effects of age on level of performance are much less pronounced for immediate memory-span measures such as the WAIS-R Digit Span or for measures of recent and remote memory such as the Presidents Test (Hamsher, 1982) than they are for supraspan (i.e., beyond immediate memory span) learning of verbal and visual materials (Craik, 1984; Davis & Bernstein, 1992). Aging effects are also much less pronounced for forgetting rates (amount lost on delay as a function of material originally acquired) in normal subjects (Trahan & Larrabee, 1992, 1993). Normative data on forgetting rates have been published for the Rey Auditory Verbal Learning Test (RAVLT) (Geffen, Moar, O'Hanlon, Clark, & Geffen, 1990; Ivnik et al., 1992c); and the original (form 1) WMS Visual Reproduction Test with delayed recall (Trahan, 1992). Table A3 of the California Verbal Learning Test manual also provides normative data relative to analysis of forgetting (Delis, Kramer, Kaplan, & Ober, 1987).

Assessment of Orientation

Orientation, typically evaluated in four spheres: time, place, person and situation, is a common

component of the mental status examination (Strub & Black, 1985). Disorientation to time frequently suggests the presence of some type of abnormal condition such as amnesia, dementia, or confusion (Benton, Sivan, et al., 1994), for orientation to time and place are actually measures of recent memory because they evaluate a patient's ability to learn and remember continuing changes in these spheres (Strub & Black, 1985).

Perhaps the best standardized measure of orientation to time is the Temporal Orientation Test of the Benton-Iowa group (Benton, Sivan, et al., 1994). This procedure requires the subject to identify the month, date, day of the week, year and to estimate the time of day. Specific error points are associated with varying magnitudes of error (e.g., being incorrect on the month is weighted more heavily than misidentifying the day of the week). Normative data are available on over 400 subjects, and there is limited association of performance with age (Benton, Sivan, et al., 1994).

Failure on the Temporal Orientation Test is more common with bilateral hemispheric disease (Benton, Sivan, et al., 1994). A screening battery which included the Temporal Orientation Test, the Benton Visual Retention Test (Sivan, 1992), and Controlled Oral Word Association Test (Benton, Sivan, et al., 1994) correctly discriminated 89 percent of normal and demented elderly (Eslinger et al., 1985).

The questions from the Temporal Orientation Test are also a major component of the Galveston Orientation and Amnesia Test (GOAT) (Levin, O'Donnell, & Grossman, 1979). The GOAT, developed to evaluate presence and duration of posttraumatic amnesia (e.g, the period of confusion and disorientation following significant closed head trauma), contains a brief series of questions assessing orientation to time, place, and person, as well as questions related to retrograde (recall of events prior to trauma) and anterograde (recall of events subsequent to trauma) amnesia. High, Levin, and Gary (1990) analyzed the pattern of recovery of components of orientation on the GOAT following head trauma of varying severity and found that the most common pattern, in 70 percent of patients studied, was return of orientation to person, followed by orientation to place, with orientation to time the last component to recover.

Assessment of Verbal Learning and Memory

A variety of methods exist for evaluating verbal learning and memory, including immediate and delayed recall of brief passages of prose (Logical Memory subtest of the WMS-R), digit supraspan learning (rote memorization, in sequence, of an eight- or nine-digit number, exceeding immediate memory span, such as the Benton-Iowa Serial Digit Learning Test; Benton, Sivan, et al., 1994), forced-choice recognition memory for words previously seen (Recognition Memory Test) (Warrington, 1984), and supraspan word-list learning (multiple-trial list-learning tasks, such as the Rey Auditory Verbal Learning Test (RAVLT), California Verbal Learning Test (CVLT), and Verbal Selective Reminding Test (VSRT) (cf. Lezak, 1995; Delis et al., 1987; Buschke, 1973). Paired Associate Learning is another modality of verbal memory testing in which the patient on the presentation trial hears a list of pairs of words, followed by a test trial in which the first word of the pair is presented, to which the patient must provide the second word (Paired Associate Learning of the WMS-R: Wechsler, 1987; Expanded Paired Associate Test [EPAT]: Trahan et al., 1989).

Three of the more widely used supraspan verbal list-learning procedures in clinical and research applications of neuropsychology are the RAVLT, CVLT, and VSRT. All three require the subject to learn a supraspan list of words over several trials, with testing of delayed recall and testing of recognition.

The RAVLT requires the subject to learn a list of 15 unrelated words over five trials, followed by a second list to serve as interference and subsequent short- and long-delay recall of the original list (Lezak, 1995; Spreen & Strauss, 1991). Lezak (1995) has also provided a 50-word list (containing the acquisition list, interference list, and 20 more words) for recognition testing following the delayed-recall trial (testing delays vary, 20 to 30 minutes after acquisition depending on the particular laboratory: Lezak, 1995; Spreen & Strauss, 1998). Analysis of patterns of performance can yield information on serial-position effect, proactive interference, retroactive interference, and forgetting over time (Larrabee & Crook, 1995; Lezak, 1995).

Performance on the RAVLT is affected by a variety of conditions including temporal lobectomy (Ivnik et al., 1993), hydrocephalus (Ogden,

1986), vertebrobasilar insufficiency (Ponsford, Donnan, & Walsh, 1980) and early Alzheimer-type dementia (Mitrushina, Satz, & Van Gorp, 1989). Powell, Cripe, and Dodrill (1991) found that the RAVLT, particularly trial 5, was more sensitive to discriminating a group of normal subjects from a mixed neurologic group than any other single test on the Halstead-Reitan or Dodrill (Dodrill, 1978) batteries. In a factor analysis of the RAVLT and other neuropsychological measures, Ryan, Rosenberg, and Mittenberg (1984) identified a factor on which the RAVLT and WMS verbal memory scores loaded. Normative data are provided by Ivnik and coworkers (1992c); Geffen and colleagues (1990); and Wiens, McMinn, and Crossen (1988). Crawford, Stewart, and Moore (1989) have developed alternate, parallel forms for the original List A and List B.

The CVLT is, on the surface, similar in general format to the RAVLT, with a five-trial supraspan learning task, followed by an interference list and short- and long-delay free recall; however, the CVLT was designed by Delis and colleagues to evaluate the process of verbal learning using an "everyday" task of learning and remembering a shopping list (Delis et al., 1987). The subject is presented with a "Monday" list of 16 items (four each, in the categories of tools, clothing, fruits, and spices/herbs), over five trials, followed by a second "Tuesday" list to serve as interference, short-delay and long-delay free recall and category-cued recall, followed by delayed multiple-choice recognition. By design, the CVLT allows for evaluation of multiple dimensions of performance including semantic clustering versus serial-learning strategies, vulnerability to proactive and retroactive interference, retention of information over time, and free versus cued recall versus recognition memory. Indeed, a factor analysis of the CVLT yielded several factors including general verbal learning, response discrimination, proactive effect, and serial position effect (Delis, Freeland, Kramer, & Kaplan, 1988). This factor structure has been replicated by Wiens, Tindall, and Crossen (1994), who have also provided additional normative data. Interestingly, these normative data yielded lower values than those published in the CVLT manual (Delis et al., 1987). The CVLT test manual contains normative data on a variety of clinical populations including Alzheimer's disease (AD), Korsakoff amnestic syndrome, multiple sclerosis and head trauma (Delis et al., 1987). Research on the CVLT has shown discrimination of patients with severe closed head trauma from control subjects in both level and pattern of performance (Crosson, Novack, Trenerry, & Craig, 1988; Millis & Ricker, 1994). A discriminant function analysis classified over 76 percent of cases of Huntington's disease, AD and Parkinson's disease (Kramer, Levin, Brandt, & Delis, 1989). Delis and collaborators (1991) have developed an alternate form of the CVLT.

The Verbal Selective Reminding Test (VSRT) was originally developed by Buschke (1973), in an attempt to separate the components of storage and retrieval inherent in verbal-list learning tasks. Unlike the RAVLT and CVLT, the only time the subject hears the examiner present all of the VSRT words is the first trial; thereafter, the examiner presents only those words which were omitted on the immediately preceding trial, yet the subject is still expected to provide all of the words (those reminded and those not reminded) on the list. Several different word lists exist for various versions of the VSRT (Spreen & Strauss, 1998). One of the most widely used versions is the 12- unrelated-word-12-trial version developed by Levin and colleagues (Hannay & Levin, 1985; Larrabee, Trahan, Curtiss, & Levin, 1988; Levin, Benton, & Grossman, 1982). Normative data are provided by Larrabee and coworkers (1988), which are reprinted in Spreen and Strauss, (1998). Additional normative data are provided by Ruff, Light, and Quayhagen (1989).

As discussed in Larrabee and colleagues (1988), the scoring criteria for the VSRT assume that once a word has been recalled at least once, without reminding, it is in long-term storage (LTS). If it is then recalled to criterion (correct recall of the entire list for three consecutive trials or to the final trial of the test), the word is considered to be in consistent long-term retrieval (CLTR). There is some debate about the validity of these assumptions (Loring & Papanicolaou, 1987). Larrabee and collaborators (1988) found that the various VSRT scores (CLTR, LTS, Short-Term Storage, Short-Term Recall, Random Long-Term Retrieval) defined only one factor when factor analyzed in the absence of any other test scores. Larrabee and Levin (1986) found separate verbal learning and retrieval factors when a reduced set of VSRT scores was factored with other memory-test measures. More recently, Beatty and coworkers (1996) demonstrated predictive validity for the various retrieval and storage indices in a sample of patients with multiple sclerosis. Words in CLTR were more

consistently recalled at delay than were words in Random Long-Term Retrieval or Short-term Storage.

The Levin VSRT exists in four forms. In normal adult subjects, forms 2, 3, and 4 are equivalent and approximately 10 percent easier than form 1 (Hannay & Levin, 1985). Because the normative data are based on form 1, this led Larrabee and colleagues (1988) to recommend reducing the raw score on Forms 2, 3, or 4 by 10 percent prior to utilizing the normative tables; however, Westerveld, Sass, Sass, and Henry (1994) found no form difference in patients with seizure disorders.

The VSRT has been widely used in research on closed head trauma (Levin et al., 1982). The procedure is sensitive to the effects of severe closed head injury in adults (Levin, Grossman, Rose, & Teasdale, 1979), children and adolescents (Levin et al., 1988). The VSRT is also sensitive to the memory decline in early-stage Alzheimer-type dementia (Larrabee, Largen, & Levin, 1985; Masur, Fuld, Blau, Crystal, & Aronson, 1990) and Sass and colleagues (1990) have correlated VSRT performance with hippocampal cell counts. Factor analyses of the VSRT show it loads on a general memory factor independent of intellectual and attentional processes (Larrabee & Curtiss, 1995; Larrabee, Trahan, & Curtiss, 1992; also see Table 12.2).

Visual Memory Performance

A variety of methodologies have been developed for evaluation of visual learning and memory including forced-choice recognition memory for facial photographs (Warrington, 1984), yes/no recognition memory for recurring familiar pictures (Hannay, Levin, & Grossman, 1979) or geometric forms (Kimura, 1963; Trahan & Larrabee, 1988), and drawing previously seen designs from memory (Meyers & Meyers, 1995; Sivan, 1992; Trahan, Quintana, Willingham, & Goethe, 1988; Wechsler, 1945, 1987). Other methodologies have included recall of object placement in a spatial array for abstract symbols (Malec, Ivnik, & Hinkeldey, 1991) or marbles (Levin & Larrabee, 1983) or learning a supraspan spatial sequence (Milner, 1971, describing a task developed by P. Corsi).

Factor analyses of purported measures of visual memory, including other cognitive tasks of attention, verbal and visuospatial intelligence, and verbal memory, frequently show high (sometimes the highest) loading on a factor assessing spatial- intellectual skills, with lower loadings on a memory factor (Larrabee, Kane, & Schuck, 1983; Leonberger et al., 1992, see Table 12.1). This poses a problem psychometrically, for when a purported measure of visual memory shows a stronger association with visuospatial intelligence and problem solving than with memory, the test is better described as a spatial problem-solving task. This problem is more pronounced when visual memory is assessed via immediate reproduction from memory (Larrabee & Curtiss, 1995; Larrabee, Kane, et al., 1985; Leonberger, et al., 1992). When delayed reproduction scores are factored, the loading for the purported visual-memory tasks increases on the memory factor (Larrabee, Kane, et al., 1985; Larrabee & Curtiss, 1995; Leonberger et al., 1992). For some visual-memory tasks, the strength of the loading pattern may actually reverse such that when immediate visual reproduction scores are factored, the strongest loading is with spatial intelligence with a secondary loading on memory, whereas when delayed visual reproduction scores are factored, the strongest loading is with memory with a secondary loading on spatial intelligence (Larrabee & Curtiss, 1995; Larrabee, Kane, et al., 1985). The factorial confound of spatial intelligence and problem solving with purported tasks of visual memory may be attenuated by use of visual-recognition memory tests (Larrabee and Curtiss, 1992, and Table12.2; Larrabee and Curtiss, 1995).

The Benton Visual Retention Test (Sivan, 1992) is also widely used as a measure of immediate design reproduction from memory. Advantages include a large normative database for adults and children, three alternate forms, and several published studies supporting sensitivity of the procedure to brain damage (see Spreen & Strauss, 1998, for a review). Factor analysis has demonstrated loadings on attention, memory, and spatial ability (Larrabee, Kane, et al., 1985). The design of the task, with 10 separate geometric patterns (six of which contain three figures per card), precludes administration of a delayed reproduction trial.

The Complex Figure Test (Lezak, 1995; Meyers & Meyers, 1995; Osterrieth, 1944; Rey, 1941; Spreen & Strauss, 1998) requires the subject to copy a spatially complex figure comprised of multiple geometric components. There are 18 scorable components which can be scored 0, .5, 1, or 2, for a total score-range of 0 to 36. Following a copy phase, the patient reproduces the complex figure from memory, with subsequent visuographic

reproduction at anywhere from a 20-to-45-minute delay, depending on the laboratory (Lezak, 1995; Spreen & Strauss, 1998). Various sets of normative data have been archived, through accrual (Lezak, 1995; Spreen & Strauss, 1998). Loring, Martin, Meador, and Lee (1990) found that 30-minute-delayed scores were higher when the copy trial was followed by an immediate reproduction trial than delayed recall without a preceding immediate recall trial. This calls for caution in application of the appropriate delayed reproduction norms.

Recently, Meyers and Meyers (1995) published a comprehensive test manual for the Complex Figure Test. This manual contains normative data on 601 normal subjects ranging from 18 to 89 years of age. Administration involves a copy trial, immediate trial (administered three minutes after the copy trial is completed), and a delayed reproduction trial completed 30 minutes after the copy trial has ended. Meyers and Meyers have also developed a recognition trial which is administered following the 30-minute-delayed trial. Specific scoring criteria are provided for the 18 different units, and the Appendix presents three fully scored examples. The explicit scoring criteria led to a median inter-rater reliability of .94.

Lezak (1995), Spreen and Strauss (1998), and Meyers and Meyers (1995) have reviewed the sensitivity of performance on the Complex Figure Test to a variety of neurologic conditions including closed head trauma, stroke, and dementia. Diamond and Deluca (1996) found that ten patients with amnesia secondary to ruptured anterior-communicating-artery aneurysms demonstrated profound loss of information on delayed reproduction of the Complex Figure, despite copy scores that were within normal limits. Lezak (1995) reviews factor-analytic data showing both a memory as well as a spatial component to Complex Figure performance. To date, there have been no published studies factoring immediate and delayed scores separately, with marker variables for verbal and visual intelligence, attention, and memory. The test manual published by Meyers and Meyers (1995) reports significant correlations of immediate and delayed Complex Figure scores with several WAIS-R PIQ subtests, the BVRT, and the RAVLT. Hence, the Complex Figure Test may be susceptible to the same spatial cognitive confounds as other design-reproduction-from-memory tasks.

Warrington (1984) has developed a forced-choice measure of facial recognition memory. This is paired with a forced-choice word-recognition task. Both tasks require the subject to make a judgement regarding how pleasant/unpleasant a word or face is. Following presentation of 50 different words, the subject is presented with 50 pairs of words and forced to choose which of the pair of words was previously seen. The same format is followed for the face memory test, that is, 50 faces presented, followed by 50 pairs of faces, with the subject required to specify which of the pair of faces was previously seen. Warrington (1984) provides data showing the expected double dissociation in performance, with right hemisphere-lesioned patients performing lower on faces referable to words, with the opposite pattern seen with left hemisphere lesions.

Two other measures of visual memory are the Continuous Recognition Memory Test (CRM) (Hannay et al., 1979) and the Continuous Visual Memory Test (CVMT) (Trahan & Larrabee, 1988). Both require the subject to detect and discriminate recurring from perceptually similar but nonrecurring figures in a yes-no recognition memory format. The CRM utilizes recognizable objects (eg., insects, seashells) whereas the CVMT employs abstract geometric patterns.

The CRM was developed for research on visual-memory deficits following closed head trauma (Hannay et al., 1979). In this original investigation, performance on the CRM differentiated persons with moderate closed head trauma from persons with mild head trauma and from non-neurological medical control patients. Levin and colleagues (1982), found that patients with mass lesions in the left temporal lobe performed defectively on the VSRT, but normally on the CRM. Hannay and Levin (1989) found that CRM performance varied as a function of head-trauma severity in adolescents who had sustained mild, moderate, or severe closed head injury. Trahan, Larrabee, and Levin (1986) reported significant effects of normal aging on CRM performance for 299 persons ages 10 to 89 years.

The CVMT, in addition to an acquisition trial, also includes a 30-minute-delay multiple-choice recognition task, followed by a match-to-sample discrimination task to rule out gross perceptual-spatial deficits (Trahan & Larrabee, 1988). Normative data are presented for 310 adults ages 18 to 91 years, with additional data on failure rates for patients with amnestic disorder, AD, and severe TBI. One hundred percent of the amnestic, 92 percent of the AD subjects, and 68 percent of the trau-

matic-brain-injury (TBI) subjects were impaired on at least two CVMT scores (Trahan & Larrabee, 1988). Patients with right hemisphere CVA performed at a significantly poorer level on the CVMT than did patients with left hemisphere CVA (Trahan, Larrabee, & Quintana, 1990). Trahan, Larrabee, Fritzsche, and Curtiss (1996) have reported on the development of an alternate form of the CVMT.

Larrabee and collaborators (1992), in a factor analysis of CVMT performance in normal subjects, found that the CVMT acquisition score for sensitivity loaded on attentional and intellectual factors. By contrast, the delayed-recognition CVMT score loaded on a visual-memory factor that was independent of the intellectual factors, as well as independent of a verbal memory factor. Larrabee and Curtiss (1995), in a factor analysis of a variety of measures of attention, memory, and intelligence, using a mixed group of neurologic and psychiatric patients, found that both the acquisition and delayed scores of the CVMT and CRM loaded on a general (verbal and visual) memory factor, in separate factor analyses of acquisition and delayed scores (also see Larrabee & Curtiss, 1992, and Table 12.2).

Altogether, the factor analyses conducted by Larrabee and coworkers (1992), Larrabee and Curtiss (1992, 1995), and Leonberger and coworkers (1992) demonstrate two important points. First, visual-recognition memory-testing procedures appear to have less of a spatial confound (with visuospatial problem solving) than drawing-from-memory visual-memory tasks. Second, as noted by Larrabee and Crook (1995), on a factor-analytic basis, the original WMS Visual Reproduction figures, utilizing a modified immediate and delayed reproduction format (Russell, 1975; Trahan et al., 1988) are a better measure of memory in delayed-recall format than are the WMS-R Visual Reproduction designs (also, compare the loadings for the WMS-R Visual Reproduction in Table 12.1 to the loadings for WMS Visual Reproduction in Table 12.2).

Recent and Remote Memory

As already noted, Temporal Orientation can be considered to be a measure of recent memory for material the patient "brings with" them to the examination (Strub & Black, 1985). It is also important to consider more remote aspects of memory. Larrabee and Crook (1995) distinguish between Tulving's (1972) constructs of *episodic* or context dependent memory as opposed to *semantic* memory (memory for facts). Using the example of recall of the identity of the "Enola Gay," Larrabee and Crook note that for a 65-year-old person who recalls the precise context of seeing a newspaper headline concerning the dropping of the atom bomb, this information is in episodic memory. By contrast, for the teen-aged history and trivia buff, this material is more likely in semantic memory.

Larrabee and Crook (1995) highlight the importance of making this distinction, which is exemplified by the normal performance of Korsakoff amnestic patients on the WAIS Information subtest (Butters & Cermak, 1980) contrasted with the marked retrograde amnesia evident on the Albert, Butters, and Levin (1979) Remote Memory Battery assessing memory for famous faces and famous events.

There are far fewer procedures available for evaluation of remote episodic memory. The original Remote Memory Battery of Albert and collaborators (1979) evaluated memory for famous faces and famous events from the 1920s through the 1970s. Using this approach with Korskoff amnestic patients, Albert and colleagues demonstrated a gradient of impairment in remote memory, which followed Ribot's (1881) law of regression, in which memories from the remote past were better preserved than those acquired during the more recent past. Administration of this battery to persons with dementing conditions has not yielded the gradient of impairment found in amnesia; rather, a global impairment is seen in remote memory for patients with Huntington's disease (Albert, Butters, & Brandt, 1981) and for patients with Alzheimer's-type dementia (Wilson, Kaszniak, & Fox, 1981). White (1987) has published a short form of the Remote Memory Battery, and Beatty, Salmon, Butters, Heindel, and Granholm (1988) have utilized a version which has been updated with material from the 1980s in an investigation of retrograde amnesia in Alzheimer's and Huntington's diseases.

Hamsher (1982) has published a brief measure of recent and remote memory: the Presidents Test. This test, derived from common mental-status examinations concerning memory for recent U.S. Presidents, has four parts: (1) Verbal Naming, requiring free recall of the current and five previous U.S. Presidents; (2) Verbal Sequencing, requiring sequencing of six cards imprinted with

the names of the last six presidents (presented in quasi-random order) in the actual order of office; (3) Photo Naming, requiring confrontation naming of photographs of each of the last six presidents (presented in the same quasi-random order as Verbal Sequencing); and (4) Photo Sequencing, requiring sequencing of the photographs in the actual order of office. Verbal Naming and Photo Naming are scored in terms of number correct. Verbal and Photo Sequencing are scored by computing the Spearman rho between the patient's sequence and the actual sequence of office.

The Presidents Test was normed on 250 hospitalized non-neurologic, nonpsychiatric medical patients, with corrections for age and education (Hamsher, 1982). Initial data suggest there is no need to re-norm the procedure each time a new president enters office. Hamsher and Roberts (1985) found that the Verbal Naming Test was the most difficult, whereas Photo Naming was the easiest, and patients with diffuse neurological disease and/or dementia performed the poorest on the various subtests. Roberts, Hamsher, Bayless, and Lee (1990) found that 88 percent of patients with diffuse cerebral disease and control subjects were correctly classified on the basis of their Presidents Test performance. In this same investigation, patients with right hemisphere disease demonstrated a selective impairment in temporal sequencing, whereas patients with left-sided lesions demonstrated selective impairment on the Verbal Naming and Photo Naming subtests. The construct validity of the Presidents Test was supported by the factor analysis of Larrabee and Levin (1986), who found a factor that was defined by self-rated change in remote memory, the Verbal Naming subtest of the Presidents Test and the Levin version (Levin et al., 1985) of Squire and Slater's (1975) Recognition Memory Test for canceled television shows.

In their review, Larrabee and Crook (1995) noted that the advantages of the Presidents Test included good standardization and brief administration time. The major disadvantage was that performance could not be analyzed for the presence of a temporal gradient of impairment.

Assessment of Intellectual and Problem-Solving Skills

Measures of intelligence and problem solving have a long history in psychology and neuropsychology. Tulsky, Zhu, & Prifitera (chapter 5, this volume) provide a comprehensive review of the evaluation of intelligence in adults. Lezak (1995) provides a thorough review of measures of concept formation and reasoning, which includes tests of proverbs, similes, verbal abstraction, and visual-concept formation such as the Proverbs Test of Gorham (1956), various subtests of the WAIS-R, the Halstead Category Test (Reitan & Wolfson, 1993), Raven's Progressive Matrices (Raven, 1982), and the Wisconsin Card Sorting Test (Grant & Berg, 1948; Heaton, Chelune, Talley, Kay, & Curtiss, 1993). Lezak (1995) devotes a separate chapter to evaluation of executive functions, which are identified as having four components: (1) volition, (2) planning, (3) purposive action, and (4) self-monitoring and regulation of performance. Goldstein and Green (1995) view problem solving and executive functions as separate, though related constructs. Problem solving is described as more specific (e.g., hypothesis generation, shifting response sets, divergent thinking, etc.) whereas executive functions are broader. Common to both are motivation, planning, execution, and evaluation of performance (Goldstein & Green, 1995). Lezak's discussion of tasks requiring executive functions covers tests also considered by others to be measures of intellectual and problem-solving skills, such as the Porteus Maze Test (Porteus, 1965). Although it is not uncommon to see a dissociation of function with preserved-intellectual and problem-solving skills in the context of impaired executive-function abilities related to frontal lobe functions, shared impairments are frequently seen, particularly with severe diffuse brain damage or disease. Tables 12.1 and 12.2 demonstrate that measures identified as requiring executive functions (e.g., the Category Test, Wisconsin Card Sorting Test, and Trailmaking B) show a high degree of association with WAIS-R Performance IQ subtests.

The Wechsler Adult Intelligence Scale, in its various revisions (i.e., Wechsler-Bellevue, WAIS, WAIS-R) is one of the most widely used measures of adult intelligence. Factor analyses of the WAIS-R and its predecessor, the WAIS, have yielded three factors: (1) Verbal Comprehension (loadings from Information, Comprehension, Vocabulary, and Similarities subtests); (2) Perceptual Organization (loadings from Picture Completion, Picture Arrangement, Block Design, and Object Assembly subtests); and (3) Freedom from Distractibility (loadings from Arithmetic and Digit Span subtests:

Sherman, Strauss, Spellacy, & Hunter, 1995; Smith et al., 1992). The Digit Symbol subtest shared loadings with Perceptual Organization and Freedom from Distractibility (Larrabee et al., 1983; Matarazzo, 1972), and Arithmetic has also demonstrated shared loadings with Verbal Comprehension (Larrabee et al., 1983; Matarazzo, 1972). The Mayo group has advocated interpretation of the WAIS-R by factor scores rather than the traditional VIQ, PIQ, FIQ analyses (Ivnik et al., 1994; Smith et al., 1992; Smith, Ivnik, Malec, Petersen, Kokmen, & Tangalos, 1994; Smith, Ivnik, Malec, & Tangalos, 1993).

Various short forms of the WAIS-R have been recommended. Smith and colleagues (1994) demonstrated adequate prediction of the Verbal Comprehension factor by Vocabulary and Information, and adequate prediction of the Perceptual Organization factor by Block Design and Picture Completion. A seven-subtest short form comprised of the WAIS-R Information, Digit Span, Arithmetic, Similarities, Picture Completion, Block Design and Digit Symbol subtests has been proposed by Ward (1990). This seven-subtest short form predicts well VIQ, PIQ, and FIQ scores based on the full WAIS-R administration, with composite reliabilities and standard errors of estimate comparable to the standard administration of the complete battery (Paolo & Ryan, 1993; Schrelten, Benedict, & Bobholz, 1994).

The sensitivity of the WAIS-R to cerebral dysfunction is widely established (Matarazzo, 1972; McFie, 1975; Reitan & Wolfson, 1993). Scores on the Wechsler scales are reduced in Alzheimer-type dementia (Fuld, 1984; Larrabee, Largen, et al., 1985), and in the context of moderate-to-severe closed head trauma (Dikmen, Machamer, Winn, & Temkin, 1995; Levin et al., 1982). Sherer, Scott, Parsons, and Adams (1994) found that the WAIS-R was as sensitive as the HRNB in discriminating brain-damaged from non-brain-damaged controls.

Lower Verbal IQ (VIQ) scores relative to Performance IQ (PIQ) scores have been associated with left hemisphere disease, with lower PIQ than VIQ scores associated with right hemisphere disease (Bornstein & Matarazzo, 1982). Some caveats are in order, because PIQ can be reduced in diffuse brain disease (Lezak, 1995) as well as in the context of aphasia (Hamsher, 1991; Larrabee, 1986). Indeed, Larrabee (1986) found that patients with left and right hemisphere CVA did not differ on PIQ until PIQ was statistically adjusted for aphasia severity. This effect extended to the most spatially "pure" WAIS subtest, Block Design.

Cautions are also indicated when considering subtest scatter. Ryan, Paolo, and Smith (1992) found that subtest scatter was no greater for brain-damaged than for normative subjects, when both samples were equivalent on IQ. Fuld (1984) has identified a pattern of WAIS-subtest performance she found to be more common in patients with Alzheimer-type dementia than in patients with multi-infarct dementia or other types of neurological dysfunction. Recently, Massman and Bigler (1993) conducted a meta-analytic review of 18 different studies covering over 3,700 subjects, and found that the sensitivity of the Fuld profile to Alzheimer-type dementia was low, 24.1 percent, although the specificity was much better, at 93.3 percent compared to normals and 88.5 percent compared to non-Alzheimer patients.

Larrabee, Largen, and colleagues (1985) found that of a combination of memory and WAIS- intelligence subtests, the VSRT was the most sensitive test in discriminating patients with AD from age, education, and gender-matched controls; however, in spite of its superiority in discriminating these two groups, the VSRT did not correlate at all with dementia severity. By contrast, WAIS Information and Digit Symbol were not only sensitive to the presence of dementia (albeit not as sensitive as the VSRT), but both of these WAIS subtests were also correlated with severity of dementia. This suggested that the primary utility of memory testing was in establishing the presence of dementia, while assessment of intellectual skills was useful in characterizing the severity of AD and determining the functional correlates of dementia.

One important outgrowth of the established sensitivity of the WAIS-R to dementia is the need to estimate pre-morbid intellectual function. One original method of analyzing the pattern of age and disease resistant ("hold") tests such as Vocabulary, relative to age and disease sensitive ("don't hold") tests such as Block Design and Digit Symbol, has not been supported in subsequent research. Although the "hold" tests show less of a decline relative to the "don't hold" tests in AD, basing the assessment of premorbid function on "hold" tests can underestimate premorbid IQ by as much as a full standard deviation (Larrabee, Largen, et al., 1985).

Investigators have taken advantage of the well-documented association of demographic factors such as educational and occupational status with

intelligence-test performance to develop regression equations for estimation of premorbid IQ. These have been developed based on the WAIS standardization data (Wilson, Rosenbaum, Brown, Rourke, Whitman, & Grisell, 1978) and based on the WAIS-R-standardization data (Barona, Reynolds, & Chastain, 1984). Recently, Paolo, Ryan, Troster, and Hilmer (1996) have extended this regression-estimation approach to estimation of WAIS-R-subtest-scales scores. Although these regression formulae can be quite useful, the standard errors of estimate are quite high (range of 12 to 13 IQ points: cf., Barona et al., 1984; range of 2.31 to 2.66 for subtest scaled scores: cf., Paolo et al., 1996).

Nelson and colleagues (Nelson, 1982; Nelson & O'Connell, 1978) have developed an estimate of pre-morbid IQ based on the ability of the patient to pronounce irregular words (e.g., "debt"), entitled the National Adult Reading Test (NART). The NART was originally standardized in comparison to the WAIS, on a sample in Great Britain. It is less sensitive to the effects of dementia than WAIS Vocabulary, but can be affected by aphasia and moderate-to-severe dementia (Crawford, 1992). One safeguard recommended by Crawford (1992) is to insure that the obtained NART score is within the expected range of a NART value estimated on the basis of demographic factors, prior to using the NART to estimate premorbid IQ. Obviously, if there is evidence that the NART has been affected by aphasia or dementia, the clinician must rely on the demographic-regression equations for estimation of premorbid IQ.

Blair and Spreen (1989) have developed a revision of the NART for a North American sample, the NART-R, for predicting WAIS-R IQ. This revised version or North American Adult Reading Test (NAART) was significantly associated with VIQ ($r = .83$), PIQ ($r = .40$) and FIQ ($r = .75$), with standard errors of estimate ranging from 6.56 for VIQ to 10.67 for FIQ. Addition of demographic variables accounted for a 3 percent increase in IQ variance, which was non-significant. Berry and colleagues (1994) published the first study to confirm the retrospective accuracy of the NART-R in predicting WAIS-R IQs obtained 3.5 years earlier in a group of normal older persons.

Two of the more widely used measures of concept formation and problem solving are the HRNB Category Test (Reitan & Wolfson, 1993) and the Wisconsin Card Sorting Test (Heaton et al., 1993). Both are thought to reflect aspects of frontal lobe function (Adams et al., 1995; Heaton et al., 1993), but performance is also affected by non-frontal dysfunction (Anderson, Damasio, Jones, & Tranel, 1991; Reitan & Wolfson, 1995). The Category Test is described in greater detail in chapter 10. The Wisconsin Cart Sorting Test (WCST) requires the patient to sort cards containing colored geometric forms of different shape and number to 4 target cards. The only examiner feedback is whether each sort is correct or incorrect. After the patient has reached a certain number correct in a row, the examiner changes the rule and the subject must switch conceptual sets. A number of scores can be computed, but the most sensitive scores are the number of perseverative responses and number of perseverative errors made (Heaton et al., 1993).

Tables 12.1 and 12.2 demonstrate an association of Category Test and WCST performance with a factor that is also defined by the WAIS-R PIQ subtests; however, Perrine (1993) found that the Category Test and WCST shared only 30 percent common variance when analyzed in the context of other concept-formation tasks. WCST performance was associated with attribute identification whereas Category Test scores were related to measures of rule learning and deduction of classification rules. More recently, Adams and coworkers (1995) found that performance on Subtest VII of the Category Test was correlated with local cerebral metabolic rate for glucose (LCMRG) in the cingulate, dorsolateral, and orbitomedial aspects of the frontal lobes in older alcoholic patients. By contrast, the Categories-achieved Score on the WCST was related to LCMRG in the cingulate region alone. Hence, the findings of Perrine (1993) and Adams and collaborators (1995) suggest that the Category Test and WCST are not interchangeable measures of problem solving and concept formation related to frontal cognitive skills.

Lezak (1995) discusses other measures of abstraction and frontal executive skills including maze problem-solving, the Tinkertoy Test, and measures of design generation. The design fluency measure developed by Jones-Gotman and Milner (1977) was intended as a nonverbal counterpart to the word-fluency procedure (see earlier discussion of Controlled Oral Word Association in the Language section of this chapter). This task requires the subject to "invent" nonsense drawings (i.e., without identifiable or recognizable meaning), under time constraints. Testing is conducted under a "free condition," and under a "fixed" condition (acceptable drawings are limited to four straight or

curved lines). Jones-Gotman and Milner (1977) originally reported an association of test impairment with right frontal excision.

Ruff (1996) has published the Ruff Figural Fluency Test. This test is a modification of an earlier procedure devised by Regard, Strauss, and Knapp (1982) to provide a measure of design fluency that was more reliable than the original Jones-Gotman and Milner (1977) procedure. Ruff's version requires the subject to produce multiple designs, connecting five symmetric and evenly spaced dots. Five different five-dot patterns are presented (two with interference). Ruff (1996) presents normative data, corrected for age and education, on 358 volunteers aged 16 to 70 years. The Ruff Figural Fluency Test loads on multiple factors including complex intelligence, planning, and arousal in normals and on planning and flexibility factors in head-injured patients (Baser & Ruff, 1987).

It is also important to evaluate academic achievement. Decline in calculational functions can be seen in dementing conditions (Cummings & Benson, 1992). As already discussed, oral reading tests have been used to predict pre-morbid ability in dementia (Blair & Spreen, 1989; Crawford, 1992; Wiens, Bryant, & Crossen, 1993). Achievement testing is also important in evaluating for learning disability. Assessment of learning disability may be the primary focus of a particular neuropsychology referral. Alternatively, it is important to rule out learning disability when evaluating young adults who have sustained closed head injury. Persons with learning disability can produce profiles suggestive of neuropsychological impairment that could be misinterpreted as secondary to trauma when these patterns actually represent preexisting problems (Larrabee, 1990).

Several measures of achievement exist (see chapter 7, this volume). The Woodcock-Johnson Psycho-Educational Battery-Revised (Woodcock & Mather, 1989) is probably one of the more comprehensive measures. Perhaps the most widely used battery in neuropsychological settings, which is more of a screening examination and shorter than the more Comprehensive Woodcock-Johnson, is the Wide Range Achievement Test (WRAT-3) (Wilkinson, 1993). Several studies have shown that learning-disabled persons perform in three reliably distinct patterns: (1) impaired oral reading and written spelling with preserved written calculations; (2) impaired reading, spelling, and arithmetic, and (3) impaired arithmetic but normal spelling and reading (Rourke, 1991). These patterns also have reliable extra-test correlates. Fletcher (1985) found the first subgroup (reading and spelling impaired) had impaired verbal relative to nonverbal learning and memory-test performance, whereas the converse was true for subgroup 3 (impaired math, normal reading and spelling), who performed poorer on nonverbal learning and memory relative to their verbal learning and memory. Rourke (1995) has reported extensively on the cognitive and emotional characteristics of the arithmetic-impaired subgroup, who frequently suffer from nonverbal learning disability.

Table 12.2 displays a complex loading pattern for the WRAT-R subtests. The primary loading of Reading, Spelling, and Math is on the first factor, which is also defined by the WAIS-R VIQ subtests. The three achievement tests also show a secondary loading on the attention and information-processing factor.

ASSESSMENT OF PERSONALITY, ADAPTIVE FUNCTIONS, AND MOTIVATION

Assessment of personality function is an important part of any comprehensive psychological or neuropsychological evaluation. The reader is referred to the chapters on personality assessment in this volume for more detailed consideration of this topic (see chapters 16 and chapter 17).

In neuropsychological settings, personality and emotional factors can relate to current status in a number of ways. Persons with preexisting psychiatric problems can have exacerbations of these problems post-injury or following disease of the central nervous system. Patients can develop personality change that is directly attributable to brain damage or disease, particularly if the frontal lobes, temporal lobes, or limbic system is involved (Heilman, Bowers, & Valenstein, 1993). Persons sustaining brain injury or disease can develop secondary emotional reactions to their disabilities or can sustain traumatic emotional reactions such as posttraumatic stress disorder in the course of sustaining their original physical injury.

As discussed in this volume, personality and emotional processes can be assessed via objective and projective instruments. In neuropsychological assessment, objective personality tests such as the MMPI or MMPI-2 are more frequently utilized

than projective measures (Butler, Retzlaff, & Vanderploeg, 1991).

Personality evaluation in patients who have brain injury or brain disease poses some unique problems. One potential consequence of significant frontal lobe trauma or degenerative conditions such as Alzheimer-type dementia is denial or minimization of deficit, termed *anosognosia* (Prigatano & Schacter, 1991). Consequently, persons with anosognosia may not endorse any personality-test or depression-test items in the clinically significant range, when indeed, symptoms are very significant. On the other hand, some (Alfano, Neilson, Paniak, & Finlayson, 1992; Gass, 1991) have advocated "neuro-correcting" the MMPI to remove those items related to neurologic factors, arguing that spurious elevations on MMPI scales may occur due to endorsing neurologically based complaints. Other research demonstrating a closer association of cognitive complaint with depression than with actual cognitive performance (Williams, Little, Scates, & Blockman, 1987; Larrabee & Levin, 1986) would argue against the need for such a correction. Indeed, Brulot, Strauss, and Spellacy (1997) recently reported that endorsement of MMPI Head Injury Scale items was related to the MMPI-2 Depression Content Scale, but not related to performance on neuropsychological tests or to measures of head-trauma severity such as loss of consciousness or posttraumatic amnesia.

One major advantage of the MMPI/MMPI-2 is that it allows an assessment of the validity of a particular patient's response pattern. Heaton and coworkers (1978) presented data on malingering which included the MMPI. Berry and colleagues (1995) present similar data on the MMPI-2. Both Heaton and collaborators (1978) and Berry and colleagues (1995) found that traditional MMPI/MMPI-2 validity scales (eg., F) were sensitive to malingering in normal subjects attempting to simulate brain injury.

One particular problem with the MMPI or MMPI-2 is that the validity scale most often relied upon to detect malingering is the F scale. This author (Larrabee, 1998) has seen patients who have been identified as malingerers by current objective measures of malingering (e.g., Portland Digit Recognition Test: Binder & Willis, 1991), who have "valid" MMPIs, with F scales below significant elevations, but have extreme elevations on scales 1 and 3. This is due to the fact that only 1 F-scale item is on either scales 1 or 3. What results is an extremely elevated "Conversion V" with scales 1 and 3 at values over T scores of 80, secondary to exaggerated somatic complaints. One way of addressing these extreme elevations on the somatic scales is to compare them to Heaton and colleagues scale 1 and 3 data for simulated malingerers on the MMPI, or to similar data for the Berry and coworkers, MMPI-2 malingerers. Elevations on scales 1 and 3 on the MMPI-2 can also be compared to the Keller and Butcher (1991) data on chronic-pain patients, particularly if pain is a feature of the presenting problems (a frequent occurrence in mild closed head trauma cases). Larrabee (1998) has demonstrated a pattern consistent with somatic malingering demonstrated by T scores at least 80 on scales 1 and 3, accompanied by an elevated score on the Lees-Haley Fake/Bad scale (Lees-Haley, 1992). Elevations on scales 1 and 3 that exceed the Keller-Butcher pain group values by over one standard deviation should be viewed as suspicious for exaggeration.

Clinician-based rating scales such as the Brief Psychiatric Rating Scale (BPRS) (Overall & Gorham, 1962), and the Neurobehavioral Rating Scales (NBRS) (Levin, High, et al., 1987) can be employed when the reliability and validity of a self-report instrument are suspect due to the patient's impaired neuropsychological status. The BPRS was developed for use with psychiatric patients and was also employed by Levin, Grossman, Rose, and Teasdale (1979) in an outcome study of patients with severe traumatic brain injury. Subsequently, Levin, High, and colleagues (1987) developed the NBRS as a measure more suited to the neurobehaviorally impaired head-injured patient. Factor analysis of the NBRS has yielded 4 factors: (1) Cognition/Energy, (2) Metacognition, (3) Somatic/Anxiety, and (4) Language. Factors 1, 2, and 4 were related to head-trauma severity, as well as to longitudinal recovery over time.

Recently, Nelson and collaborators (1989) and Nelson, Mitrushina, Satz, Sowa, and Cohen (1993) have developed the Neuropsychology Behavior and Affect Profile (NBAP). The NBAP is completed by relatives rating pre-illness behavior and emotional status as well as current functioning, on 106 items comprising five scales: (a) Indifference, (b) Inappropriateness, (c) Depression, (d) Mania, and (e) Pragnosia (a defect in the pragmatics of communications; e.g., "My relative often seems to miss the point of a discussion"). Nelson and coworkers (1989) provide evidence for high internal consistency and good discriminative validity

between normal elderly and dementia patients and between normal subjects and stroke patients.

Measures of adaptive functioning assess the patient's capacity to function effectively in their own environment. These measures include the more comprehensive rating scales such as the Cognitive Behavioral Rating Scale (Williams, 1987) which assesses a variety of functional areas, via family or friend ratings of the patient, including areas such as language, higher cognitive functions, orientation, skilled motor movement, agitation and memory, to the more specifically focused scales such as the Memory Assessment Clinic's Self and Family rating scales (MAC-S) (Crook & Larrabee 1990, 1992), (MAC-F) (Feher, Larrabee, Sudilovsky, & Crook, 1994). Scales such as the MAC-S and MAC-F, which include parallel self- and family-appraisal rating forms, allow for assessment of the patient's awareness of deficit and can allow quantification of anosognosia in patients who under-report difficulties (Feher et al., 1994). By contrast, greater self-report of impairment compared to ratings by relatives could suggest a potential depressive pseudo-dementia or a somatoform basis to cognitive complaint.

Assessment of motivation and cooperation has assumed an increasingly important role in the medico-legal arena. Over the past several years, there has been a significant increase in research on malingering, or the intentional production of false or exaggerated symptoms for secondary gain (American Psychiatric Association, 1994). Brandt (1988) has indicated that the only way a clinician can be certain of malingering is if the patient confesses. Obviously, confession rarely occurs. Malingering can involve both symptom report (Berry et al., 1995) and/or neuropsychological test performance (Brandt, 1988).

Several procedures have been developed to assist in detection of malingering. Malingering of symptom report has been discussed relative to the MMPI/MMPI-2. One of the major advances in detection of malingered neuropsychological test performance has been the application of forced-choice methodology and the binomial theorem to assess malingering (Binder, 1990; Binder & Pankratz, 1987; Hiscock & Hiscock, 1989). In a forced-choice task (e.g., identifying whether one was touched once or twice; identifying which of two five-digit numbers was previously presented), it is conceivable that someone with severe brain damage could perform at chance level; however, if someone does significantly worse-than-chance based on the binomial distribution, the assumption is made that they had to know the correct answer to perform at such an improbably poor level.

This interpretation is rationally and statistically appealing. Unfortunately, many persons whose behavior is suspicious for malingering may not perform at worse-than-chance level on forced-choice symptom-validity procedures. Hence, Binder and Willis (1991) performed a study contrasting the performance of persons with documented brain damage who were not seeking compensation, with a similar group seeking compensation, a group without brain damage but suffering major affective disorder, a group of minor head-trauma patients seeking compensation, a group of non-patient control subjects, and a group of non-patient subjects instructed to feign brain impairment, on the Portland Digit Recognition test (PDRT), a two-alternative forced-choice recognition-memory task. The lowest performance of all the groups was achieved by the non-patient simulators who averaged 50 percent correct on the 30-second-delay (Hard) condition of the PDRT. Using a cutoff of below the worst performance of the documented brain-damage group, up to 26 percent of the minor head trauma (MHT) group seeking compensation performed more poorly than all of the subjects who had documented brain damage. Binder and Willis also contrasted the performance of MHT subjects divided on PDRT performance into extreme groups (high vs. low motivation), on a variety of standard neuropsychological tests. They found significantly poorer performance for the low-motivation group on a variety of cognitive (eg., IQ, Digit Span), motor (eg., Fingertapping; Grooved Pegboard), and personality (SCL-90-R) measures. Subsequently, Binder, Villanueva, Howieson, and Moore (1993) demonstrated that MHT patients with poor PDRT performance also performed poorly on the Recognition trial of the RAVLT (mean score was 8, just above chance).

Lee, Loring, and Martin (1992) established cut-off scores for performance on the Rey-15 Item Test, an older measure of motivation developed by the French Psychologist Rey (1964). This task presents the subject with 15 items arranged in three columns by five rows. These items are presented for 10 seconds, then withdrawn with instructions to reproduce them from memory. Although there are 15 items, they can be grouped and clustered rather easily (eg., upper and lower case letters; Roman and Arabic numerals) so that even patients with significant brain injury can perform normally.

Indeed, 42 of 100 patients with temporal lobe epilepsy and documented memory impairment performed perfectly in the Lee and colleagues' (1992) study. Rey (1964) originally suggested that a score of nine or less was suggestive of malingering. Based on the distribution of performance for the temporal lobe epilepsy group, Lee and coworkers determined a cutoff of seven or less items for identification of malingered performance (only 4 percent of their memory-impaired epileptics performed this poorly). Six of 16 outpatient subjects in litigation, the majority with history of mild head trauma, performed at a level of seven or less correct. More recently, Greiffenstein, Baker, and Gola (1996) have provided data suggesting that Rey's original cutoff of nine or less, was both sensitive and specific to malingering, provided that a true organic amnestic disorder could be excluded on the basis of medical records.

Millis (1992) found that 50 percent of minor head-trauma patients performed more poorly than patients with moderate-to-severe closed head trauma on the Word Recognition subtest of the Warrington Recognition Memory Test. The MHT mean score approached chance, but was not worse-than-chance. Obviously, given the two-alternative forced-choice format, particularly poor performance on the Warrington Recognition Memory Test can also be evaluated with the binomial theorem, and the current author has seen two patients who performed at a significantly worse-than-chance level who also failed the Rey-15 Item Test and PDRT.

Other methodologies used in establishing patterns suspicious for malingering include comparison of normal persons instructed to feign impairment on standard psychological and neuropsychological tests with performance of persons having sustained moderate-to-severe traumatic brain injury (Heaton et al., 1978; Mittenberg et al., 1993; Mittenberg et al., 1995). Heaton and collaborators (1978) found that experimental malingerers performed more poorly than head-injured subjects on selected cognitive (eg., Digit Span), motor (tapping speed, grip strength), and personality (MMPI F scale, and scales 1, 3, 7, and 8) variables, despite out-performing the head injured on several sensitive tasks, including the Category Test and Tactual Performance Test.

Mittenberg and colleagues have contrasted the performance of normal subject (i.e., noninjured) simulators with that of non-litigating head- injured patients on the WAIS-R, WMS-R, and HRNB (Mittenberg, et al., 1993, 1995; Mittenberg, Rotholc, Russell, & Heilbronner, 1996). Although Mittenberg and colleagues derived discriminant functions for these various different tests, they also found that simple difference scores between WMS-R Attention Concentration (AC) and General Memory (GM) (AC lower than GM) and WAIS-R Digit Span significantly lower than Vocabulary, were nearly as effective as the complete discriminant functions in differentiating the experimental malingerers from head-injured subjects.

SCREENING BATTERIES

Earlier, in the discussion of test batteries such as the HRNB versus ability-focused, flexible approaches to neuropsychology, the issue of screening or core batteries was raised. As noted, Bauer (1994) recommended development of multiple fixed batteries, depending on the population being assessed. Earlier, Benton (1992) recommended development of a core battery of neuropsychological tests that could be administered in an hour or less. Subsequently, Parsons (1993a), as President of Division 40 (the Neuropsychology division of the APA), invited input from the membership regarding development of a 1 1/2-to-2 hour core test battery. Due to a primarily negative response, further investigation into the development of a core battery was dropped (Parsons, 1993b).

Since this time, the practice of clinical psychology in general, and neuropsychology specifically, has come under significant economic pressure from the impact of managed care companies on reimbursement for services. In a recent survey, Sweet, Westergaard, and Moberg (1995) found that 64 percent of respondents believed that national health-care reform would reduce patient evaluation time. Hence, it appears timely to reconsider Benton and Parson's previous recommendations for development of a core battery.

One of the biggest concerns regarding establishment of a core battery is that clinicians will be "forced" by insurance companies to administer a limited set of procedures to each patient, and there will be insufficient examination of the complexity of brain functions; however, it is quite possible to establish a core battery that is flexible and adaptable to the patient's needs by developing statistical and psychometric guidelines based on an integra-

tion of Bauer's "screening," "population specific" and "domain-specific" multiple battery approach, with what he terms a "tiered" approach (also described as a step battery by Tarter and Edwards, 1986). For example, in a patient with a history of left hemisphere stroke, one is already alerted to a population-specific need for screening of language and sensorimotor abilities. If this patient scores above average on screening measures of semantic and phonemic fluency, and visual-confrontation naming, there may be no need for more comprehensive, "domain specific" aphasia examination. Similarly, if the patient demonstrated average to above average dominant-hand fine-motor skills on the Grooved Pegboard, and normal dominant-hand performance on the Benton Tactile Form Perception Test, there may be no further testing required of more basic manual motor and manual tactile functions.

Several recent investigations are pertinent to the establishment of a core battery. The meta-analysis of Chouinard and Braun (1993), contrasted the relative sensitivity of various neuropsychological procedures in cases of diffuse cerebral dysfunction. Several investigators have shown that the WAIS-R can be reduced from the original eleven subtests to seven or fewer, with little appreciable loss in diagnostic or descriptive information (Paolo & Ryan, 1993; Smith et al., 1994; Ward, 1990). Sherer and colleagues' (1994) research, demonstrating equivalent sensitivity of the WAIS-R subtests to the HRNB in detecting brain dysfunction, and recent factor analyses of complex test batteries (Tables 12.1, 12.2) suggest a model for developing a core neuropsychological battery.

A core battery should cover the basic domains outlined in the current chapter as well as by other authors (eg., Chouinard & Braun, 1993; Lezak, 1995;), including language, perceptual/spatial, sensorimotor, attention, verbal and visual memory, and intellectual and problem-solving skills. In addition, multi-factorial tests such as WAIS-R Digit Symbol or Trailmaking B should be included, which are sensitive (i.e., to the presence of impairment), but not necessarily specific (i.e., as to which of several cognitive functions might be impaired). Initial development would require an over-sampling of each domain (e.g., for verbal memory, including several supraspan list-learning tasks such as the CVLT, AVLT, Selective Reminding, paired-associate learning, prose recall, and verbal recognition memory).

An ideal patient population for test development would be patients who have suffered moderate-to-severe closed head trauma. This population would encompass both diffuse central nervous system function as well as cases of focal injury superimposed on diffuse damage. This population would also have known biological markers of severity (Glasgow Coma Scale; duration of posttraumatic amnesia) which can be correlated with test performance (cf., Dikmen et al., 1995). Subgroups can be formed of subacute and chronic samples, and patients with and without mass lesions.

The over-inclusive battery would be administered to the subacute and chronic groups of patients. Validity could be established in a variety of fashions. Construct validity would be established through factor analysis. Criterion validity could be established through demonstration of associations of test performance with initial admission Glasgow Coma scale values, and by evaluating the association of different tests with relative-rating scales such as the Cognitive Behavior Rating Scales (CBRS) (Williams, 1987), and/or self- and family-rating scales such as the MAC-S or MAC-F (Crook & Larrabee, 1990, 1992; Feher et al., 1994). Discriminant validity could be evaluated by contrasting patient versus control-subject performance on the different tasks. Internal consistency and test-retest reliability would also be examined.

The above procedure would ideally result in identification of the most valid and reliable measures of each of the major neurobehavioral areas. Tests could be ranked by the size and purity of factor loadings, by sensitivity to trauma severity, by ecological validity (prediction by relative ratings) and by reliability. Certain inherently less reliable domains such as attention/information processing may require two measures.

Concurrently, subtests from "domain-specific" batteries could be directly compared in patient populations with left and right hemisphere CVA. Hence, the Multilingual Aphasia Examination, the Boston Diagnostic Aphasia Examination, and the Western Aphasia Battery subtests could be directly compared as to their respective sensitivities to language impairment in lefthemisphere CVA. The Benton Visual Form Discrimination, Facial Recognition, Line Orientation, and 3-Dimensional Constructional Praxis Test could be compared with measures such as the Hooper Visual Organizational Test, the Boston Parietal Lobe Battery, and the Gollin Figures as to sensitivity in right hemi-

sphere CVA. Both left and right- hemisphere CVA patients would be administered multiple measures of motor and tactile function, with determination of the most sensitive measures. Factor analyses could be conducted of performance on these domain-specific batteries. Cluster analysis could also be conducted on patterns of performance within each subgroup, followed by discriminant function analysis to identify the tasks contributing the most to cluster definition (c.f., Larrabee and Crook, 1989). Of additional interest would be analysis of the spatial/perceptual tasks that best discriminate left and right CVA patients, given the known association of aphasic comprehension deficit with performance on "nonverbal" tests, to establish which of these procedures were least likely to be failed by patients with left hemisphere CVA.

The domain-specific tests established as the most sensitive and having the best construct validity in the CVA groups could then be administered with the core procedures established in the head-trauma sample to explore interrelationships and contingencies of performance. Hence, it could be determined that a left CVA patient with normal WAIS-R Vocabulary and Boston Naming would not need to be evaluated further for language impairment and that same patient who has normal Block Design would not have to be administered Line Orientation.

Finally, the contingencies of performance on the core battery could be examined in probable Alzheimer-type dementia. For example assume the core-battery study identified the RAVLT as the most appropriate supraspan learning test relative to the CVLT and Selective Reminding. Further, assume a measure of paired-associate learning and the Warrington Word Recognition Memory score also loaded on the verbal memory factor, but were not as sensitive to impairment as the RAVLT. These contingencies could be evaluated in the Alzheimer group, such that if the RAVLT was failed, one would need to explore in addition, paired-associate learning and word-recognition memory. If a certain level of recognition memory was necessary for possible success on RAVLT or paired-associate learning, then on reassessment a year later, the recognition memory test alone would need to be administered, if this critical level was not exceeded. Once performance reached a certain floor on recognition memory, there would be no further examination of memory on future follow-up of the patient.

The above discussion suggests an approach for developing a core examination, in three groups of patients most frequently seen for neuropsychological evaluation. The resultant core battery would not pose unnecessarily restrictive limitations on the evaluation of a particular patient. For clinicians employing a fixed battery approach such as the HRNB or LNNB, or domain-specific batteries such as the WMS-R, failure on subtests of the core examination would justify administration of the more comprehensive battery. For clinicians employing an individualized approach, additional assessment can also be justified on the basis of performance patterns on core battery subtests. Assume that in addition to the RAVLT, the core battery contains Controlled Oral Word Association, the Grooved Pegboard, Trailmaking B, the PASAT, Rey-Osterrieth Complex Figure Test (CFT), WAIS-R Block Design and Digit Symbol. The head-injured patient who fails RAVLT would also need to be examined on the less sensitive verbal memory measures to explore completely their verbal learning difficulties. By contrast, the patient who performs normally on COWA, Grooved Pegboard, Trailmaking B, PASAT, RAVLT, CFT, Block Design, and Digit Symbol, would not need more detailed exploration of other language, perceptual, sensorimotor, attentional, memory, or intellectual and problem-solving skills.

AUTHOR NOTES

Glenn J. Larrabee, Center for Neuropsychological Studies, University of Florida, and Sarasota Memorial Hospital.

The author gratefully acknowledges the assistance of Susan Towers and Kristin Kravitz in the preparation of this chapter.

Correspondence concerning this chapter should be addressed to Glenn J. Larrabee, Ph.D., 630 South Orange Avenue, Suite 202, Sarasota, FL 34236.

REFERENCES

Adams, K. M., Gilman, S., Koeppe, R., Kluin, K., Junck, L., Lohman, M., Johnson-Greene, D., Berent, S., Dede, D., & Kroll, P. (1995). Correlation of neuropsychological function with cerebral metabolic rate in subdivisions of the frontal lobes of older alcoholic patients measured with [$_{18}$F] fluo-

rodeoxy-glucose and positron emission tomography. *Neuropsychology, 9,* 275–280.

Agnew, J., Bolla-Wilson, K., Kawas, C. H., & Bleecker, M. L. (1988). Purdue Pegboard age and sex norms for people 40 years old and older. *Developmental Neuropsychology, 4,* 29–26.

Albert, M. S., Butters, N., & Brandt, J. (1981). Patterns of remote memory in amnestic and demented patients. *Archives of Neurology, 38,* 495–500.

Albert, M. S., Butters, N., & Levin, J. (1979). Temporal gradients in the retrograde amnesia of patients with alcoholic Korsakoff's disease. *Archives of Neurology, 36,* 211–216.

Alfano, D. P., Neilson, P. M., Paniak, C. E., & Finlayson, M. A. J. (1992). The MMPI and closed head injury. *The Clinical Neuropsychologist, 6,* 134–142.

American Psychiatric Association (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Anderson, S. W., Damasio, H., Jones, R. D., & Tranel, D. (1991). Wisconsin Card Sorting Test performance as a measure of frontal lobe damage. *Journal of Clinical and Experimental Neuropsychology, 13,* 909–922.

Axelrod, B. N., Ricker, J. H., & Cherry, S. A. (1994). Concurrent validity of the MAE Visual Naming Test. *Archives of Clinical Neuropsychology, 9,* 317–321.

Barona, A., Reynolds, C. R., & Chastain, R. (1984). A demographically based index of premorbid intelligence for the WAIS-R. *Journal of Consulting and Clinical Psychology, 5,* 885–887.

Baser, C. A., & Ruff, R. M. (1987). Construct validity of the San Diego Neuropsychological Test Battery. *Archives of Clinical Neuropsychology, 2,* 13–32.

Bauer, R. M. (1993). Agnosia. In K. M. Heilman & E. Valenstein (Eds.), *Clinical neuropsychology* (3rd ed., pp. 215–278). New York: Oxford.

Bauer, R. M. (1994). The flexible battery approach to neuropsychology. In R. D. Vanderploeg (Ed.), *Clinician's guide to neuropsychological assessment* (pp. 259–290). Hillsdale, NJ: Lawrence Erlbaum Associates.

Beatty, W. W., Krull, K. R., Wilbanks, S. L., Blanco, C. R., Hames, K. A., & Paul, R. H. (1996). Further validation of constructs from the Selective Reminding Test. *Journal of Clinical and Experimental Neuropsychology, 18,* 52–55.

Beatty, W. W., Salmon, D. P., Butters, N., Heindel, W. C., & Granholm, E. (1988). Retrograde amnesia in patients with Alzheimer's disease and Huntington's disease. *Neurobiology of Aging, 9,* 181–186.

Benson, D. F. (1979). *Aphasia, alexia, and agraphia.* New York: Churchill Livingstone.

Benson, D. F. (1993). Aphasia. In K. M. Heilman and E. Valenstein (Eds.), *Clinical Neuropsychology* (3rd ed., pp. 17–36). New York: Oxford.

Benton, A. L. (1968). Differential behavioral effects in frontal lobe disease. *Neuropsychologia, 6,* 53–60.

Benton, A. L. (1992). Clinical neuropsychology: 1960–1990. *Journal of Clinical and Experimental Neuropsychology, 14,* 407–417.

Benton, A. L., Hamsher, K. deS., and Sivan, A. B. (1994). *Multilingual Aphasia Examination* (3rd ed.). Iowa City, IA: AJA Associates, Inc.

Benton, A. L., Sivan, A. B., Hamsher, K. deS., Varney, N. R., & Spreen, O. (1994). *Contributions to neuropsychological assessment. A clinical manual* (2nd ed.). New York: Oxford.

Berry, D. T. R., Carpenter, G. S., Campbell, D. A., Schmitt, F. A., Helton, K., & Lipke-Molby, T. (1994). The New Adult Reading Test-Revised: Accuracy in estimating WAIS-R IQ scores obtained 3.5 years earlier from normal persons. *Archives of Clinical Neuropsychology, 9,* 239–250.

Berry, D. T. R., Wetter, M. W., Baer, R. A., Youngjohn, J. R., Gass, C. S., Lamb, D. G., Franzen, M. D., MacInnes, W. D., & Bucholz, D. (1995). Overreporting of closed-head injury symptoms on the MMPI-2. *Psychological Assessment, 7,* 517–523.

Binder, L. M. (1990). Malingering following minor head trauma. *The Clinical Neuropsychologist, 4,* 25–36.

Binder, L. M., & Pankratz, L. M. (1987). Neuropsychological evidence of a factitious memory complaint. *Journal of Clinical and Experimental Neuropsychology, 9,* 167–171.

Binder, L. M., Villanueva, M. R., Howieson, D., & Moore, R. T. (1993). The Rey AVLT Recognition Memory Task measures motivational impairment after mild head trauma. *Archives of Clinical Neuropsychology, 8,* 137–147.

Binder, L. M., & Willis, S. C. (1991). Assessment of motivation after financially compensable minor head trauma. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 3,* 175–181.

Blair, J. R., & Spreen, O. (1989). Predicting premorbid IQ: A revision of the National Adult Reading Test. *The Clinical Neuropsychologist, 3,* 129–136.

Bornstein, R. A., & Matarazzo, J. D. (1982). Wechsler VIQ versus PIQ differences in cerebral dysfunction: A literature review with emphasis on sex differences. *Journal of Clinical Neuropsychology, 4,* 319–334.

Brandt, J. (1988). Malingered amnesia. In R. Rogers (Ed.), *Clinical assessment of malingering and deception.* (pp. 65–83). New York: Guilford Press.

Braun, C. M. J., Daigneault, H. S., & Gilbert, B. (1989). Color discrimination testing reveals early printshop solvent neurotoxicity better than a neuropsychological test battery. *Archives of Clinical Neuropsychology, 4,* 1–13.

Brittain, J. L., LaMarche, J. A., Reeder, K. P., Roth, D. L., & Boll, T. J. (1991). Effects of age and IQ on Paced Auditory Serial Addition Task (PASAT) performance. *The Clinical Neuropsychologist, 5,* 163–175.

Brown, J. (1958). Some tests of the decay theory of immediate memory. *Quarterly Journal of Experimental Psychology, 10,* 12–21.

Brulot, M. M., Strauss, E. H., & Spellacy, F. (1997). Validity of Minnesota Multiphasic Personality Inventory-2 correction factors for use with patients with suspected head injury. *The Clinical Neuropsychologist, 11,* 391–401.

Buschke, H. (1973). Selective reminding for analysis of memory and learning. *Journal of Verbal Learning and Verbal Behavior, 12,* 543–550.

Butler, M., Retzlaff, P., & Vanderploeg, R. (1991). Neuropsychological test usage. *Professional Psychology: Research and Practice, 22,* 510–512.

Butler, R. W., Rorsman, I., Hill, J. M., & Tuma, R. (1993). The effects of frontal brain impairment on fluency: Simple and complex paradigms. *Neuropsychology, 7,* 519–529.

Butters, N., & Cermak, L. S. (1980). *Alcoholic Korsakoff's syndrome: An information processing approach.* New York: Academic Press.

Butters, N., Grant, I., Haxby, J., Judd, L. J., Martin, A., McClelland, J., Pequegnat, W., Schacter, D., & Stover, E. (1990). Assessment of AIDS-related cognitive changes: Recommendations of the NIMH work group on neuropsychological assessment approaches. *Journal of Clinical and Experimental Neuropsychology, 12,* 963–978.

Butters, N., Salmon, D. P., Cullum, C. M., Cairns, P., Troster, A. I., Jacobs, D., Moss, M., & Cermak, L. S. (1988). Differentiation of amnestic and demented patients with the Wechsler Memory Scale-Revised. *The Clinical Neuropsychologist, 2,* 133–148.

Capruso, D. X., Hamsher, K. deS., & Benton, A. L. (1995). Assessment of visuocognitive processes. In R. L. Mapou & J. Spector (Eds.), *Clinical neuropsychological assessment. A cognitive approach* (pp. 137–183). New York: Plenum.

Chouinard, M.J., & Braun, C.M.J. (1993). A meta-analysis of the relative sensitivity of neuropsychological screening tests. *Journal of Clinical and Experimental Neuropsychology, 15,* 591–607.

Cohen, R. A. (1993). *The neuropsychology of attention.* New York: Plenum.

Costa, L. D. (1975). The relation of visuospatial dysfunction to digit span performance in patients with cerebral lesions. *Cortex, 11,* 31–36.

Craik, F. I. M. (1984). Age differences in remembering. In L. R. Squire & N. Butters (Eds.), *Neuropsychology of memory* (pp. 3–12). New York: Guilford Press.

Crawford, J. R. (1992). Current and premorbid intelligence measures in neuropsychological assessment. In J. R. Crawford, D. M. Parker, & W. W. McKinlay (Eds.), *A handbook of neuropsychological assessment* (pp. 21–49). Hillsdale, NJ: Lawrence Erlbaum.

Crawford, J. R., Stewart, L. E., & Moore, J. W. (1989). Demonstration of savings on the AVLT and development of a parallel form. *Journal of Clinical and Experimental Neuropsychology, 11,* 975–981.

Crook, T. H., & Larrabee, G. J. (1990). A self-rating scale for evaluating memory in everyday life. *Psychology and Aging, 5,* 48–57.

Crook, T. H., & Larrabee, G. J. (1992). Normative data on a self-rating scale for evaluating memory in everyday life. *Archives of Clinical Neuropsychology, 7,* 41–51.

Crosson, B. (1992). *Subcortical functions in language and memory.* New York: Guilford.

Crosson, B., Novack, T. A., Trenerry, M. R., & Craig, P. L. (1988). California Verbal Learning Test (CVLT) performance in severely head-injured and neurologically normal adult males. *Journal of Clinical and Experimental Neuropsychology, 10,* 754–768.

Cummings, J. L., & Benson, D. F. (1992). *Dementia. A clinical approach* (2nd ed.). Boston: Butterworth-Heinemann.

Davis, H. P., & Bernstein, P. A. (1992). Age-related changes in explicit and implicit memory. In L. R. Squire & N. Butters (Eds.), *Neuropsychology of memory* (2nd ed., pp. 249–261). New York: Guilford.

Delis, D. C., Freeland, J., Kramer J. H., & Kaplan, E. (1988). Integrating clinical assessment with cognitive neuroscience: Construct validation of a multivariate verbal learning test. *Journal of Consulting and Clinical Psychology, 56,* 123–130.

Delis, D. C., Kramer, J. H., Kaplan, E., & Ober, B. A. (1987). *California Verbal Learning Test. Research edition. Manual.* San Antonio, TX: Psychological Corporation.

Delis, D. C., McKee, R., Massman, P. J., Kramer, J. H., Kaplan, E., & Gettman, D. (1991). Alternate form of the California Verbal Learning Test: Development and Reliability. *The Clinical Neuropsychologist, 5,* 154–162.

DeRenzi, E., & Vignolo, L. A. (1962). The Token Test: A sensitive test to detect disturbances in aphasics. *Brain, 85,* 665–678.

Diamond, B. J., & Deluca, J. (1996). Rey-Osterrieth Complex Figure Test performance following anterior communicating artery aneurysm. *Archives of Clinical Neuropsychology, 11*, 21–28.

Dikmen, S. S., Machamer, J. E., Winn, H. R., & Temkin, N. R. (1995). Neuropsychological outcome at 1-year post head injury. *Neuropsychology, 9*, 80–90.

Dodrill, C. B. (1978). A neuropsychological battery for epilepsy. *Epilepsia, 19*, 611–623.

Erickson, R. C., & Scott, M. L. (1977). Clinical memory testing: A review. *Psychological Bulletin, 84*, 1130–1149.

Eslinger, P. J., & Benton, A. L. (1983). Visuoperceptual performances in aging and dementia: Clinical and theoretical implications. *Journal of Clinical Neuropsychology, 5*, 213–220.

Eslinger, P. J., Damasio, A. R., Benton, A. L., & VanAllen, M. (1985). Neuropsychologic detection of abnormal mental decline in older persons. *Journal of the American Medical Association, 253*, 670–674.

Feher, E. P., Larrabee, G. J., Sudilovsky, A., & Crook, T. H. (1994). Memory self-report in Alzheimer's disease and in age-associated memory impairment. *Journal of Geriatric Psychiatry and Neurology, 7*, 58–65.

Fletcher, J. M. (1985). External validation of learning disability typologies. In B. P. Rourke (Ed.), *Neuropsychology of learning disabilities* (pp. 187–211). New York: Guilford.

Fuld, P. A. (1984). Test profile of cholinergic dysfunction and of Alzheimer-type dementia. *Journal of Clinical Neuropsychology, 5*, 380–392.

Gass, C. S. (1991). MMPI-2 interpretation and closed head injury: A correction factor. *Psychological Assessment, 3*, 27–31.

Geffen, G., Moar, K. J., O'Hanlon, A. P., Clark, C. R., & Geffen, L. B. (1990). Performance measures of 16-to-86-year-old males and females on the Auditory Verbal Learning Test. *The Clinical Neuropsychologist, 4*, 45–63.

Golden, C. J. (1978). *Stroop Color and Word Test*. Chicago: Stoelting.

Golden, C. J., Purisch, A. D., & Hammeke, T. A. (1985). *Luria-Nebraska Neuropsychological Battery: Forms I and II*. Los Angeles, CA: Western Psychological Services.

Goldstein, F. C., & Green, R. C. (1995). Assessment of problem solving and executive functions. In R. L. Mapou & J. Spector (Eds.), *Clinical neuropsychological assessment. A cognitive approach* (pp. 49–81). New York: Plenum.

Goodglass, H., & Kaplan, E. F. (1983). *Assessment of aphasia and related disorders* (2nd ed.). Philadelphia: Lea and Febiger.

Gorham, D. R. (1956). A Proverbs Test for clinical and experimental use. *Psychological Reports, 2*, 1–12.

Grant, D. A., & Berg, E. A. (1948). A behavioral analysis of the degree of reinforcement and ease of shifting to new responses in a Weigl-type card sorting problem. *Journal of Experimental Psychology, 38*, 404–411.

Greiffenstein, M. F., Baker, J. W., & Gola, T. (1996). Comparison of multiple scoring methods for Rey's malingered amnesia measures. *Archives of Clinical Neuropsychology, 11*, 283–293.

Gronwall, D. M. A. (1977). Paced auditory serial-addition task: A measure of recovery from concussion. *Perceptual and Motor Skills, 44*, 367–373.

Hamsher, K. deS. (1982). *Presidents Test: Manual of instructions*. Milwaukee, WI: University of Wisconsin Medical School at Mt. Sinai Medical Center.

Hamsher, K. deS. (1990). Specialized neuropsychological assessment methods. In G. Goldstein & M. Hersen (Eds.), *Handbook of psychological assessment* (2nd ed., pp. 256–279). New York: Pergamon.

Hamsher, K. deS. (1991). Intelligence and aphasia. In M.T. Sarno (Ed.), *Acquired aphasia* (2nd ed., pp. 339–372). San Diego, CA: Academic Press.

Hamsher, K. deS., & Roberts, R. J. (1985). Memory for recent U.S. presidents in patients with cerebral disease. *Journal of Clinical and Experimental Neuropsychology, 7*, 1–13.

Hannay, H. J., & Levin, H. S. (1985). Selective Reminding Test: An examination of the equivalence of four forms. *Journal of Clinical and Experimental Neuropsychology, 7*, 251–263.

Hannay, H. J., & Levin, H. S. (1989). Visual continuous recognition memory in normal and closed-head-injured adolescents. *Journal of Clinical and Experimental Neuropsychology, 11*, 444–460.

Hannay, H. J., Levin, H. S., & Grossman, R. G. (1979). Impaired recognition memory after head injury. *Cortex, 15*, 269–283.

Heaton, R. K., Chelune, G. J., Talley, J. L., Kay, G. G., & Curtiss, G. (1993). *Wisconsin Card Sorting Test manual. Revised and expanded*. Odessa, FL: Psychological Assessment Resources.

Heaton, R. K., Grant, I., & Matthews, C. G. (1991). *Comprehensive norms for an expanded Halstead-Reitan Battery: Demographic corrections, research findings, and clinical applications*. Odessa, FL: Psychological Assessment Resources.

Heaton, R. K., Smith, H. H., Jr., Lehman, R. A., & Vogt, A. J. (1978). Prospects for faking believable deficits on neuropsychological testing. *Journal of Consulting and Clinical Psychology, 46*, 892–900.

Heilman, K. M., Bowers, D., & Valenstein, E. (1993). Emotional disorders associated with neurological dis-

eases. In K.M. Heilman & E. Valenstein (Eds.), *Clinical neuropsychology* (3rd ed., pp. 461–497). New York: Oxford.

Heilman, K. M., Watson, R. T., & Valenstein, E. (1993). Neglect and related disorders. In K. M. Heilman & E. Valenstein (Eds.), *Clinical neuropsychology* (3rd ed., pp. 279–336). New York: Oxford.

Hermann, B. P., Seidenberg, M., Wyler, A., & Haltiner, A. (1993). Dissociation of object recognition and spatial localization abilities following temporal lobe lesions in humans. *Neuropsychology, 7,* 343–350.

High, W. M., Levin, H. S., & Gary, H. E. (1990). Recovery of orientation following closed head injury. *Journal of Clinical and Experimental Neuropsychology, 12,* 703–714.

Hiscock, M., & Hiscock, C. K. (1989). Refining the forced-choice method for the detection of malingering. *Journal of Clinical and Experimental Neuropsychology, 11,* 967–974.

Hom, J., & Reitan, R. M. (1982). Effect of lateralized cerebral damage upon contralateral and ipsilateral sensorimotor performance. *Journal of Clinical Neuropsychology, 4,* 249–268.

Ivnik, R. J., Malec, J. F., Sharbrough, F. W., Cascino, G. D., Hirschorn, K. A., Crook, T. H., & Larrabee, G. J. (1993). Traditional and computerized assessment procedures applied to the evaluation of memory change after temporal lobectomy. *Archives of Clinical Neuropsychology, 8,* 69–81.

Ivnik, R .J., Malec, J. F., Smith, G. E., Tangalos, E. G., Petersen, R. C., Kokmen, E., & Kurland, L. T. (1992a). Mayo's Older Americans Normative Studies: WAIS-R Norms for ages 56 to 97. *The Clinical Neuropsychologist, 6,* (suppl.), 1–30.

Ivnik, R .J., Malec, J. F., Smith, G. E., Tangalos, E. G., Petersen, R. C., Kokmen, E., & Kurland, L. T. (1992b). Mayo's Older Americans Normative Studies: WMS-R Norms for ages 56 to 97. *The Clinical Neuropsychologist, 6*(Suppl.), 48–81.

Ivnik, R. J., Malec, J. F., Smith, G. E., Tangalos, E. G., Petersen, R. C., Kokmen, E., & Kurland, L. T. (1992c). Mayo's Older Americans Normative Studies: Updated AVLT Norms for ages 56 to 97. *The Clinical Neuropsychologist, 6*(Suppl.), 82–103.

Ivnik, R. J., Smith, G. E., Malec, J. F., Kokmen, E., & Tangalos, E. G. (1994). Mayo Cognitive Factor Scales: Distinguishing normal and clinical samples by profile variability. *Neuropsychology, 8,* 203–209.

Jarvis, P. E., & Barth, J. T. (1994). *The Halstead-Reitan Neuropsychological Battery. A guide to interpretation and clinical applications.* Odessa, FL: Psychological Assessment Resources.

Jastak, S., & Wilkinson, G. S. (1984). *The Wide Range Achievement Test-Revised. Administration manual.* Wilmington, DE: Jastak Associates, Inc.

Jones, R. D., & Benton, A. L. (1994, February). *Use of the Multilingual Aphasia Examination in the detection of language disorders.* Paper presented at the Twenty Second Annual Meeting of the International Neuropsychological Society, Cincinnati, OH.

Jones-Gotman, M., & Milner, B. (1977). Design fluency: The invention of nonsense drawings after focal cortical lesions. *Neuropsychologia, 15,* 653–674.

Kane, R. L. (1991). Standardized and flexible batteries in neuropsychology: An assessment update. *Neuropsychology Review, 2,* 281–339.

Kaplan, E., Goodglass, H., & Weintraub, S. (1983). *The Boston Naming Test.* Philadelphia: Lea & Febiger.

Keller, L. S., & Butcher, J. N. (1991). *Assessment of chronic pain patients with the MMPI-2.* Minneapolis, MN: University of Minnesota Press.

Kertesz, A. (1982). *Western Aphasia Battery.* San Antonio, TX: The Psychological Corporation.

Kimura, D. (1963). Right temporal lobe damage. *Archives of Neurology, 8,* 264–271.

Kramer, J. H., Levin, B. E., Brandt, J., & Delis, D. C. (1989). Differentiation of Alzheimer's, Huntington's and Parkinson's disease patients on the basis of verbal learning characteristics. *Neuropsychology, 3,* 111–120.

Larrabee, G. J. (1986). Another look at VIQ-PIQ scores and unilateral brain damage. *International Journal of Neuroscience, 29,* 141–148.

Larrabee, G. J. (1990). Cautions in the use of neuropsychological evaluation in legal settings. *Neuropsychology, 4,* 239–247.

Larrabee, G. J. (1998). Somatic malingering on the MMPI and MMPI-2 in personal injury litigants. *The Clinical Neuropsychologist, 12,* 179-188.

Larrabee, G. J., & Crook, T. H., III. (1989). Performance subtypes of everyday memory function. *Developmental Neuropsychology, 5,* 267–283.

Larrabee, G. J., & Crook, T. H., III. (1995). Assessment of learning and memory. In R.L. Mapou & J. Spector (Eds.), *Clinical neuropsychological assessment. A cognitive approach.* (pp. 185–213). New York: Plenum.

Larrabee, G. J., & Curtiss, G. (1992). Factor structure of an ability-focused neuropsychological battery [Abstract]. *Journal of Clinical and Experimental Neuropsychology, 14,* 65.

Larrabee, G. J., & Curtiss, G. (1995). Construct validity of various verbal and visual memory tests. *Journal of Clinical and Experimental Neuropsychology, 17,* 536–547.

Larrabee, G. J., & Kane, R. L. (1986). Reversed digit repetition involves visual and verbal processes. *International Journal of Neuroscience, 30,* 11–15.

Larrabee, G. J., Kane, R. L., & Schuck, J. R. (1983). Factor analysis of the WAIS and Wechsler Memory Scale: An analysis of the Construct Validity of the Wechsler Memory Scale. *Journal of Clinical Neuropsychology, 5,* 159–168.

Larrabee, G. J., Kane, R. L., Schuck, J. R., & Francis, D. J. (1985). The construct validity of various memory testing procedures. *Journal of Clinical and Experimental Neuropsychology, 7,* 239–250.

Larrabee, G. J., Largen, J. W., & Levin, H. S. (1985). Sensitivity of age-decline resistant ("Hold") WAIS subtests to Alzheimer's disease. *Journal of Clinical and Experimental Neuropsychology, 7,* 497–504.

Larrabee, G. J., & Levin, H. S. (1984, February). Verbal visual and remote memory test performance in a normal elderly sample. Paper presented at the Twelfth Annual Meeting of the International Neuropsychological Society, Houston, TX.

Larrabee, G. J., & Levin, H. S. (1986). Memory self-ratings and objective test performance in a normal elderly sample. *Journal of Clinical and Experimental Neuropsychology, 8,* 275–284.

Larrabee, G. J., Levin, H. S., Huff, J., Kay, M. C., & Guinto, F. C. (1985). Visual agnosia contrasted with visual-verbal disconnection. *Neuropsychology, 23,* 1–12.

Larrabee, G. J., Trahan, D. E., & Curtiss, G. (1992). Construct validity of the Continuous Visual Memory Test. *Archives of Clinical Neuropsychology, 7,* 395–405.

Larrabee, G. J., Trahan, D. E., Curtiss, G., & Levin, H. S. (1988). Normative data for the Verbal Selective Reminding Test. *Neuropsychology, 2,* 173–182.

Larrabee, G. J., Youngjohn, T. R., Sudilovsky, A., & Crook, T. H., III. (1993). Accelerated forgetting in Alzheimer-type dementia. *Journal of Clinical and Experimental Neuropsychology, 14,* 701–712.

Lee, G. P., Loring, D. W., & Martin, R. C. (1992). Rey's 15-item Visual Memory Test for the detection of malingering: Normative observations on patients with neurological disorders. *Psychological Assessment, 4,* 43–46.

Lees-Haley, P. R. (1992). Efficacy of MMPI-2 validity scales and MCMI-II modifier scales for detecting spurious PTSD claim: F, F-K, Fake Bad Scale, Ego Strength, Subtle-Obvious subscales, Dis, and Deb. *Journal of Clinical Psychology, 48,* 681–688.

Leonberger, F. T., Nicks, S. D., Larrabee, G. J., & Goldfader, P. R. (1992). Factor structure of the Wechsler Memory Scale-Revised within a comprehensive neuropsychological battery. *Neuropsychology, 6,* 239–249.

Levin, H. S., Benton, A. L., & Grossman, R. G. (1982). *Neurobehavioral consequences of closed head injury.* New York: Oxford.

Levin, H. S., Grossman, R. G., Rose, J. E., & Teasdale, G. (1979). Long-term neuropsychological outcome of closed head injury. *Journal of Neurosurgery, 50,* 412–422.

Levin, H. S., High, W. M., Ewing-Cobbs, L., Fletcher, J. M., Eisenberg, H. M., Miner, M. E., & Goldstein, F. C. (1988). Memory functioning during the first year after closed head injury in children and adolescents. *Neurosurgery, 22,* 1043–1052.

Levin, H. S., High, W. M., Goethe, K. E., Sisson, R. A., Overall, J. E., Rhoades, H. M., Eisenberg, H. M., Kalisky, Z., & Gary, H. E. (1987). The neurobehavioral rating scale: Assessment of the behavioral sequelae of head injury by the clinician. *Journal of Neurology, Neurosurgery, and Psychiatry, 50,* 183–193.

Levin, H. S., High, W. M., Meyers, C. A., Von Laufen, A., Hayden, M. E., & Eisenberg, H. M. (1985). Impairment of remote memory after closed head injury. *Journal of Neurology, Neurosurgery, & Psychiatry, 48,* 556–563.

Levin, H. S., & Larrabee, G. J. (1983). Disproportionate decline in visuo-spatial memory in human aging. Society for Neurosciences Abstracts, 9, 918.

Levin, H. S., Mattis, S., Ruff, R. M., Eisenberg, H. M., Marshall, L. F., Tabaddor, K., High, W. M., & Frankowski, R. F. (1987). Neurobehavioral outcome following minor head injury: A three-center study. *Journal of Neurosurgery, 66,* 234–243.

Levin, H. S., O'Donnell, V. M., & Grossman, R. G. (1979). The Galveston Orientation and Amnesia Test: A practical scale to assess cognition after head injury. *Journal of Nervous and Mental Disease, 167,* 675–684.

Lezak, M. D. (1976). *Neuropsychological assessment.* New York: Oxford.

Lezak, M. D. (1995). *Neuropsychological assessment* (3rd ed.). New York: Oxford.

Loring, D. W., Martin, R. C., Meador, K. J., & Lee, G. P. (1990). Psychometric construction of the Rey-Osterrieth Complex Figure: Methodological considerations and interrater reliability. *Archives of Clinical Neuropsychology, 5,* 1–14.

Loring, D. W., & Papanicolaou, A. C. (1987). Memory assessment in neuropsychology: Theoretical considerations and practical utility. *Journal of Clinical and Experimental Neuropsychology, 9,* 340–358.

Malec, J. F., Ivnik, R. J., & Hinkeldey, N. S. (1991). Visuospatial Learning Test. *Psychological Assessment: A*

Journal of Consulting and Clinical Psychology, 3, 82–88.

Mapou, R. L., & Spector, J. (Eds.). (1995). *Clinical neuropsychological assessment. A cognitive approach.* New York: Plenum.

Marsh, N. V., & Knight, R. G. (1991). Relationship between cognitive deficits and social skill after head injury. *Neuropsychology, 5,* 107–117.

Marson, D. C., Cody, H. A., Ingram, K. K., & Harrell, L. E. (1995). Neuropsychologic predictors of competency in Alzheimer's disease using a rational reasons legal standard. *Archives of Neurology, 52,* 955–959.

Martin, R. C., Loring, D. W., Meador, K. J., & Lee, G. P. (1988). Differential forgetting in patients with temporal lobe dysfunction. *Archives of Clinical Neuropsychology, 3,* 351–358.

Massman, P. J., & Bigler, E. D. (1993). A quantitative review of the diagnostic utility of the WAIS-R Fuld profile. *Archives of Clinical Neuropsychology, 8,* 417–428.

Masur, D. M., Fuld, P. A., Blau, A. D., Crystal, H., & Aronson, M. K. (1990). Predicting development of dementia in the elderly with the Selective Reminding Test. *Journal of Clinical and Experimental Neuropsychology, 12,* 529–538.

Matarazzo, J. D. (1972). *Wechsler's measurement and appraisal of adult intelligence* (5th ed.). Baltimore, MD: Williams and Wilkins.

McFie, J. (1975). *Assessment of organic intellectual impairment.* London: Academic Press.

Meyers, J. E., & Meyers, K. R. (1995). *Rey Complex Figure Test and recognition trial. Professional Manual.* Odessa, FL: Psychological Assessment Resources.

Mickanin, J., Grossman, M., Onishi, K., Auriacombe, S., & Clark C. (1994). Verbal and non-verbal fluency in patients with probable Alzheimer's disease. *Neuropsychology, 8,* 385–394.

Milberg, W. P., Hebben, N., & Kaplan, E. (1996). The Boston process approach to neuropsychological assessment. In I. Grant & K. M. Adams (Eds.), *Neuropsychological assessment of neuropsychiatric disorders* (2nd ed., pp. 58–80). New York: Oxford.

Millis, S. R. (1992). The Recognition Memory Test in the detection of malingered and exaggerated memory deficits. *The Clinical Neuropsychologist, 6,* 406–414.

Millis, S. R., & Ricker, J. H. (1994). Verbal learning patterns in moderate and severe traumatic brain injury. *Journal of Clinical and Experimental Neuropsychology, 16,* 498–507.

Milner, B. (1971). Interhemispheric differences in the localization of psychological processes in man. *British Medical Bulletin, 27,* 272–277.

Mirsky, A. F., Anthony, B. J., Duncan, C. C., Ahearn, M. B., & Kellam, S. G. (1991). Analysis of the elements of attention: A neuropsychological approach. *Neuropsychology Review, 2,* 109–145.

Mishkin, M., Ungerleider, L. G., & Macko, K. A. (1983). Object vision and spatial vision: Two cortical pathways. *Trends in Neurosciences, 6,* 414–417.

Mitrushina, M., Satz, P., & Van Gorp, W. (1989). Some putative cognitive precursors in subjects hypothesized to be at-risk for dementia. *Archives of Clinical Neuropsychology, 4,* 323– 333.

Mittenberg, W., Azrin, R., Millsaps, C., & Heilbronner, R. (1993). Identification of malingered head injury on the Wechsler Memory Scale-Revised. *Psychological Assessment, 1993, 5,* 34–40.

Mittenberg, W., Rotholc, A., Russell, E., & Heilbronner, R. (1996). Identification of malingered head injury on the Halstead-Reitain Battery. *Archives of Clinical Neuropsychology, 11,* 271–281.

Mittenberg, W., Theroux-Fichera, S., Zielinski, R. E., & Heilbronner, R. L. (1995). Identification of malingered head injury on the Wechsler Adult Intelligence Scale-Revised. *Professional Psychology: Research and Practice, 26,* 491–498.

Monsch, A. U., Bondi, M. W., Butters, N., Paulsen, J. S., Salmon, D. P., Brugger, P., & Swenson, M. R. (1994). A comparison of category and letter fluency in Alzheimer's disease and Huntington's disease. *Neuropsychology, 8,* 25–30.

Nelson, H. E. (1982). *National Adult Reading Test (NART) test manual.* Windsor, Great Britain: NFER-Nelson Publishing Company.

Nelson, H. E., & O'Connell, A. (1978). Dementia: The estimation of premorbid intelligence levels using the New Adult Reading Test. *Cortex, 14,* 234–244.

Nelson, L. D., Mitrushina, M., Satz, P., Sowa, M., & Cohen, S. (1993). Cross-validation of the Neuropsychology Behavior and Affect Profile in stroke patients. *Psychological Assessment, 5,* 374–376.

Nelson, L. D., Satz, P., Mitrushina, M., VanGorp, W., Cicchetti, D., Lewis, R., & Van Lancker, D. (1989). Development and validation of the Neuropsychology Behavior and Affect Profile. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 1,* 266–272.

Ogden, J. A. (1986). Neuropsychological and psychological sequelae of shunt surgery in young adults with hydrocephalus. *Journal of Clinical and Experimental Neuropsychology, 8,* 657–679.

Osterrieth, P. A. (1944). Le test de copie d'une figure complex: Contribution a l'etude de la perception et de la memoire. *Archives de Psychologie, 30,* 206–356.

Overall, J. E., & Gorham, D. R. (1962). The brief psychiatric rating scale. *Psychological Reports, 10,* 799–812.

Paolo, A., & Ryan, J. J. (1993). WAIS-R abbreviated forms in the elderly: A comparison of the Satz-Mogel with a seven-subtest short form. *Psychological Assessment, 5,* 425–429.

Paolo, A. M., Ryan, J. J., Troster, A. I., & Hilmer, C. D. (1996). Demographically based regression equations to estimate WAIS-R subtest scales scores, *The Clinical Neuropsychologist, 10,* 130–140.

Parsons, O. A. (1993a). President's message. *Division of Clinical Neuropsychology, Newsletter 40, 11,* No. 1, 1–2.

Parsons, O. A. (1993b). President's message. *Division of Clinical Neuropsychology, Newsletter 40, 11,* No. 3, 1.

Perrine, K. (1993). Differential aspects of conceptual processing in the Category Test and Wisconsin Card Sorting Test. *Journal of Clinical and Experimental Neuropsychology, 15,* 461–473.

Peterson, L. R. & Peterson, M. J. (1959). Short-term retention of individual verbal items. *Journal of Experimental Psychology, 58,* 193–198.

Ponsford, J. L., Donnan, G. A., & Walsh, K. W. (1980). Disorders of memory in vertebrobasilar disease. *Journal of Clinical Neuropsychology, 2,* 267–276.

Porteus, S. D. (1965). *Porteus Maze Test. Fifty years application.* New York: Psychological Corporation.

Powell, J. B., Cripe, L. I., & Dodrill, C. B. (1991). Assessment of brain impairment with the Rey Auditory Verbal Learning Test: A comparison with other neuropsychological measures. *Archives of Clinical Neuropsychology, 6,* 241–249.

Prigatano, G. P., & Schacter, D. L. (1991). *Awareness of deficit after brain injury. Clinical and theoretical issues.* New York: Oxford.

Raven, J. C. (1982). *Revised manual for Raven's Progressive Matrices and Vocabulary Scale.* Windsor, United Kingdom: NFER-Nelson.

Regard, M., Strauss, E., & Knapp, P. (1982). Children's production on verbal and non-verbal fluency tasks. *Perceptual and Motor Skills, 55,* 839–844.

Reitan, R. M., & Wolfson, D. (1993). *The Halstead-Reitan Neuropsychological Test Battery. Theory and clinical interpretation* (3rd ed.). Tucson, AZ: Neuropsychology Press.

Reitan, R. M., & Wolfson, D. (1995). Category Test and Trailmaking Test as measures of frontal lobe functions. *The Clinical Neuropsychologist, 9,* 50–56.

Rey, A. (1964). *L'examen clinique en psychologie.* Paris: Presses Universitaires de France.

Rey, A. (1941). L'examen psychologique dans le cas de'encephalopathie traumatique. *Archives de Psychologie, 28,* 286–340.

Ribot, T. (1881). *Diseases of memory.* New York: Appleton.

Roberts, R. J., Hamsher, K.deS., Bayless, J. D., & Lee, G. P. (1990). Presidents Test performance in varieties of diffuse and unilateral cerebral disease. *Journal of Clinical and Experimental Neuropsychology, 12,* 195–208.

Roman, D. C., Edwall, G. E., Buchanan, R. J., & Patton, J. H. (1991). Extended norms for the Paced Auditory Serial Addition Task. *The Clinical Neuropsychologist, 5,* 33–40.

Rourke, B. P. (Ed.). (1991). *Neuropsychological validation of learning disability subtypes.* New York: Guilford.

Rourke, B. P. (Ed.). (1995). *Syndrome of non-verbal learning disabilities.* New York: Guilford.

Ruff, R. M. (1996). *Ruff Figural Fluency Test.* Odessa, FL: Psychological Assessment Resources.

Ruff, R. M., Light, R. H., & Quayhagen, M. (1989). Selective reminding tests: A normative study of verbal learning in adults. *Journal of Clinical and Experimental Neuropsychology, 11,* 539–550.

Russell, E. W. (1975). A multiple scoring method for the evaluation of complex memory functions. *Journal of Consulting and Clinical Psychology, 43,* 800–809.

Ryan, J. J., Paolo, A. M., & Smith, A. J. (1992). Wechsler Adult Intelligence Scale-Revised intrasubtest scatter in brain-damaged patients: A comparison with the standardization sample. *Psychological Assessment, 4,* 63–66.

Ryan, J. J., Rosenberg, S. J., & Mittenberg, W. (1984). Factor analysis of the Rey Auditory Verbal Learning Test. *The International Journal of Clinical Neuropsychology, 6,* 239–241.

Sass, K. J., Spencer, D. D., Kim, J. H., Westerveld, M., Novelly, R. A., & Lencz, T. (1990). Verbal memory impairment correlates with hippocampal pyramidal cell density. *Neurology, 40,* 1694–1697.

Satz, P., Taylor, H. G., Friel, J., & Fletcher, J. M. (1978). Some developmental and predictive precursors of reading disabilities: A six year follow-

up. In A. L. Benton & D. Pearl (Eds.), *Dyslexia* (pp. 313–347). New York: Oxford.

Schmidt, M., Trueblood, W., Merwin, M., & Durham, R. L. (1994). How much do attention tests tell us? *Archives of Clinical Neuropsychology, 9,* 383–394.

Schretlen, D., Benedict, R. H. B., & Bobholz, J. H. (1994). Composite reliability and standard errors of measurement for a seven subtest short form of the Wechsler-Adult Intelligence Scale-Revised. *Psychological Assessment, 6,* 188–190.

Schwartz, B. S., Ford, D. P., Bolla, K. I., Agnew, J., Rothman, N., & Bleecker, W. L. (1990). Solvent-associated decrements in olfactory function in paint manufacturing workers. *American Journal of Industrial Medicine, 18,* 697–706.

Semmes, J. (1965). A non-tactual factor in stereognosis. *Neuropsychologia, 3,* 295–315.

Sherer, M., Scott, J. G., Parsons, O. A., & Adams, R. L. (1994). Relative sensitivity of the WAIS-R subtests and selected HRNB measures to the effects of brain damage. *Archives of Clinical Neuropsychology, 9,* 427–436.

Sherman, E. M. S., Strauss, E., Spellacy, F., & Hunter, M. (1995). Construct validity of WAIS-R factors: Neuropsychological test correlates in adults referred for evaluation of possible head injury. *Psychological Assessment, 7,* 440–444.

Shum, D. H. K., McFarland, K. A., & Bain, J. D. (1990). Construct validity of eight tests of attention: Comparison of normal and closed head injured samples. *The Clinical Neuropsychologist, 4,* 151–162.

Sivan, A. B. (1992). *Benton Visual Retention Test* (5th ed). San Antonio, TX: Psychological Corporation.

Ska, B., Poissant, A., & Joanette, Y. (1990). Line orientation judgment in normal elderly and subjects with dementia of Alzheimer's type. *Journal of Clinical and Experimental Neuropsychology, 12,* 695–702.

Smith, G. E., Ivnik, R. J., Malec, J. F., Kokmen, E., Tangalos, E. G., & Kurland, L. T. (1992). Mayo's older Americans normative studies (MOANS): Factor structure of a core battery. *Psychological Assessment, 4,* 382–390.

Smith, G. E., Ivnik, R. J., Malec, J. F., Petersen, R. C., Kokmen, E., & Tangalos, E. G. (1994). Mayo cognitive factor scores: Derivation of a short battery and norms for factor scores. *Neuropsychology, 8,* 194–202.

Smith, G. E., Ivnik, R. J., Malec, J. F., & Tangalos, E. G. (1993). Factor structure of the Mayo older Americans normative sample (MOANS) core battery: Replication in a clinical sample. Psychological Assessment, 5, 121–124.

Spreen, O., & Strauss, E. (1998). *A compendium of neuropsychological tests. Administration, norms, and commentary* (2nd ed.). New York, NY: Oxford.

Squire, L. R., & Slater, P. C. (1975). Forgetting in very long-term memory as assessed by an improved questionnaire technique. *Journal of Experimental Psychology: Human Learning and Memory, 104,* 50–54.

Strub, R. L., & Black, F. W. (1985). *The mental status examination in neurology* (2nd ed.). Philadelphia: F.A. Davis.

Stuss, D. T., Kaplan, E. F., Benson, D. F., Weir, W. S., Chirilli, S., & Sarazin, F. F. (1982). Evidence for the involvement of orbitofrontal cortex in memory functions: An interference effect. *Journal of Comparative and Physiological Psychology, 96,* 913–925.

Stuss, D. T., Stethem, L. L., Hugenholtz, H., & Richard, M. T. (1989). Traumatic brain injury: A comparison of three clinical tests, and analysis of recovery. *The Clinical Neuropsychologist, 3,* 145–146.

Stuss, D. T., Stethem, L. L., & Poirer, C. A. (1987). Comparison of three tests of attention and rapid information processing across six age groups. *The Clinical Neuropsychologist, 1,* 139–152.

Sweet, J. J., Westergaard, C. K., & Moberg, P. J. (1995). Managed care experiences of clinical neuropsychologists. *The Clinical Neuropsychologist, 9,* 214–218.

Tarter, R. E., & Edwards, K. L. (1986). Neuropsychological batteries. In T. Incagnoli, G. Goldstein, & C. J. Golden (Eds.), *Clinical application of neuropsychological test batteries* (pp. 135–153). New York: Plenum.

Thompson, L. L., Heaton, R. K., Matthews, C. G., & Grant, I. (1987). Comparison of preferred and nonpreferred hand performance on four neuropsychological motor tasks. *The Clinical Neuropsychologist, 1,* 324–334.

Trahan, D. E. (1992). Analysis of learning and rate of forgetting in age-associated memory differences. *The Clinical Neuropsychologist, 6,* 241–246.

Trahan, D. E., & Larrabee, G. J. (1988). *Professional manual: Continuous Visual Memory Test.* Odessa, FL: Psychological Assessment Resources.

Trahan, D. E., & Larrabee, G. J. (1992). Effect of normal aging on rate of forgetting. *Neuropsychology, 6,* 115–122.

Trahan, D. E., & Larrabee, G. J. (1993). Clinical and methodological issues in measuring rate of forgetting with the verbal selective reminding test. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 5,* 67–71.

Trahan, D. E., Larrabee, G. J., Fritzsche, B., & Curtiss, G. (1996). Continuous Visual Memory Test: Alternate form and generalizability estimates. *The Clinical Neuropsychologist, 10,* 73–79.

Trahan, D. E., Larrabee, G. J., & Levin, H. S. (1986). Age-related differences in recognition memory for pictures. *Experimental Aging Research, 12,* 147–150.

Trahan, D. E., Larrabee, G. J., & Quintana, J. W. (1990). Visual recognition memory in normal adults and patients with unilateral vascular lesions. *Journal of Clinical and Experimental Neuropsychology, 12,* 857–872.

Trahan, D. E., Larrabee, G. J., Quintana, J. E., Goethe, K. E., & Willingham, A. C. (1989). Development and clinical validation of an expanded paired associate test with delayed recall. *The Clinical Neuropsychologist, 3,* 169–183.

Trahan, D. E., Quintana, J., Willingham, A. C., & Goethe, K. E. (1988). The Visual Reproduction Subtest: Standardization and clinical validation of a delayed recall procedure. *Neuropsychology, 2,* 29–39.

Trenerry, M. R., Crosson, B., Deboe, J., & Leber, W. R. (1989). *The Stroop Neuropsychological Screening Test.* Odessa, FL: Psychological Assessment Resources.

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving (Ed.), *Organization of memory* (pp. 381–403). New York: Academic Press.

Varney, N. R. (1988). Prognostic significance of anosmia in patients with closed-head trauma. *Journal of Clinical and Experimental Neuropsychology, 10,* 250–254.

Walsh, K. W. (1987). *Neuropsychology. A clinical approach* (2nd ed.). Edinburgh, Scotland: Churchill Livingstone.

Walsh, K. W. (1995). A hypothesis-testing approach to assessment. In R. L. Mapou & J. Spector (Eds.), *Clinical neuropsychological assessment. A cognitive approach* (pp. 269–291). New York: Plenum.

Ward, L. C. (1990). Prediction of verbal, performance, and Full Scale IQs from seven subtests of the WAIS-R. *Journal of Clinical Psychology, 46,* 436–440.

Warrington, E. K., (1984). *Recognition Memory Test.* Windsor, United Kingdom: NFER-Nelson.

Wechsler, D. (1945). A standardized memory scale for clinical use. *Journal of Psychology, 19,* 87–95.

Wechsler, D. (1981). *WAIS-R manual.* New York: The Psychological Corporation.

Wechsler, D. (1987). *Wechsler Memory Scale-Revised manual.* San Antonio, TX: The Psychological Corporation.

Westerveld, M., Sass, K. J., Sass, A., & Henry, H. G. (1994). Assessment of verbal memory in temporal lobe epilepsy using the Selective Reminding Test: Equivalence and reliability of alternate forms. *Journal of Epilepsy, 7,* 57–63.

White, R. F. (1987). Differential diagnosis of probable Alzheimer's disease and solvent encephalopathy in older workers. *The Clinical Neuropsychologist, 1,* 153–160.

Wiens, A. N., Bryant, J. E., & Crossen, J. R. (1993). Estimating WAIS-R FIQ from the National Adult Reading Test-Revised in normal subjects. *The Clinical Neuropsychologist, 7,* 70–84.

Wiens, A. N., McMinn, M. R., & Crossen, J. R. (1988). Rey Auditory Verbal Learning Test: Development of norms for healthy young adults. *The Clinical Neuropsychologist, 2,* 67–87.

Wiens, A. N., Tindall, A. G., & Crossen, J. R. (1994). California Verbal Learning Test: A Normative study. *The Clinical Neuropsychologist, 8,* 75–90.

Wilkinson, G. S. (1993). *WRAT-3. Wide Range Achievment Test. Administration manual.* Wilmington, DE: Wide Range Inc.

Williams, J. M. (1987). *Cognitive Behavior Rating Scales. Research edition Manual.* Odessa, FL: Psychological Assessment Resources, Inc.

Williams, J. M. (1991). *Memory Assessment Scales.* Odessa, FL: Psychological Assessment Resources.

Williams, J. M., Little, M. M., Scates, S., & Blockman, N. (1987). Memory complaints and abilities among depressed older adults. *Journal of Consulting and Clinical Psychology, 55,* 595–598.

Wilson, R. S., Kaszniak, A. W., & Fox, J. H. (1981). Remote memory in senile dementia. *Cortex, 17,* 41–48.

Wilson, R. S., Rosenbaum, G., Brown, G., Rourke, D., Whitman, D., & Grisell, J. (1978). An index of premorbid intelligence. *Journal of Consulting and Clinical Psychology, 46,* 1554–1555.

Woodcock, R. W., & Mather, N. (1989). *Woodcock-Johnson Psycho-Educational Battery-Revised.* Allen, TX: DLM Teaching Resources.

Yeudall, L. T., Fromm, D., Reddon, J. R., & Stefanyk, W. O. (1986). Normative data stratified by age and sex for 12 neuropsychological tests. *Journal of Clinical Psychology, 42,* 918–946.

This Page Intentionally Left Blank

PART VI

INTERVIEWING

This Page Intentionally Left Blank

CHAPTER 13

CONTEMPORARY CLINICAL INTERVIEWING: INTEGRATION OF THE DSM-IV, MANAGED CARE CONCERNS, MENTAL STATUS, AND RESEARCH

Shawn Christopher Shea

INTRODUCTION

Interviewing is the backbone of all mental health professions. It is a dynamic and creative process, which represents a somewhat elusive set of skills. The importance of this set of skills has been highlighted by Langsley and Hollender (1982). Their survey of 482 psychiatric teachers and practitioners revealed that 99.4 percent ranked conducting a comprehensive interview as an important requirement for a psychiatrist. This represented the highest ranking of 32 skills listed in the survey. Seven of the top 10 skills were directly related to interviewing technique, including skills such as the assessment of suicide and homicide potential, the ability to make accurate diagnoses, and the ability to recognize countertransference problems and other personal idiosyncrasies as they influence interactions with patients. These results were replicated in a follow-up survey (Langsley & Yager, 1988).

It can be seen from this list that the contemporary clinician is being asked to combine an impressive list of complex skills, ranging from structuring techniques and diagnostic explorations using the DSM-IV, to more classic psychodynamic approaches and engagement skills. This clinical challenge has been made even more difficult by yet another new influence, the powerful presence of managed care and the constant ticking of "the clock" concerning the number of sessions available to the client. In the past a skilled clinican could perform a sound diagnostic assessment within an hour, although many chose to take longer. The difference is that today the clinician does not have a choice; managed-care principles dictate that he or she must complete the assessment within an hour and subsequently rapidly write up the document as well.

Such a daunting integrative task, performed under tight time constraints, can represent a major hurdle for the developing clinician. This educational expectation was somewhat wryly stated by Sullivan (1970) decades ago when he wrote: "The psychiatric expert is presumed, from the cultural definition of an expert, and from the general rumors and beliefs about psychiatry, to be quite able to handle a psychiatric interview." But the ability to handle the initial assessment interview has become a considerably more complicated task

since the time of Sullivan's quote, for there has been an evolution in psychiatry and mental-health care of immense proportions in the past 40 years.

This chapter is about this ongoing evolution and its impact on assessment interviewing. Perhaps the single most striking legacy of the evolution is the disappearance of *the* psychiatric interview. Instead of a single style of interviewing, the contemporary clinician must learn to perform an impressive array of interviews suited to the specific clinical task at hand, including assessments as diverse as those required in an emergency room; an inpatient unit; a psychotherapy practice, consultation and liaison setting; and a managed-care clinic.

This chapter is designed for both academicians interested in the theoretical and research underpinnings of the interview process and clinical students concerned with practical interviewing techniques. It makes no attempt to be an exhaustive overview; instead, the reader is provided with a conceptual guide that provides a wealth of references for more in-depth study.

The following areas are discussed: (a) an historical overview and description of the influences that have shaped the evolution of clinical interviewing mentioned earlier; (b) a practical introduction to two of the major clinical cornerstones of current assessment interviewing: the mental status examination and the DSM-IV; and (c) a review of some of the major research efforts with regard to interviewing, including clinician phrasing of responses, nonverbal concerns, alliance issues and empathy, structured interviews, and educational research.

Before proceeding it will be of use to define a few terms that clarify many of the complicated issues regarding interviewing style. The style of any specific clinical or research interview is greatly determined by the following structural factors: (a) specific content areas required to make a clinical decision or to satisfy a research data base, (b) quantity of data required, (c) importance placed on acquiring valid historical and symptomatic data as opposed to patient opinion and psychodynamic understanding, and (d) time constraints placed upon the interviewer.

With regard to these structural concerns of the interview, two concepts outlined by Richardson, Dohrenwend, and Klein (1965) are useful: *standardization* and *scheduling*. Standardization refers to the extent to which informational areas or items to be explored are specified in the interview procedure. Scheduling refers to the prespecification of the wording and sequence of the interview process.

By utilizing these two concepts, several interview types can be defined. In the free-format interview, the interviewer has little standardization of database and is highly interested in the spontaneous content produced by the patient. Such free-format interviews place little emphasis on scheduling and tend to follow the natural wanderings of the patient. These interviews are valuable for uncovering patient psychodynamics and revealing patient feelings, opinions, and defenses.

At the opposite end of the spectrum is the fully structured interview that is highly standardized and strictly scheduled. In fully structured interviews the required informational areas are specified in detail and the ways of exploring them are also prescribed. An example of this type of interview is the Diagnostic Interview Schedule (Robins, Helzer, Croughan, & Ratcliff, 1981), developed for community surveys by lay interviewers.

Semistructured interviews represent procedures in which the informational areas to be explored are specified, but the sequence and wording to be used in data gathering are only moderately predetermined. In these interviews, general guidelines about the interview sequence, such as beginning with the chief complaint and following with episodes of the present illness, may be provided, but the clinician is given some latitude to move within this framework. Semistructured interviews are of value in both research and clinical settings. They frequently can provide standardized databases as pioneered by Mezzich in the Initial Evaluation Form (Mezzich, Dow, Rich, Costello, & Himmelhoch, 1981; Mezzich, Dow, Ganguli, Munetz, & Zettler-Segal, 1986).

The last major type of interview is the flexibly structured interview. The flexibly structured interview represents the most popular clinical interview, and when performed by an experienced clinician, holds promise as a research tool. With the flexibly structured interview, the clinician has a standardized database (pre-determined by the clinical or research task at hand) but is given total freedom in scheduling. The interview begins with a free-format style in which the clinician moves with whatever topics appear to be most pressing for the patient. Once the engagement is secured the clinician begins to structure the interview sensitively.

With flexibly structured interviews the actual scheduling will be relatively unique to each clinician-patient dyad, for the interviewer fluidly alters the style of scheduling to gather the standardized database most effectively while working with the

specific needs and defenses of the patient. These interviews require a high degree of sophistication from the clinician and allow him or her to insert areas of free format and dynamic questioning whenever expedient. Most experienced clinicians, whether consciously or by habit, utilize a flexibly structured format. The complexities and nuances of the flexibly structured clinical interview have been most recently explored in detail by Shea (1998) and Othmer and Othmer (1994). A comprehensive annotated bibliography on the literature concerning clinical interviewing and training appears in *Core Readings of Psychiatry* (Shea, 1995).

Historically, clinical interview styles have varied in popularity; they have ranged from semistructured interviews that were partially based on the medical model to more free-form analytic interviews and flexibly structured styles. It is to this evolution that attention is now turned.

HISTORICAL FOUNDATIONS

When studying the historical evolution of the interview, it is helpful to look for underlying principles of development. Perhaps the most useful principle is that interview styles tend to evolve out of whatever theoretical knowledge base is most popular in a given age. In particular, the more numerous and syndrome-specific the available treatment modalities are, the more likely it is that a standardized database will be sought. If the standardized database requirements become large, there is a gradual shifting toward methods of structuring, whether done by rigid schedule or by flexible maneuvering. This relationship between the availability of treatment modality and interview style is seldom noted but represents a powerful and unifying historical principle.

Early in the century the approach to clinical assessment was rooted in the medical model. Kraepelin had attempted to classify mental illnesses and indeed had been able to differentiate manic depression from dementia praecox (Kaplan, Freedman, & Sadock, 1980). Although there was not an abundance of treatment modalities present, the gestalt of the moment was toward a careful detailing of behaviors and symptoms in an effort to determine specific syndromes and diseases.

At this time the gifted psychiatrist and educator Adolf Meyer proved to be a catalyst in the development of the psychiatric interview. Paradoxically, his interests would move forward both the free-format style and a more semistructured approach. Meyer professed a psychobiological approach to the patient, in which it was deemed important to determine a "biography" of the patient that included biological, historical, psychological, and social influences on the patient's current behavior (Kaplan, Freedman, & Sadock, 1980). His interest in psychological and social influences further advanced a style of interviewing in which there was an appreciation for the value of the free-format style (Siassi, 1984).

On the other hand, Meyer's interest in determining a sharp conceptualization of biological influences as well as a clear presentation of the patient's immediate symptomatology moved him toward an appreciation of semistructured or flexibly structured formats. For instance, Meyer believed that the clinician should begin the interview with a careful exploration of the patient's chief complaint (Kaplan, Freedman, & Sadock, 1980). In his work "Outlines of Examinations" (Meyer, 1951), which was printed privately in 1918, Meyer was the first to define the term "mental status" (Donnelly, Rosenberg, & Fleeson, 1970).

By the end of the first quarter of the century many of the major components of the psychiatric interview had been established. These key content regions included chief complaint, history of the present illness, social history, family history, medical history, and mental status. All of these were related to an underlying attempt to arrive at a diagnostic overview. But a diagnostic system based on mutually agreed-upon criteria was not well established, and consequently, most of the interview was not directed primarily toward establishing a specific diagnosis.

Such lack of diagnostic specificity, coupled with a relative paucity of treatment interventions, resulted in a database that did not require a high degree of scheduling. In the first place, because there were few diagnostic-related interventions, there was not a pressing need to complete the initial assessment quickly. The clinician could spend many hours over many days eliciting data for the initial interview. In the second place, the diagnostic schema were so limited that there was not a significant need to cover large areas of symptomatology quickly. The resulting relative lack of scheduling and structure was to have a major thrust toward even more emphasis on free format.

Psychoanalysis arrived on the shores of America like a native-born son. By the 1940s it had become

well established. Freud's pioneering work had an enormous impact on interviewing technique. His basic theories seemed to move away from emphasis on diagnosis in a medical sense toward a more probing investigation of actual psychological processes. With the development of ego psychology and a further investigation of defense mechanisms by theorists such as Heinz Hartman and Anna Freud, the emphasis further shifted toward an understanding of how the patient's defenses were manifested in the context of the interview itself. Interviewing and therapy seemed to become less distinct.

A free-format style of interviewing became more common. Clinicians became increasingly aware of the value of spontaneous speech as a fertile ground for uncovering patient defenses and conflicts. The elicitation and description of these defenses and a basic description of the patient's ego structure became goals of the interview. Important advances in interviewing technique evolved during this time. Emphasis was placed not only on what the patient said but what the patient either consciously or unconsciously did not say. Resistance came to be seen as a golden door for entering the dynamics and conflicts of the patient. A free-format style of interview provided a rich psychological milieu in which to observe directly the maneuverings of the patient's unconscious defenses.

Several books helped clinicians to adapt to this new emphasis in interviewing style. One was *Listening With the Third Ear* by Theodore Reik (1952). In the section entitled "The Workshop," Reik provides a variety of insights concerning issues such as freefloating attention, conscious and unconscious observation, and the therapist-patient alliance.

Another important analytic contribution was *The Clinical Interview* (Vols. I & 2) written by Felix Deutsch and William Murphy (1955a, 1955b). Working out of Boston, Deutsch and Murphy described the technique of *associative anamnesis.* This technique emphasizes a free-format style in which free association and gentle probing by the clinician open a window into the symbolic world that lies "between the lines" of the patient's report.

But perhaps the most influential book dedicated to interviewing from an analytic point is the classic text, *The Psychiatric Interview in Clinical Practice,* by MacKinnon and Michels (1971). This book provided an easily read yet highly rewarding introduction to understanding dynamic principles as they revealed themselves in the initial interview and subsequent therapy. Few, if any, books describe more lucidly and insightfully the subtle relationships between patient defense mechanisms and clinician style.

In the early 1950s another major force was to have an impact on the psychiatric interview. That force would be a single man: Harry Stack Sullivan. During his life, Sullivan proved to be one of the most gifted interviewers of all time. His book *The Psychiatric Interview* was published posthumously in 1954 (Sullivan, 1970). The book would establish forever the importance of the interpersonal matrix as one of the major areas through which to understand the interview process. Sullivan stressed the importance of viewing the interview as a sociological phenomenon in which the patient and the clinician form a unique and dynamic dyad, with the behavior of each affecting the other.

One of Sullivan's key terms was *participant observation.* This concept emphasized the need of the interviewer to "step aside" during the interview itself in the sense of viewing his or her own behavior and the impact of that behavior upon the patient. Sullivan saw that the measuring instrument itself, in this case the interviewer, could actually change the database, that is, the patient's behaviors and degree of distortion in relating symptomatology.

Sullivan was also one of the first interviewers to emphasize the importance of structuring, and he discussed specific methods of making transitions during the interview from one topic to another. In this sense Sullivan recognized the importance of free-format style as well as scheduling issues, and essentially developed a flexibly structured style of interviewing in which these various techniques could be intermixed at the will of the clinician.

Near the time of Sullivan's book, another work appeared entitled *The Initial Interview in Psychiatric Practice* by Gill, Newman, and Redlich (1954). This work was strongly influenced by Sullivan's interpersonal perspective. Innovatively, the book includes three fully annotated transcripts of interviews which were accompanied by phonographic records of the actual patient/physician dialogue. It also contains an excellent history of interviewing technique.

With regard to the interpersonal perspective, Sullivan's classic text had been predated by J. C. Whitehorn. In 1944 Whitehorn published an influential article in the *Archives of Neurology and Psychiatry* entitled, "Guide to Interviewing and

Clinical Personality Study." One of Whitehorn's contributions lay in his emphasis on eliciting patient opinion as both a powerful engagement technique, and a method of looking at unconscious dynamics. In particular, the patient's opinions concerning interpersonal relations and reasons for caring for others represented major avenues for exploration.

Closely related to the analytic and interpersonal schools was the European-based school of phenomenological psychiatry and psychology. Giants in the field during the first half of this century, such as Karl Jaspers and Medard Boss, emphasized an approach to the patient in which the focus was on developing an understanding of the exact ways in which the patient experienced "being in the world" (Hall & Lindzey, 1978). In this approach, while utilizing a phenomenological style of interview, the clinician delicately probes the patient for careful descriptions of the patient's symptoms, feelings, perceptions, and opinions. Through a shared process of precise questioning and at times, self-transparency, the clinician arrives at a vivid picture of the patient's universe, a picture which sometimes even surprises the patient as defenses and distortions are worked through by the clinician's style of questioning.

In more recent years, authors such as Alfred Margulies and Leston Havens have reemphasized the importance of a phenomenological approach (Havens, 1978, 1979; Margulies, 1984; Margulies & Havens, 1981). A particularly fascinating technique, known as counterprojection, has been described by Havens. The counterprojective technique deflects paranoid projections before they manifest onto the interviewer. Such techniques are valuable in consolidating an alliance with frightened, hostile, angry, or actively paranoid patients (Havens, 1980).

In summary, it can be seen that during the middle years of this century and later, psychiatrists from the analytic, interpersonal, and phenomenological schools exerted a strong influence on interviewing technique. The next impact would come from a nonmedical tradition.

In the 1950s, 1960s, 1970s, and 1980s the fields of psychology and counseling had an enormous impact on clinical interviewing. More than psychiatry, these fields emphasized the need for empirical research concerning the flexibly structured clinical interview, which will be discussed in more detail later. These research approaches opened up an increased awareness of the specific factors, both verbal and nonverbal, which allow the clinician to relate favorably to the interviewee.

Carl Rogers represents one of the most powerful influences in this regard. His "client-centered approach" emphasized empathic techniques. He described empathy as the clinician's ability "to perceive the internal frame of reference of another with accuracy, and with the emotional components and meanings which pertain thereto, as if one were the other person, but without ever losing the 'as if' condition (Rogers, 1951, 1959)."

Rogers is also well known for his concept of "unconditional positive regard." A clinician conveys this value when he or she listens without passing judgment on the patient's behaviors, thoughts, or feelings. These ideas were pivotal in conveying the idea that the clinician should not appear remote or distant during the interviewing process, for such artificial remoteness could seriously disengage patients. Interviewers were allowed to utilize in a naturalistic sense their social skills and personality.

Other counselors, such as Alfred Benjamin (1969) in *The Helping Interview* and Gerard Egan (1975) in *The Skilled Helper,* carried on this tradition of emphasizing genuineness and common sense in the therapeutic relationship. Benjamin emphasized the need to develop a trusting relationship with the patient, avoiding the tendency to hide behind rules, position, or sense of authority. Egan attempted to help clinicians develop these abilities by describing a concrete language with which to help convey these ideas in an educational sense, highlighted by a self-programmed manual to accompany his text. Most of the interviewing techniques developed by these authors and other counselors are distinctly non-diagnostically focused. Consequently, as one would expect, they tend to be neither highly standardized nor scheduled.

But the fields of counseling and psychology did not ignore the importance of the database. To the contrary, an emphasis on developing an increasingly sophisticated understanding of the impact of interviewing technique on the validity of data, grew out of the empirical studies and behavioral approaches pioneered by nonmedical researchers.

For example, Richardson, Dohrenwend, and Klein (1965) whose schema of standardization and scheduling was mentioned earlier, produced a particularly incisive work entitled *Interviewing: Its Forms and Functions.* The power of the text lies in the authors' attempts systematically to define and study various characteristics of the interview process ranging from the style of questioning (such as

open-ended versus closed-ended) to the impact of patient and clinician characteristics on the interviewing process. Their work transformed a process that was heretofore somewhat nebulous in nature into a process that could be studied behaviorally.

Another psychologist, Gerald Pascal (1983), described a technique known as the *behavioral incident*. Although simple in nature, this technique represents one of the most significant and easily taught of all interviewing techniques in the last several decades. The technique is based on the premise that questions can range on a continuum from those that request patient opinion to those that ask for historical or behavioral description. The latter style of questioning is more apt to yield valid information, whereas questions which request patient opinion are dangerously prone to patient distortion.

The behavioral incident provides a more reliable tool for exploring areas of particular sensitivity where patient distortion may be high, as in the assessment of suicide potential, child abuse, substance abuse, and antisocial behavior. For example, a clinician may phrase a lethality probe in this fashion, "Have you ever had any serious suicide attempts?" It is then up to the patient to interpret the notion of what constitutes a "serious" attempt. To the patient, an overdose of 20 pills may not seem serious and consequently may not be reported. Using the behavioral-incident technique the clinician asks a series of questions focused directly on patient behaviors: "Tell me exactly what methods of killing yourself you have ever tried, even if only in a small way," and "When you took the pills how many did you take?" With these types of questions the patient is asked to provide concrete information. It is up to the clinician to arrive at an opinion as to what is "serious".

Another significant book concerning the specific phrasing of questions with regard to their impact on data gathering was *The Structure of Magic I* by Grinder and Bandler (1975). Although some of their latter work has been controversial in nature, this early volume was sound, penetrating, and to the point. They described a variety of techniques for phrasing questions in such a manner that the patient's hidden thoughts would be gradually pulled to the surface. The work is based on an understanding of transformational grammar and is enhanced by the self-programmed layout of the book, which literally forces the reader to make actual changes in style of questioning.

Ironically, the most powerful forces to operate on interviewing style in the last several decades were not directly related to attempts to advance interviewing per se but were related to the diagnostic and therapeutic advances occurring within psychiatry proper and the other disciplines of mental-health care. The evolution of the psychiatric interview was the direct result of a revolution in three areas: (1) treatment modalities, (2) diagnostic systems, and (3) managed-care principles.

Concerning the first factor, in the past 30 years an impressive array of new therapeutic interventions has emerged. These revolutionary advances include modalities such as tricyclic antidepressants, serotonin-selective reuptake inhibitors, mood stabilizers, antipsychotic medications, cognitive behavioral therapy, family therapy, group therapy, and more sophisticated forms of dynamic and behavioral therapies such as interpersonal-dynamic psychotherapy and dialectical behavioral therapy, to name only a few. In the same fashion that the rapid acceptance of analytic thinking resulted in further development of the free-format style of interviewing, these new interventions, which frequently are chosen in relation to a DSM-IV diagnosis, have led interviewers to reexamine the importance of developing both a thorough and valid data base.

The development of new treatment interventions directly spawned the second major force molding the contemporary interview. Researchers and clinicians quickly realized that better diagnostic systems, which would decrease variability and unreliability, needed to be developed, for treatment modalities were being increasingly determined by diagnosis.

One of the most influential of the modern diagnostic systems that resulted was the Feighner criteria (Feighner et al., 1972). These criteria were developed in the Department of Psychiatry at Washington University in St. Louis. This system delineated 15 diagnostic categories by using both exclusion and inclusion criteria. Building on this base, Spitzer, Endicott, and Robins (1978) developed the Research Diagnostic Criteria (RDC). With the RDC system the psychopathological range was increased to include 23 disorders.

Of particular note to the history of interviewing was the subsequent development of a semistructured interview designed to delineate the diagnoses described by the Research Diagnostic Criteria. This interview, the Schedule for Affective Disorders and Schizophrenia (SADS), was developed by

Endicott and Spitzer (1978). It was a powerful tool with good reliability and it became popular as a research instrument. A second interview that was both highly standardized and rigidly scheduled was the Diagnostic Interview Schedule (DIS) developed by Robins (Robins, Helzer, Croughan, & Ratcliff, 1981). This interview was designed to be used by lay interviewers and hence was highly scheduled to ensure interrater reliability.

The semistructured and structured formats displayed by the SADS and the DIS, respectively, were not overly popular with clinicians. Such lack of enthusiasm demonstrated that even though clinicians were progressively required to obtain a highly standardized database, the method to achieve this goal while flexibly engaging the patient and handling resistance was not clear.

The movement toward the need for a highly standardized database with regard to diagnostic information was given further momentum in the United States by the publication of the third edition of the *Diagnostic and Statistical Manual of Mental Disorders* (DSM-III) by the American Psychiatric Association (1980). This manual emphasized a multi-axial approach which will be described in more detail later in this chapter. Seven years later the revised edition, the DSM-III-R, appeared (APA, 1987) and was followed by the DSM-IV (APA,1994). With the advent of these widely accepted diagnostic systems, interviewers were faced with the necessity of gathering sensitively the data that would be required for a sophisticated differential diagnosis. This would prove to be no easy task.

From the arena of semi-structured interviews, this task was approached through the development of ever improving and more "user friendly" formats such as the Structured Clinical Interview for the DSM-III (the SCID-III, SCID-III-R, and the SCID-IV), developed by Spitzer and Williams (1983). The SCID-IV has seen wide use, but has some limitations due to its length and its restriction to the DSM-IV system. The Mini-International Neuropsychiatric Interview (M.I.N.I.) pioneered in the United States by Sheehan and in France by Lecrubier (1999) has seen wider international acceptance. The M.I.N.I. is elegant, practical and remarkably brief to administer with a median duration of 15 minutes—all qualities that have enhanced its acceptance by clinicians. But from a practical front-line clinician's standpoint, who must arrive at much more than merely a diagnosis in 60 minutes, the task of integrating a sound DSM-IV differential into an equally sound biopsychosocial evaluation remained daunting.

The task would prove to be further complicated by the advent of a philosophical/economic paradigm shift, representing the third significant factor molding the contemporary psychiatric interview. Managed care, an approach that gained enormous power in the early 1990s, placed a heavy emphasis upon efficient use of resources. In its healthy functioning it pushes clinicians to always work with a sound cost-mindfulness. In its unhealthy functioning it results in inadequate treatment, sometimes caused by clinicians missing critical treatable diagnoses in the initial session, in an effort to prematurely begin therapy and cut corners. For the most part, clinicians no longer have an option; the initial diagnostic session must be completed in sixty minutes with roughly another 30 minutes allotted, if the clinician is lucky, for the written document.

In order to determine a correct DSM-IV diagnosis, perform a sound biopsychosocial assessment, and spot the client's strengths to capitalize upon in brief therapy formats, the contemporary interviewer must gather an amount of concrete information in 60 minutes which might have seemed quite unmanageable to an interviewer of 40 years ago. Consequently, interviewers have reexamined their approaches to scheduling, moving toward partially scheduled interviews, as seen in the semistructured format, or toward a method of tracking the database while maximizing interviewer spontaneity, as seen in the flexibly structured format.

The lead toward resolving some of the complex integrative tasks facing the contemporary psychiatric clinician came from the Western Psychiatric Institute and Clinic at the University of Pittsburgh. In 1985, Hersen and Turner edited an innovative book entitled *Diagnostic Interviewing*. This book represented one of the first attempts to acknowledge fully that interviewers should become familiar with specific techniques for sensitively exploring the diagnostic criteria from the various diagnoses in the DSM-III. To accomplish this educational task, various experts contributed chapters on a wide range of DSM-III categories from schizophrenia to sexual disorders.

At the same time, also at the Western Psychiatric Institute and Clinic, a variety of innovations, both with regard to interviewing technique and interviewing training, were presented in *Psychiatric Interviewing: The Art of Understanding* (Shea, 1988). This book represented the first attempt to synthesize smoothly the divergent streams of inter-

viewing knowledge developed in the various mental-health fields over the previous 50 years. In particular, it acknowledged the confusing task facing contemporary clinicians who had to synthesize a wide range of important information (including regions such as the chief complaint, history of the present illness, the social history, the family history, the medical history, the mental status, and the DSM-III-R diagnostic regions) into an interview that was naturally flowing and to which appropriate energy could be given to dynamic issues and resistance concerns.

To accomplish this task, Shea developed a supervision language that would provide a readily understandable system with which to study the structuring, flow, and time management of the interview. This widely accepted supervision system would later prove to be of immediate value in graduate training across disciplines and in the development of quality assurance programs in the fast-paced world of the managed-care asssessment. This study of the structure and flow of the interview process was named *facilics* (Shea, 1988) derived from the Latin root *facilis* (ease of movement). The study of facilics emphasized a rigorous examination of the overall structuring of the interview as it related to time constraints and clinical tasks. To enhance learning further, a schematic supervision system was designed, made up of symbols which depict the various transitions and types of topical expansions utilized by the trainee. This shorthand system clarifies educational concepts and highlights structural elements, while presenting an immediately understandable and permanent record of what took place in the interview.

In subsequent years facilic principles were applied to specific interviewing tasks such as the elicitation of suicidal ideation (Shea, 1998). This resulted in the development of innovative interviewing strategies such as the Chronological Assessment of Suicide Events (the CASE Approach) and its counterpart for uncovering violent ideation the Chronological Assessment of Dangerous Events (the CADE Approach). The CASE Approach, delineated in *The Practical Art of Suicidal Assessment* (Shea, 1999) was a flexible, practical, and easily learned inteview strategy for eliciting suicidal ideation, planning, and intent. It was designed to increase validity and decrease potentially dangerous errors of omission. Because the techniques of the CASE Approach were behaviorally concrete it could be readily taught and the skill level of the clinician easily documented for quality assurance purposes.

While these advances were being made, an outstanding and highly influential textbook, *The Clinical Interview Using DSM-III-R* was written by Othmer and Othmer (1989) which, more than any other textbook on interviewing, firmly secured the DSM-III-R system as an accepted clinical tool by front-line clinicians. This text, filled with practical tips and model questions, has secured itself as a classic. It was later expanded to two volumes to further address interview situations involving complicated clinical interactions with borderline and resistant patients (Othmer and Othmer, 1994). Other outstanding inteviewing texts soon followed including Morrison's *The First Interview: A Guide for Clinicians* (1993), Shea's *Psychiatric Interviewing: The Art of Understanding,* 2nd edition (1998), Sommers-Flanagan's *Clinical Interviewing,* 2nd edition (1999), and Carlat's *The Psychiatric Interview: A Practical Guide* (1999).

From a historical perspective, clinical interviewing has continually evolved and undoubtedly will continue to change as clarifying theories and treatment modalities grow in number and depth. Clinicians have been forced to cope with the realization that the contemporary clinical interview frequently requires a high degree of standardization, as exemplified by the demand for larger amounts of data, of both diagnostic and psychosocial importance, to be gathered in relatively short periods of time. At first it appeared that these requirements would necessitate clinical interviews to be tightly scheduled or semistructured in nature. But with the advent of sensitive structuring approaches, such as facilics, clinicians remain free to utilize flexibly structured styles of interviewing, as exemplified by the CASE strategy and the work of Othmer and Othmer. Such styles provide clinicians with methods of gathering thorough data-bases in relatively short periods of time, scheduling the interview as they go along, each interview representing a unique creative venture.

With the historical review completed it is valuable to provide a practical introduction to two of the most powerful influences mentioned earlier. The first influence, which dates back to the pioneering work of Adolf Meyer, is the mental status. The second, and much more recent influence, is development of the DSM-III, the DSM-III-R, and the DSM-IV diagnostic systems.

THE MENTAL STATUS EXAMINATION

The mental status represents an attempt to describe objectively the behaviors, thoughts, feelings, and perceptions of the patient during the course of the interview itself. These observations are usually written as a separate section of the patient's evaluation. The general topics covered by the mental status are categorized as follows: appearance and behavior, speech characteristics and thought process, thought content, perception, mood and affect, sensorium, cognitive ability, and insight. Clinicians may vary on the exact categories that are used, and some clinicians collect all of these observations into a single narrative paragraph. In any case, the clinician attempts to convey the state of the patient during the interview itself, as if a cross-section were being taken of the patient's behavior for 60 minutes.

In a sense, the mental status consists of a variety of different clinician activities ranging from observation to the written record. Part of the mental status occurs informally as the clinician observes various characteristics of the patient while the patient spontaneously describes symptoms or history. The clinician may note whether the patient appears to be shabbily dressed or able to concentrate. Other aspects of the mental status are more formal in nature as the clinician asks direct questions concerning areas of psychopathology, such as inquiries regarding hallucinations or delusions. Finally, certain aspects of the mental status may be quite formalized as is seen during the formal cognitive examination. During this part the patient is asked to perform tasks, such as calculations or digit spans.

All of these clinician activities are synthesized into the written mental status. Indeed, it is by examining the thought processes required to produce a sound written mental status that one can best discuss the more intangible processes at work during the "gathering of the mental status information." Consequently, in this chapter each segment of the mental status is examined as it might appear in a standard written evaluation. An effort is made to summarize commonly utilized descriptive terms, to clarify confusing terms, to point out common mistakes, and to provide an example of a well-written mental status.

Appearance and Behavior

In this section the clinician attempts to describe accurately the patient's outward behavior and presentation. One place to start is with a description of the patient's clothes and self-care. Striking characteristics such as scars and deformities should be noted, as well as any tendencies for the patient to look older or younger than his or her chronological age. Eye contact is usually mentioned. Any peculiar mannerisms are noted, such as twitches or the patient's apparent responses to hallucinations, which may be evident through tracking movements of the eyes or a shaking of the head as if shutting out an unwanted voice. The clinician should note the patient's motor behavior; common descriptive terms include *restless, agitated, subdued, shaking, tremulous, rigid, pacing,* and *withdrawn*. Displacement activities such as picking at a cup or chain smoking are frequently mentioned. An important, and often forgotten characteristic, is the patient's apparent attitude toward the interviewer. With these ideas as a guide, the following excerpt represents a relatively poor description.

> *Clinician A:* The patient appeared disheveled. Her behavior was somewhat odd and her eye contact did not seem right. She appeared restless and her clothing seemed inappropriate.

Although this selection gives some idea of the patient's appearance, one does not come away with a sense of what it would be like to meet this patient. Generalities are used instead of specifics. The following description of the same patient captures her presence more aptly.

> *Clinician B:* The patient presents in tattered clothes, all of which appear filthy. Her nails are laden with dirt, and she literally has her soiled wig on backwards. She is wearing two wrist watches on her left wrist and tightly grasps a third watch in her right hand, which she will not open to shake hands. Her arms and knees moved restlessly throughout the interview, and she stood up to pace on a few occasions. She did not give any evidence of active response to hallucinations. She smelled badly but did not smell of alcohol. At times she seemed mildly uncooperative.

This passage presents a more vivid picture of her behavior, a pattern that may be consistent with a manic or psychotic process. Her "odd" behaviors have been concretely described. The clinician has included pertinent negatives, indicating that she shows no immediate evidence of hallucinating, as might be seen in a delirium.

Speech Characteristics and Thought Process

The clinician can address various aspects of the patient's speech, including the speech rate, volume, and tone of voice. At the same time, the clinician attempts to describe the thought process of the patient, as it is reflected in the manner with which the patient's words are organized. The term formal *thought disorder* is utilized to suggest the presence of abnormalities in the form and organization of the patient's thought. The less commonly used term, *content thought disorder,* refers specifically to the presence of delusions, and is addressed in a subsequent section of the mental status. The more generic term *thought disorder* includes both the concept of a formal thought disorder and of a content thought disorder. In this section of the mental status the emphasis is on the process of the thought (presence of a formal thought disorder), not the content of the speech. Terms frequently used by clinicians include the following:

Pressured speech: This term refers to an increased rate of speech, which may possibly best be described as a "speech sans punctuation." Sometimes it is only mildly pressured, whereas at other times, the patient's speech may virtually gush forth in an endless stream. It is commonly seen in mania, agitated psychotic states, or during extreme anxiety or anger.

Tangential thought: The patient's thoughts tend to wander off the subject as he or she proceeds to take tangents off his or her own statements. There tends to be some connection between the preceding thought and the subsequent statement. An example of fairly striking tangential thought would be as follows: "I really have not felt very good recently. My mood is shot, sort of like it was back in Kansas. Oh boy, those were bad days back in Kansas. I'd just come up from the Army and I was really homesick. Nothing can really beat home if you know what I mean. I vividly remember my mother's hot cherry tarts. Boy, they were good. Home cooking just can't be beat." *Circumstantial thought* is identical in nature to tangential thought but differs in that the patient returns to the original topic.

Loosening of associations: The patient's thoughts at times appear unconnected. Of course, to the patient, there may be obvious connections, but a normal listener would have trouble making them. In mild forms, loosening of associations may represent severe anxiety or evidence of a schizotypal character structure. In moderate or severe degrees, it is generally an indicator of psychosis. An example of a moderate degree of loosening would look like this: "I haven't felt good recently. My mood is shot, fluid like a waterfall that's black, back home I felt much better, cherry tarts and Mom's hot breath keeps you going and rolling along life's highways." If loosening becomes extremely severe it is sometimes referred to as a *word salad.*

Flight of ideas: In my opinion this is a relatively weak term, for it essentially represents combinations of the above terms. This is why most trainees find it confusing. For flight of ideas to occur, the patient must demonstrate tangential thought or a loosening of associations in conjunction with a significantly pressured speech. Usually there are connections between the thoughts, but, at times, a true loosening of associations is seen. A frequently, but not always, seen characteristic of flight of ideas is the tendency for the patient's speech to be triggered by distracting stimuli or to demonstrate plays on words. When present, these features represent more distinguishing hallmarks of a flight of ideas. Flight of ideas is commonly seen in mania, but can appear in any severely agitated or psychotic state.

Thought blocking: The patients stop in mid-sentence and never return to the original idea. These patients appear as if something has abruptly interrupted their train of thought and, indeed, something usually has, such as an hallucination or an influx of confusing ideation. Thought blocking is very frequently a sign of psychosis. It is not the same as exhibiting long periods of silence before answering questions. Some dynamic theorists believe it can also be seen in neurotic conditions, when a repressed impulse is threatening to break into consciousness.

Illogical thought: The patient displays illogical conclusions. This is different from a delusion, which represents a false belief but generally has logical reasoning behind it. An example of a mildly illogical thought follows: "My brother has spent a lot of time with his income taxes so he must be extremely wealthy. And everyone knows this as a fact because I see a lot of people deferring to him." These conclusions may be true, but they do not necessarily logically follow. Of course, in a more severe form, the illogical pattern may be quite striking: "I went to Mass every Sunday, so my boss should have given me a raise. That bum didn't even recognize my religious commitment."

In the following excerpt, the speech and thought process of the woman with two watches on her wrist is depicted. Once again the description demonstrates some areas in need of improvement.

Clinician A: Patient positive for loosening of associations and tangential thought. Otherwise grossly within normal limits.

This clinician has made no reference to the degree of severity of the formal thought disorder. Specifically, does this patient have a mild loosening of associations or does she verge upon a word

salad? Moreover, the clinician makes no reference to her speech rate and volume, characteristics that are frequently abnormal in manic patients. The following brief description supplies a significantly richer data base:

> *Clinician B:* The patient demonstrates a moderate pressure to her speech accompanied at times by loud outbursts. Even her baseline speech is slightly louder than normal. Her speech is moderately tangential, with rare instances of a mild loosening of associations. Without thought blocking or illogical thought.

Slowly one is beginning to develop a clearer picture of the degree of this patient's psychopathology. More evidence is mounting that there may be both a manic-like appearance and a psychotic process. In any case, the patient's speech coupled with her strikingly disheveled appearance, may lead the clinician to suspect that the patient is having trouble managing herself.

Thought Content

This section refers primarily to four broad issues: ruminations, obsessions, delusions, and the presence of suicidal or homicidal ideation. Ruminations are frequently seen in a variety of anxiety states and are particularly common in depressed patients. Significantly depressed patients will tend to be preoccupied with worries and feelings of guilt, constantly turning the thoughts over in their minds. The thinking process itself does not appear strange to these patients, and they do not generally try to stop it. Instead, they are too caught up in the process to do much other than talk about their problems. In contrast, obsessions have a different flavor to them, although they may overlap with ruminations at times.

An obsession is a specific thought that is repeated over and over by the patient as if he or she is seeking an answer to some question. Indeed, the patient frequently demonstrates obsessions over a question and its answer. As soon as the question is answered, the patient feels an intense need to ask it again, as if some process had been left undone. The patient may repeat this process hundreds of times in a row until it "feels right." If one interrupts the patient while this process is occurring, the patient will frequently feel a need to start the whole process again. Unlike the case with ruminations, patients find these obsessive thought processes to be both odd and painful. They frequently have tried various techniques to interrupt the process. Common themes for obsessions include thoughts of committing violence, homosexual fears, issues of right and wrong, and worries concerning dirt or filth. Obsessions may consist of recurrent ideas, thoughts, fantasies, images, or impulses. If the clinician takes the time to listen carefully to the patient, bearing the above phenomenological issues in mind, he or she can usually differentiate between ruminations and obsessions.

Delusions represent strongly held beliefs, that are not correct or held to be true by the vast majority of the patient's culture. They may range from bizarre thoughts, such as invasion of the world by aliens, to delusions of an intense feeling of worthlessness and hopelessness.

The fourth issue consists of statements concerning lethality. Because all patients should be asked about current lethality issues, these issues should always be addressed in the written mental status. In general, the clinician should make some statement regarding the presence of suicidal wishes, plans, and degree of intent to follow the plans in an immediate sense. If a plan is mentioned, the clinician should state the degree to which any action has been taken on it. He or she should also note whether any homicidal ideation is present and to what degree, as with suicidal ideation.

> *Clinician A:* The patient is psychotic and can't take care of herself. She seems delirious.

This excerpt is just simply sloppy. The first statement has no place in a mental status, for it is the beginning of the clinician's clinical assessment. The description of the delusion is threadbare and unrevealing. The clinician has also omitted the questioning concerning lethality. Assuming the clinician asked but forgot to record this information, he or she may sorely regret this omission if this patient were to kill herself and the clinician was taken to court to face a malpractice suit for possible negligence. A more useful description is given below.

> *Clinician B:* The patient appears convinced that if the watch is removed from her right hand, the world will come to an end. She relates that, consequently, she has not bathed for three weeks. She also feels that an army of rats is following her and is intending to enter her intestines to destroy "my vital essence."

She denies current suicidal ideation or plans. She denies homicidal ideation. Without ruminations or obsessions.

With this description it has become clear that the patient is psychotic, as evidenced by her delusions. The next question is whether hallucinations play a role in her psychotic process.

Perception

This section refers to the presence or absence of hallucinations or illusions. It is of value to note that there is sometimes a close relationship between delusions and hallucinations. It is not uncommon for the presence of hallucinations eventually to trigger the development of delusional thinking, but the two should not be confused. Let us assume that a patient is being hounded by a voice screaming, "You are possessed. You are a worthless demon." If the patient refuses to believe in the reality of the voice, then one would say that the patient is hearing voices but is not delusional. If, on the other hand, the patient eventually begins to believe in the existence of the voice and feels that the devil is planning her death, then the patient is said to have developed a delusion as well. The following description of perceptual phenomena is obviously threadbare.

Clinician A: Without abnormal perceptions.

There is a question concerning the appropriateness in the mental status of using phrases such as, "grossly within normal limits" or "without abnormality." Generally, the mental status is improved by the use of more precise and specific descriptions, but sometimes clinical situations require flexibility. For example, if the clinician is working under extreme time constraints, such global statements may be appropriate; in most situations, however, it is preferable to state specifically the main entities that were ruled out, for this tells the reader that the clinician actually looked for these specific processes. Stated differently, with these global phrases, the reader does not know whether they are accurate or the end result of a sloppy examination. If one has performed a careful examination, it seems best to let the reader know this.

There is another problem with the phrasing used by Clinician A: he or she has stated that the patient does not, in actuality, have hallucinations. It is possible, however, that this patient is simply withholding information out of fear that the voices represent a sickness. Numerous reasons exist for a patient to avoid sharing the presence of hallucinations with a clinician, including instructions to the patient from the voices not to speak about them to the clinician. It may be more accurate to state that the patient denied having hallucinations rather than to report categorically that the patient is without them. A more sophisticated report would be as follows:

Clinician B: The patient denied both visual and auditory hallucinations and any other perceptual abnormality.

Mood and Affect

Mood is a symptom, reported by the patient, concerning how he or she has generally been feeling recently, and it tends to be relatively persistent. Affect is a physical indicator noted by the clinician as to the immediately demonstrated feelings of the patient. Affect is demonstrated by the patient's facial expressions and other nonverbal clues during the interview itself; it is frequently of a transient nature. Mood is a self-reported symptom; affect is a physical sign. If a patient refuses to talk, the clinician can say essentially nothing about mood in the mental status itself, except that the patient refused to comment on mood. Later in the narrative assessment, the clinician will have ample space to describe his or her impressions of what the patient's actual mood has been. In contrast to mood, in which the clinician is dependent upon the patient's self-report, the clinician can always say something about the patient's affect.

Clinician A: The patient's mood is fine and her affect is appropriate but angry at times.

This statement is somewhat confusing. In which sense is her affect appropriate? Is it appropriately fearful for a person who believes that rats are invading her intestines or does the clinician mean that her affect is appropriate for a person without a delusional system? The clinician should always first state what the patient's affect is and then comment upon its appropriateness. Typical terms used to describe affect include *normal (broad) affect with full range of expression, restricted affect* (some decrease in facial animation), *blunted affect* (a fairly striking decrease in facial expression), *a flat affect* (essentially no sign of spontaneous facial activity), *buoyant affect, angry affect, suspicious*

affect, frightened affect, flirtatious affect, silly affect, threatening affect, labile affect, and *edgy affect.* The following description gives a much clearer feeling for this patient's presentation:

> *Clinician B:* When asked about her mood, the patient abruptly retorted, "My mood is just fine, thank you!" Throughout much of the interview she presented a guarded and mildly hostile affect, frequently clipping off her answers tersely. When talking about the nurse in the waiting area she became particularly suspicious and seemed genuinely frightened. Without tearfulness or a lability of affect.

Sensorium, Cognitive Functioning, and Insight

In this section the clinician attempts to convey a sense of the patient's basic level of functioning with regard to the level of consciousness, intellectual functioning, insight, and motivation. It is always important to note whether a patient presents with a normal level of consciousness, using phrases such as "The patient appeared alert with a stable level of consciousness," or "The patient's consciousness fluctuated rapidly from somnolence to agitation."

It should be noted that this section of the mental status examination may have evolved from two processes: the informal cognitive examination and the formal cognitive examination. The informal cognitive examination is artfully performed throughout the interview in a noninvasive fashion. The clinician essentially "eyeballs" the patient's concentration and memory by noting the method by which he or she responds to questions. If the clinician chooses to perform a more formal cognitive examination, it can range from a brief survey of orientation, digit spans, and short term memory, to a much more comprehensive examination, perhaps lasting 20 minutes or so. Clinical considerations will determine which approach is most appropriate. For a fast reading and penetrating discussion of the use of the formal cognitive exam, the reader is referred to *The Mental Status Examination in Neurology,* (Strub & Black, 1979). The reader may also be interested in becoming familiar with the 30-point Folstein Mini-Mental State Exam. This exam can be given in about 10 minutes, provides a standardized set of scores for comparison, and is extremely popular (Folstein, Folstein, & McHugh, 1975). There are two further outstanding resources on the mental status for the interested reader. Arguably the single best introduction to the mental status is Robinson's and Chapman's *Brian Calipers: A Guide to a Successful Mental Status Exam* (1997). This very readable primer is written with both wit and a keen eye for practicality. On a more comprehensive level, Trzepacz and Baker's book, *The Psychiatric Mental Status* (1993) is, in my opinion, the single best reference book on the mental status currently available, filled with concise definitions and clinical applications.

With regard to the patient in question, the following description is a weak one and could use some polishing:

> *Clinician A:* The patient seemed alert. She was oriented. Memory seemed fine and cognitive functioning was grossly within normal limits.

Once again this clinician's report is vague. Most importantly, the reader has no idea how much cognitive testing was performed. No mention has been made regarding the patient's insight or motivation. The following excerpt provides a more clarifying picture:

> *Clinician B:* The patient appeared alert with a stable level of consciousness throughout the interview. Indeed, at times, she seemed hyperalert and overly aware of her environment. She was oriented to person, place, and time. She could repeat six digits forward and four backward. She accurately recalled three objects after five minutes. Other formal testing was not performed. Her insight was very poor as was her judgment. She does not want help at this time and flatly refuses the use of any medication.

When done well, as described above, the mental status can provide a fellow clinician with a reliable image of the patient's actual presentation over the course of the interview. It should be openly acknowledged that, in actual practice, the written mental status may need to be significantly briefer, but the principles outlined above remain important and can help prevent the briefer mental status from being transformed into an inept mental status.

THE DSM-IV

The importance, with respect to interviewing, of the DSM-IV system does not pertain to any specific interviewing technique or mode of questioning. The DSM-IV is not a style of interview; it is a diagnostic system. Its impact on interviewing derives from its having established an important

set of symptoms that must be covered in order for a thorough assessment of the stated criteria to take place. In this sense, the DSM-IV has become an important factor in determining the type and amount of data that contemporary clinicians must address. With the advent of the DSM-III, the DSM-III-R, and the DSM-IV, the degree of standardization required in a typical "intake interview" has increased significantly, for the required database has grown significantly.

In this section a brief outline of the DSM-IV system is provided as an introduction to utilizing the system in practice. An attempt is also made to highlight some of the more important conceptual advances of the DSM-III system as it was revised and ultimately developed into the DSM-IV itself. For the more interested reader, Frances, First, and Ross (1995) have written an excellent brief review of the changes in the DSM-IV from the DSM-III-R.

When the DSM-III appeared in 1980, it represented several major advances. First, as compared to the Feighner criteria or the SADS, it was a system designed primarily for clinical practice rather than for application to a research setting. This clinical orientation mandated that all areas of psychopathology be delineated. The actual diagnoses were intended to be distinct from one another. Consequently, a second major advance, in comparison to the DSM-I and the DSM-II systems, was an emphasis on well-defined criteria for almost all the diagnostic categories.

The third major advance, and perhaps the most important, was the utilization of a multi-axial system, in which the patient's presentation was not limited to a single diagnosis. The clinician was pushed to look at the patient's primary psychiatric diagnosis within the context of a variety of interacting systems, such as the patient's physical health, level of stress, and level of functioning. As Mezzich (1985) has pointed out, the DSM-III system evolved from pioneering work with multiaxial systems across the world including England (Rutter, Shaffer, & Shepherd, 1975; Wing, 1970), Germany (Helmchen, 1975; von Cranach, 1977), Japan (Kato, 1977), and Sweden (Ottosson & Perris, 1973).

There are very few major changes in the DSM-IV over its immediate predecessor the DSM-III-R. This is because of the extensive work that went into the preparation of the DSM-II-R itself, with a heavy emphasis on empirical trials and data as opposed to expert opinion. This tradition was carried on with the DSM-IV Task Force, which took a conservative stance towards change, essentially making changes only when such changes would simplify the system or bring it into agreement with available empirical data.

Perhaps the most important change was a philosophical clarification regarding the etiology of some major mental disorders. The term, "Organic Disorders", which was used for diseases such as dementia and delirium in the DSM-III-R was removed. This term gave the misleading impression that other mental disorders such as schizophrenia and bipolar disorder were not caused by organic dysfunction, despite the fact that there is substantial data suggesting that biochemical dysfunction plays an etiologic role in such disorders.

A second set of changes was the attempt to simplify diagnoses if possible. Perhaps the best example of this is the criteria list for Somatization Disorder , that decreased from an intimidating and hard-to-remember list of 35 symptoms to a list that only must include four pain symptoms, two gastrointestinal symptoms, one sexual symptom, and one pseudoneurological symptom.

A third set of changes was the attempt to clarify certain variable characteristics of disorders that may have a direct impact upon treatment decisions or that might push a clinician to hunt more aggressively for a specific disorder. For example, in Mood Disorders, Bipolar II Disorder (at least one major depressive disorder plus at least one hypomanic episode without overt mania) was added. Specifiers such as "with rapid cycling" and "with seasonal pattern" were added as qualifiers that could indicate treatment interventions and have prognostic significance. Another nice example was the splitting of Attention-Deficit/Hyperactivity Disorder into three subtypes: Predominantly Hyperactive-Impulsive Type, Predominantly Inattentive Type, and Combined Type. This categorization will push clinicians to look for the inattentive subtype that was probably underdiagnosed in the past.

The fourth set of changes was the effort that was put into expanding and enriching the narrative text sections. This enrichment helps the clinician to better understand the phenomenology of these disorders and to make a better differential diagnosis. It will also serve as an excellent introduction to these disorders for the more novice clinician. An effort has also been made to help the clinician understand the cross-cultural variations of these disorders.

In the DSM-IV the clinical formulation is summarized on the following five axes:

Axis I: All clinical disorders and other conditions that may be a focus of clinicial attention (except for personality disorders and mental retardation)
Axis II: Personality disorders and mental retardation
Axis III: General medical conditions
Axis IV: Psychosocial and environmental problems
Axis V: Global assessment of functioning

Each of these axes is examined in more detail below.

Axis I

At first glance, Axis I may appear somewhat intimidating because of the large number of diagnostic entities it contains. But the clinician can approach the system in a two-step manner which greatly simplifies the task. In the first step or *primary delineation,* the clinician determines whether the patient's symptoms suggest one or more of the major diagnostic regions of Axis I which are confined to the following 16 relatively easily remembered categories:

1. Disorders usually first diagnosed in infancy, childhood, or adolescence
2. Delirium, dementia, amnestic and other cognitive disorders
3. Mental disorders due to a general medical condition (e.g., personality change secondary to a frontal lobe tumor)
4. Substance-related disorders
5. Schizophrenia and other psychotic disorders
6. Mood disorders
7. Anxiety disorders
8. Somatoform disorders
9. Factitious disorders
10. Dissociative disorders
11. Sexual and gender-identity disorders
12. Eating disorders
13. Sleep disorders
14. Impulse-control disorders not otherwise classified (e.g., kleptomania or pathological gambling)
15. Adjustment disorders
16. Other conditions that may be a focus of clinical attention (includes V codes and entities such as psychological symptoms affecting a medical condition or medication-induced movement disorders such as tardive dyskinesia)

Clues to which general categories of disorders are most relevant to the patient in question will arise as the clinician explores the patient's history of the present illness, both spontaneously and with the use of probe questions. The clinician should keep in mind that there are childhood diagnoses, that may first come to clinical attention in adulthood, such as attention-deficit disorder.

It should also be kept in mind that developmental disorders are coded on Axis I and may reflect limited cognitive delays or pervasive developmental disorders involving serious cognitive, social, motor, and language disturbances. Examples include mathematics disorder, developmental coordination disorder, expressive language disorder, autistic disorder, and Rett's disorder.

The second step or *secondary delineation* consists of delineating the specific diagnoses under each broad category. In the secondary delineation the clinician clarifies the database so that an exact DSM-IV diagnosis can be determined. Thus, if the clinician suspects a mood disorder, he or she will eventually search for criteria substantiating specific mood diagnoses, such as major depression, bipolar disorder, dysthymic disorder, cyclothymic disorder, mood disorder due to a medical condition, mood disorder due to substance abuse, bipolar disorder not otherwise specified, depressive disorder not otherwise specified, and mood disorder not otherwise specified. This secondary delineation is performed in each broad diagnostic area deemed pertinent.

With regard to the interview process itself, the trained clinician performs these delineations in a highly flexible manner, always patterning the questioning in the fashion most compatible with the needs of the patient. Utilizing a flexibly structured format, the clinician can weave in and out of these diagnostic regions, as well as any other areas such as the social history or family history, in whatever fashion is most engaging for the patient. With the flexibly structured format the only limiting factor is that the standard database must be thoroughly explored by the end of the available time. It is up to the clinician to schedule the interview creatively. When done well, the interview

feels unstructured to the patient, yet delineates an accurate diagnosis.

One diagnostic area that warrants further explanation is the concept of the *V code.* V codes represent conditions not attributable to a mental disorder that have nevertheless become a focus of therapeutic intervention. Examples include academic problems, occupational problems, uncomplicated bereavement, partner relational problems, and others. The DSM-IV has specific V Codes for various types of abuse including physical and sexual abuse of both children and adults as well as neglect, emphasizing the importance of these areas for questioning in the initial interview. Sometimes V codes are used because no mental disorder is present and the patient is coping with one of the stressors just listed. They can also be used if the clinician feels that insufficient information is available to rule out a psychiatric syndrome, but in the meantime, an area for specified intervention is being highlighted.

Axis II

Axis II emphasizes the realization that all the Axis I disorders exist in the unique psychological milieu known as *personality.* Many mental health problems are primarily related to the vicissitudes of personality development. Moreover, the underlying personality of the patient can greatly affect the manner in which the clinician chooses to relate to the patient both in the interview and in subsequent therapy.

On Axis II the following diagnostic categories are utilized:

1. Paranoid personality disorder
2. Schizoid personality disorder
3. Schizotypal personality disorder
4. Antisocial personality disorder
5. Borderline personality disorder
6. Histrionic personality disorder
7. Narcissistic personality disorder
8. Avoidant personality disorder
9. Dependent personality disorder
10. Obsessive-compulsive personality disorder
11. Personality disorder not otherwise specified (NOS)

Mental retardation is also included on Axis II.

This axis also functions in many respects as an important integration center in which diagnostic concerns can be related to psychodynamic principles. For instance, the clinician is asked to look carefully for evidence not only of personality disorders but also of maladaptive personality traits. These traits can also be listed on Axis II. Along similar lines, the clinician may list specific defense mechanisms that may have been displayed during free-format areas of the interview or as methods of avoiding certain topics raised by the clinician. These defense mechanisms may range from those commonly seen in neurotic disorders, such as rationalization and intellectualization, to those seen in more severe disorders, such as denial, projection, and splitting.

Axis III

On this axis the clinician considers the role of physical disorders and conditions, especially those that are potentially relevant to the understanding or management of the individual's mental disorders. The clinician is asked to view the patient's psychiatric problems within the holistic context of the impact of these problems on physical health, and vice versa. This axis reinforces the idea that a sound medical review of systems and past medical history should be a component of any complete initial assessment by a mental-health professional.

In addition, other physical conditions that are not diseases may provide important information concerning the holistic state of the patient. For instance, it is relevant to know whether the patient is pregnant or is a trained athlete, for these conditions may point toward germane psychological issues and strengths.

Axis IV

This axis concerns itself with an examination of the current psychosocial and environmental problems affecting the interviewee. It examines the crucial interaction between the client and the environment in which he or she lives. Sometimes interviewers are swept away by diagnostic intrigues and fail to uncover the reality based problems confronting the patient. This axis helps to keep this important area in focus.

By way of illustration, on this axis the interviewer may discover that, secondary to a job lay-

off, the home of the patient is about to be foreclosed. Such information may suggest the need to help the patient make contact with a specific social agency or the utility of a referral to a case manager. When the clinical task focuses upon crisis-intervention techniques and solution-focused strategies, as are commonly utilized in managed-care settings, Axis IV becomes of primary importance, for it points directly towards possible areas for immediate intervention and support.

Axis V

A variety of changes were made in Axis V when the DSM-III was revised into the DSM-III-R. In DSM-III this axis delineated only the highest functioning of the patient over a two-month period in the preceding year. This relatively narrow perspective did not provide an abundance of practical information. Consequently, in the DSM-III-R this axis was broadened. It included not only a rating of the highest functioning in the past year, but also a rating of the current functioning, which provided immediate data pertinent to treatment planning and the decision as to whether hospitalization was warranted. These ratings were to be made by combining both symptoms and occupational and interpersonal functioning on a 90-point scale, the Global Assessment Functioning Scale (GAF Scale).

In DSM-IV the same procedure and scale are utilized (scale range is now 0–100), except the process has been streamlined to only require a GAF rating of "current functioning". Other time frames can be added, in which case the additional time frame is indicated in parentheses after the additional score. Examples would be as follows, 45 (highest level in past year) or 70 (at discharge from hospital).

Probably of even more practical importance to the clinician is the window that this axis opens into the patient's adaptive skills and coping mechanisms as reflected in the rating of current functioning. Looking for strengths to capitalize upon and to utilize as foundations for solution-focused problem solving is equally important as finding out areas of malfunction and pathology. The gifted initial interviewer is equally adept at uncovering what is right and what is wrong. Both regions of knowledge are critical in order to provide the most rapid and long-lasting relief for the client seeking help.

This brief review of the DSM-IV system shows that the impact of this new diagnostic system on the interviewing process has been manyfold. The multi-axial approach and the thoroughness of the diagnostic schema require the clinician to cover a lot of ground, especially during a one-session intake, as is often necessary in a managed-care setting. But the resulting standardized database is an illuminating one that highlights a holistic and rigorous approach to understanding the patient's problems, strengths, and needs. Moreover, the skilled use of a flexibly structured interview allows this informational base to be gathered in an empathic and flowing manner.

RESEARCH ON INTERVIEWING

It is not an exaggeration to state that it would require an entire book to review comprehensively the vast literature related to interviewing. On a more modest level, an attempt is made in this chapter to introduce the reader to the main currents of this research area, providing a simplifying schema for categorizing the available literature while referencing specific articles that can be used as stepping stones into the categories described.

One of the confusions facing the reader, as he or she attempts to approach the research on interviewing, is the significant overlap between interviewing research and research done with regard to psychotherapy. This overlap is a healthy one, for it demonstrates that alliance issues are in some respects inseparable from data-gathering issues. There is an intimate relationship between the strength of the initial alliance and the resulting ability to gather valid information and structure the flow of the conversation effectively.

On the other hand, there are differences in emphasis between intake interviews and psychotherapy sessions. As the degree of standardization has increased with the advent of new therapies and new diagnostic systems, these differences have become more apparent. Eventually, such differentiation between interviewing and psychotherapy will probably be reflected more distinctly in the research literature as increased research occurs on structuring techniques and validity concerns. With these qualifications observed, the following major research areas will be discussed: (a) clinician response modes, (b) nonverbal behavior and paralanguage, (c) clinician characteristics as related to alliance issues and empathic communication, (d)

reliability and validity concerns as related to structured interviews, and (e) educational techniques.

With regard to the first category, response-mode research attempts to examine the type of verbal exchange occurring between clinician and patient, focusing on styles of response such as open-ended questions, reflections, and interpretations. Stiles (1978) notes that it is important to separate response-mode research from content research, which focuses on the actual meaning of the words spoken, and research on extra-linguistic areas, such as speech characteristics, pauses, and laughter.

It has been estimated that 20 to 30 response-mode systems have been developed (Elliott et al., 1987). Much of the pioneering research on response modes was done during the 1950s, 1960s, and 1970s (Aronson, 1953; Danish & D'Augelli, 1976; Goodman & Dooley, 1976; Hackney & Nye, 1973; Hill, 1975; Ivey, 1971; Robinson, 1950; Snyder, 1945, 1963; Spooner & Stone, 1977; Strupp, 1960; Whalen & Flowers, 1977).

In 1978 Clara Hill developed a system that integrated many of the best features of the earlier systems. Her system consisted of 14 categories, including response types such as minimal-encourager, direct-guidance, closed-question, open-question, self-disclosure, confrontation, approval-reassurance, and restatement. Hill's system was further developed to include three supercategories that focused on the degree of structuring as seen with low structure (encouragement/approval/reassurance, reflection/restatement, and self-disclosure), moderate structure (confrontation, interpretation, and provision of information) and high structure (direct guidance/advice and information seeking) (Friedlander, 1982).

In 1987, six of the major-rating systems were compared when applied to therapy sessions by well-known clinicians such as Albert Ellis and Carl Rogers (Elliott et al., 1987). Interrater reliability was found to be high; when categories in different rating systems were collapsed to the same level of specificity, moderate to strong convergence was found. Studies such as the above point to a bright future for response-mode research when systems with high interrater-reliability are applied to various interviewing situations. Different systems appear to shed slightly different light on the database, and the complementary use of various systems will probably become the preferred approach in the future.

In this regard the next logical step was to utilize response-mode systems to study patterns of clinician-response modes in actual clinical practice, attempting to delineate their possible impact on client engagement or behavior. Longborg demonstrated an increase in the use of information giving, confrontation, and minimal respones (encouragers and silence) over the time course of the initial interview of counseling trainees (Longborg, Daniels, Hammond, Houghton-Wenger, & Brace, 1991). Chang (1994) demonstrated a strong direct positive correlation between positive feedback in the initial session of a weight-reduction study and specific behavioral indicators of client compliance including return for second session, number of weekly report sheets completed, and amount of time spent on meditation homework. This excellent study is one of the few that focuses directly on positive client-outcome behaviors as opposed to client engagement.

Research with respect to nonverbal communication, the second major research category, spans a variety of perspectives that can best be separated into three areas known as proxemics, kinesics, and paralanguage. Edward T. Hall (1966) first coined the term *proxemics* in his classic book *The Hidden Dimension.* Proxemics represents the study of how humans conceptualize and utilize interpersonal space. Hall was particularly interested in the impact of culture on an individual's sense of interpersonal space. Kinesics is the study of the body in movement including movements of the torso, head, limbs, face, and eyes as well as the impact of posture. The field was pioneered by Ray T. Birdwhistell (1952) in the book, *Introduction to Kinesics: An Annotation System for Analysis of Body Motion and Gesture.* The final realm of nonverbal study is *paralanguage*, which focuses on how messages are delivered, including elements such as tone of voice, loudness of voice, pitch of voice, and fluency of speech (Cormier & Cormier, 1979).

The impact of these three areas of nonverbal behavior on the issue of social control has received much attention. Ekman has devoted considerable time to the nonverbal constituents of the act of lying (Ekman, 1985; Ekman & Friesen, 1974; Ekman & Rosenberg, 1998). In a concise review of the literature concerning nonverbal behavior and social control, including areas such as status, persuasion, feedback, deception, and impression formation, it appears that gaze and facial expression are the most telling factors (Edinger & Patterson, 1983). Scheflen (1972) has described *kinesic reciprocals* which represent display behaviors between two organisms that convey intent, such as

mating rituals, parenting behavior, and fighting behavior, all of which also reflect the role of nonverbal behavior in social control.

Another area of active research concerns those nonverbal behaviors that can facilitate the therapeutic alliance. Tepper and Haase (1978) emphasized the importance of considering a multichannel approach to understanding this subtle set of relationships. In one study they reviewed the impact of various factors including verbal message, trunk lean, eye contact, vocal intonation, and facial expression on facilitative concerns such as empathy, respect, and genuineness. Nonverbal components appeared to play a major role in these facilitative processes. Attempts have been made to determine methods of measuring clinician ability to decode the nonverbal behavior of patients. Rosenthal, Hall, DiMatteo, Rogers, and Archer (1979) developed the Profile of Nonverbal Sensitivity (PONS) in this regard. The original PONS consisted of 220 two-second film segments for which subjects were asked to read accurately nonverbal clues, such as facial expression and tone of voice.

The issue of decoding nonverbal cues immediately raises the concept of cross-cultural differences with regard to interpretation of nonverbal behavior. As mentioned earlier, Hall was fascinated by this process and, in more recent times, Sue has studied these ramifications in detail (Sue, 1981; Sue & Sue, 1977).

Further issues concerning the complicated nature of how clinicians decode nonverbal language was more recently addressed by Hill and Stephany (1990). They studied the presence of nonverbal behaviors, such as speech hesitancies, vertical head movements, horizontal head movements, arm movements, leg movements, postural shifts, adaptors, illustrators, and smiles with recognition by clinicians of moments of therapeutic importance to clients.

Issues such as paralanguage and temporal-speech characteristics have been carefully studied. Matarazzo and Wiens (1972) have developed a concrete system of exploring such interactions. They have delineated three major temporal-speech characteristics: duration of utterance (DOU), response time latency (RTL), and percentage of interruptions (Wiens, 1983). In conjunction with Harper, these same authors provide an insightful review of nonverbal behavior in *Nonverbal Communication: The State of the Art* (Harper, Wiens, & Matarazzo, 1978). As Tepper and Haase (1978) have emphasized, the future of nonverbal research probably lies in an integrative approach combining paralanguage concerns, such as those delineated by Matarazzo and Wiens (1972), with other proxemic and kinesic elements as they have impact on the interviewing relationship.

The interviewing relationship is further defined by the third major area of research which focuses on characteristics of the interviewer that affect the therapeutic alliance, such as communication style, race, physical attractiveness, and the ability to convey empathy. Because of its broad area of investigation, this type of research overlaps with some of the areas already described. For instance, response modes have been used to correlate client perceptions of clinician empathy with clinician phrasing, responses focused on exploration being strongly associated with perceived empathy (Barkham & Shapiro, 1986). In a similar vein, the child psychiatrist, Rutter, has developed a system of training clinicians to display four distinct styles ranging from a "sounding-board" style to a "structured" style. The impact of these styles on the interview process was then examined (Rutter, Cox, Egert, Holbrook, & Everitt, 1981).

The concept of empathy has received as much, if not more, emphasis than any other single clinician characteristic. As mentioned earlier, Rogers was pivotal in the development of thought related to the empathic process. Historically, another major contribution was made by Truax and Carkhuff (1967) who emphasized qualities such as accurate empathy, nonpossessive warmth, and interpersonal genuineness as critical to the development of a sound therapeutic alliance (Truax & Carkhuff, 1967). The Truax scale itself was a popular measure of empathy but has been attacked on numerous grounds ranging from a lack of specificity concerning the clinician behaviors in question, to the claim that the scale may be measuring more than one thing (Cochrane, 1974; Lambert, DeJulio, & Stein, 1978; Wenegrat, 1974; Zimmer & Anderson, 1968).

One of the more powerful unifying theories is the *empathy cycle* proposed by G. T. Barrett-Lennard (1981). The empathy cycle delineates the empathic process in five specific phases, including such processes as the clinician's ability to perceive the patient's feelings and the patient's ability to provide feedback that the empathic message has been received. The empathy cycle provides a framework from which differing components of the empathic process can be studied (Harmon,

1986). Over the years, numerous articles and reviews concerning empathy have spun off from the works previously described as well as the viewpoints espoused by the psychoanalytic community (Berger, 1987; Elliott et al., 1982; Elliott et al., 1987; Gladstein, 1983; Smith-Hanen, 1977).

When considering the broad region of the impact of clinician characteristics on alliance, one area of progress has been in the development of rating forms with regard to patient satisfaction with the interviewer. In 1975 Barak and LaCrosse developed the Counselor Rating Form which is also available in a shortened form (Barak & LaCrosse, 1975; Corrigan & Schmidt, 1983). Other scales have followed that emphasize the alliance as it develops in the psychotherapeutic relationship (Alexander & Luborsky, 1986; Marmar, Horowitz, Weiss, & Marziali, 1986). More recently, Mahalik (1994) has developed a scale for actually measuring client resistance along five continua: Opposing Expression of Painful Affect, Opposing Recollection of Material, Opposing Therapist, Opposing Change, and Opposing Insight (Client Resistance Scale [CRS]). In the same paper Mahalik used Hill's Response Modes Verbal Category System (Hill, 1978) to study correlations between clinician response modes and specific forms of resistance.

Degree of alliance has also been creatively approached by Stiles (1984), who developed the idea of measuring the *depth* of an interview and the *smoothness* of the interview with the Session Evaluation Questionnaire (SEQ). Depth was measured on five bipolar scales: deep-shallow, full-empty, powerful-weak, valuable-worthless, and special-ordinary. The smoothness index is the mean rating on the following five bipolar scales: comfortable-uncomfortabe, smooth-rough, easy-difficult, pleasant-unpleasant, and relaxed-tense. Utlizing the SEQ, Tryon (1990) correlated a higher engagement with deeper interviews and longer interviews as rated by both the client and the counselor. This work was also based on her concept of the *engagement quotient* (EQ), representing the percentage of clients who return to a counselor following the initial assessment (Tryon, 1985). Using the SEQ as well as four other engagement/outcome rating scales, Mallinckrodt (1993) examined the impact of session satisfaction and alliance strength over the course of time in a brief therapy format.

Attempts to focus on the specific clinician/client feelings, expectations, reactions, and behaviors during pivotal moments of the initial interview or ongoing therapy have become known as *significant events* research. Cummings, Slemon, and Hallberg (1993) have produced a good example of such work based partially upon their development of the Important Events Questionnaire (Cummings, Martin, Hallberg, & Slemon, 1992). This questionnaire has five questions, such as, "What was the most important thing that happened in this session for you?" and "Why was it important and how was it helpful or not helpful?" and can be completed by both clinician and client. An attempt to uncover some underlying general principles in significant-events research, using grounded theory-research technique, was done by Frontman and Kunkel (1994).

A major thrust in research dealing with clinician characteristics evolves from the work of Strong. His work with the interpersonal-influence theory of counseling has focused attention on the idea that counselors who were perceived as expert, attractive, and trustworthy would possess a more effective means of influencing the behaviors of their clients (Paradise, Conway, & Zweig, 1986; Strong, 1968; Strong, Taylor, Bratton, & Loper, 1971). For example, the physical attractiveness of the clinician appears to have a positive impact in certain situations (Cash, Begley, McCown, & Weise, 1975; McKee & Smouse, 1983; Vargas & Borkowski, 1982).

One of the major areas of recent research has been the impact of race and cultural sensitivity to both client outcome and clinician perception (Atkinson, Matshushita, & Yashiko, 1991; Helms, Carter, & Robert, 1991; Paurohit, Dowd, & Cottingham, 1982). Tomlinson-Clarke, using an archival study, does a nice job of delineating some of the stumbling blocks in research design and interpretation of results, that is inherent in research focusing upon the bias caused by race (Tomlinson-Clarke & Cheatham, 1993).

Other characteristics that have been studied include religious background (Keating & Fretz, 1990), body movement (LaCrosse, 1975), spontaneity and fluency of speech (Strong & Schmidt, 1970), and the role of displays of accreditation, such as diplomas, on the walls of the clinician's office (Siegel & Sell, 1978).

In concluding a review of the literature, describing the impact of clinician characteristics on alliance, it is natural to mention some of the work based on the ultimate measure of clinician impact as shown by impact on compliance and follow-up. A number of issues have been studied, such as the

impact of the degree of directiveness, counselor gender, and counselor experience, as well as the clinician's willingness to negotiate a therapeutic contract. It appears that the ability to convey accurately a sensitive understanding of the patient's problem and the ability to negotiate future treatment plans flexibly are powerful predictors of compliance (Eisenthal, Koopman, & Lazare, 1983; Eisenthal & Lazare, 1977a, 1977b; Epperson, Bushway, & Warman, 1983; Heilbrun, 1974). Finally, two good reviews on process- and outcome-research techniques have been done by Hill (Hill, 1990; Hill and Corbett, 1993).

The fourth major area of interviewing research leaves the arena of interpersonal dynamics and focuses more on the issue of structured and semistructured interviews and their impact on the thoroughness, reliability, and validity of the database. Whereas much of the process research previously described has evolved from the fields of counseling and psychology, a large part of the work on structured interviews has been undertaken in the field of psychiatry.

In many respects structured and semistructured interviews grew out of the tradition of psychiatric epidemiology (Helzer, 1983). Examples include the Home Interview Survey (HIS) used in the Midtown Manhattan Study (Srole, Langer, Michael, Opler, & Rennie, 1962) and the Psychiatric Epidemiological Research Interview (PERI) developed by Dohrenwend (Dohrenwend, Shrout, Egri, & Mendelsohn, 1980). One of the most influential interviews that dealt directly with psychiatric symptomatology and diagnosis was the Present State Examination (PSE) developed by Wing in England (Wing, Cooper, & Sartorius, 1974).

The PSE combines elements of both the recent psychiatric history and the mental status. It represents a semistructured interview which emphasizes the need for the interviewer to cross-examine in a flexible manner when attempting to delineate the presence and severity of a symptom. The PSE has undergone numerous editions, and the ninth edition can be used in conjunction with a computer program, CATEGO, which will delineate a diagnosis from the data gathered during the interview. The ninth version contains 140 principal items and its phenomenological approach creates a Western European feel in the interview format (Hedlund & Vieweg, 1981). Numerous studies have been undertaken with regard to the reliability of the PSE (Cooper, Copeland, Brown, Harris, & Gourlay, 1977; Wing, Nixon, Mann, & Leff, 1977).

Several important interviews have already been mentioned during the historical survey earlier in the chapter including the Diagnostic Interview Schedule (DIS) and the Schedule for Affective Disorders and Schizophrenia (SADS). All of these interview formats were developed with the idea of increasing the thoroughness, reliability, and validity of the database. In some respects, these goals have been at least partially realized. But Sanson-Fisher and Martin (1981) have emphasized an important point. Because these interviews have been shown to be reliable, researchers tend to assume that the interviews will automatically be reliable in the hands of the clinicians working in their protocols. This assumption is not necessarily the case. It is important that reliability studies be used at each research site and in an ongoing fashion if, indeed, the interview format is to function with a high degree of reliability.

Before leaving the area of structured interviews and their impact on reliability and validity concerns, it is important to mention the major role that child psychiatrists have had in the development of interview formats. A variety of interviews have been developed including the Diagnostic Interview for Children and Adolescents (DICA) (Herjanic & Campbell, 1977; Herjanic & Reich, 1982), the Interview Schedule for Children (ISC) (Kovacs, 1983), the Kiddie-SADS (Puig-Antich & Chambers, 1978), the Diagnostic Interview Schedule for Children (DISC) (Costello, Edelbrock, Kalas, & Dulcan, 1984), and the semistructured interview developed by Rutter and Graham (1968). The development of such interviewing tools has allowed researchers to address the intriguing questions concerning the correlation between developmental age and the validity of information provided by children (Edelbrock, Costello, Dulcan, Kalas, & Conover, 1985).

The fifth, and final major area in interviewing research concerns developments in educational techniques. This field is both exciting and broad, with contributions from all disciplines of mental health. For the sake of simplicity, it is best to group this research into two large areas: the development of improved supervision techniques and the development of tools for measuring student learning with regard to interviewing skills.

In the same fashion that there has been a striking evolution in the number of treatment modalities now available, there has been an equally remarkable advancement in training techniques over the past several decades. For many years, interviewing

training seemed to be stuck on the model of indirect supervision that had evolved from the psychoanalytic tradition. With indirect supervision the trainee sees the patient alone and then reports on "what happened" to the supervisor. Indirect supervision, when done well, can be very effective, providing an intimate and carefully individualized supervision, but it has obvious limitations.

The idea that the supervisor could actually "sit in" with the patient and the interviewer probably developed from a variety of disciplines. For instance, the idea of direct supervision is a popular style of supervision in family therapy. With regard to interviewing an individual patient, numerous advantages appear when comparing direct to indirect supervision (Digiacomo, 1982; Stein, Karasu, Charles, & Buckley, 1975).

Direct supervision removes many of the distorting mechanisms at work with the secondhand information provided by indirect supervision. In direct supervision the supervisor can more accurately evaluate nonverbal interaction, the structuring of the interview, and the handling of resistance. It also provides the trainee with the all-too-rare opportunity to model a more experienced clinician, if the supervisor chooses to demonstrate a technique. Rarely does direct supervision appear to hamper engagement with the patient significantly. In one study more than twice as many patients with direct supervision, compared with indirect supervision, remained in active treatment or successively completed therapy (Jaynes, Charles, Kass, & Holzman, 1979).

On the heels of direct supervision, the closely related concept of videotape supervision was developed. Such supervision complements both indirect and direct supervision. Like direct supervision it provides an excellent opportunity for feedback on nonverbal and structuring techniques. It also offers the advantage of helping the clinician to develop a more effective observing ego by literally experiencing the process of observing and analyzing his or her own behavior (Dowrick & Biggs, 1983; Jackson & Pinkerton, 1983; Maguire et al., 1978; Waldron, 1973).

The advent of recording technologies, such as audiotape and videotape, provided the foundation for an innovative style of supervision known as Interpersonal Process Recall (IPR). Bloom (1954) was one of the first to experiment with the technique in his attempt to explore the thought processes of college students during discussion sections. Kagan (1975) was the first to apply the technique to the clinical interview in the mental-health professions and coined the term *Interpersonal Process Recall.* In IPR the students are asked to reflect upon their internal feelings, thoughts, and reactions that are associated with specific clinical situations observed on videotapes of their own clinical interviews. It is an excellent tool for uncovering countertransference issues and other psychodynamic concerns. IPR is also a powerful method of helping trainees to recapture fleeting impressions that would normally be lost or distorted (Elliott, 1986).

Role playing provides yet another complementary and widely accepted avenue for enhancing specific interviewing skills (Canada, 1973; Errek & Randolpf, 1982; Hannay, 1980; Hutter et al., 1977). It may represent the single most effective manner by which to familiarize trainees with various methods of handling hostile or awkward patient questions.

Ward and Stein (1975) pioneered the concept of group supervision by colleagues. In this format the patient is interviewed by the trainee while fellow trainees observe in the same room. It provides a format in which the group identifies emotionally with both the patient and the interviewer, providing a unique window into the processes of engagement and empathy

Combining many of the advances just described, Ivey (1971) developed the innovative process of microtraining. Microtraining probably represents one of the most extensively studied and empirically proven of all the training techniques currently utilized. In this format, specific skills, such as the use of empathic statements or open-ended questions, are taught in an individualized fashion with a heavy emphasis on behavioral reinforcement. The trainee is videotaped. Immediate feedback is given and problem areas delineated. Concise, goal-directed reading material, related to a well-circumscribed skill, is provided. The trainee then immediately performs further role-plays during which the newly acquired skill is practiced until it is perfected, the trainee constantly being given concrete feedback from the supervisor.

Shea and colleagues would ultimately combine all of the above techniques into an innovative training program in interviewing (Shea & Mezzich, 1988; Shea, Mezzich, Bohon, & Zeiders, 1989). The first innovation in the training program was the idea that a unified block of highly supervised time, with an emphasis on direct mentorship and observation, should be set aside for trainees in

which interviewing skills, as opposed to psychotherapy skills, were intensively studied in an immediately relevant setting, such as an assessment clinic or emergency room. The second innovation was to focus, not only on traditional skills such as empathy and engagement, but on utilizing these skills in conjunction with real-life clinicial demands such as DSM-III-R differential diagnosis and suicide assessment and also performed with real-life time limitations. The third innovation consisted of integrating both theory and supervision techniques (such as facilics, videotaping, direct supervision, role playing, microtraining, macrotraining, and behavioral self-monitoring) from a variety of disciplines into a specialized training package that was designed into an individualized program for each specific trainee. Individualized learning goals were established as well as matching the specific training techniques to the needs and preferences of the trainee.

The second major area, with regard to research on interviewing in education, focuses less on the educational techniques themselves and more on methods of evaluating interviewing skills and determining whether or not educational goals have been achieved. It is interesting to note that much of the empirical work in this area has been done with medical-student education as opposed to psychiatric-resident education or mental-health-professional training.

One test technique consists of providing trainees with videotape vignettes followed by three possible physician responses. The trainee is asked to select the most appropriate response (Adler, Ware, & Enelow, 1968; Cline & Garrard, 1973). A written test that attempts to examine interviewer decision making with regard to the interview process has been described by Smith (Smith, Hadac, & Leversee, 1980). This instrument, called the Helping Relationship Inventory, consists of 10 brief patient statements. Each statement is followed by five alternative responses categorized as understanding, probing, interpretive, supportive, or evaluative. Liston has developed a tool for assessing the acquisition of psychotherapy skills known as the Psychotherapy Competence Assessment Schedule (PCAS) (Liston & Yager, 1982; Liston, Yager, & Strauss, 1981).

With regard to assessing the skills demonstrated in the initial medical or psychiatric interview, the vast majority of work has moved away from pencil-and-paper tests, focusing instead on direct or videotaped evaluation of actual clinical interviews. This body of literature is relatively large and is well reviewed by Ovadia, Yager, and Heinrich (1982). A representative example of one such format is the Queen's University Interview Rating Scale (QUIRS). This rating process was developed to test the psychiatric-interviewing skills of medical students as they rotated on third and fourth year clerkships (Jarrett, Waldron, Burra, & Handforth, 1972). The QUIRS consists of 23 items collapsed from a list of 75 skills drawn from the literature. The test items are organized into three supercategories: interview structure, interviewer role, and communication skills. With regard to medical interviewing, Brockway (1978) developed an extensive system for evaluating interviewing skills. This system includes over 50 items ranging from process items, such as the use of silence, to content items, such as eliciting the patient's rationale for making an appointment.

Levinson and Roter (1993) demonstrated that physicians who participated in a two and one-half day continuing-medical-education program, when compared to physicians who participated in a four and one-half hour workshop, showed significantly more improvement in interviewing skills. In the study five sequential patient visits were audiotaped one month before and one month after the trainings. The short-program group showed essentially no improvements. The long-program cohort showed the use of more open-ended questions, more frequently asked for the patient's opinions, and gave more biomedical information. Levinson and Roter emphasize the value of a patient-centered style of interviewing and a corresponding learner-centered style of teaching.

Kivlighan (1989) demonstrated improvements in psychotherapy skill in trainees who completed a course on interpersonal-dynamic therapy, including an increase in the use of minimal encouragers and the reported depth of the sessions as reported by the clients. The strength of this study was the use of both a good control group and a battery of scales that have been documented to have good reliablity and validity, including the Intentions List, Client Reactions System, Hill Counselor Verbal Response Category System (Hill, 1978), and the Session Evaluation Questionnaire. It has not been common for researchers studying interviewing training programs, to utilize well-tested rating instruments, such as those used by Kivlighan.

Along these lines, a paper written by Sanson-Fisher, Fairbairn, and Maguire (1981) provides a good ending point for this section, albeit a some-

what sobering one. In a review of 46 papers dealing with the teaching of communication and interviewing skills to medical students, the majority of the papers revealed methodological flaws. According to Sanson-Fisher, the future of research in this area should include a consolidated effort toward the use of standard research techniques including control groups, reliability studies, student characteristics, patient characteristics, and more sophisticated statistical analyses.

SUMMARY

In this chapter an attempt has been made to provide a sound introduction to the art and craft of initial assessment interviewing and diagnosis, including its history, core clinical concerns, and research. It can be seen that the historical currents of initial assessment interviewing are varied and rich. These currents include medical traditions such as the mental status, diagnostic systems, and psychoanalytic techniques. But they also include a remarkable array of contributions from nonmedical fields such as counseling and psychology.

At the present moment there is a cross-pollination among fields that is unusually promising. Research teams from different disciplines can be assembled to study the interviewing process from a variety of perspectives. These interdisciplinary teams can analyze engagement techniques, nonverbal processes, and structuring principles in the context of specific styles of interaction and characterological functioning, as determined by psychological testing and diagnosis by DSM-IV criteria. For the first time the role of response modes, empathic statements, and nonverbal techniques can be studied in relation to specific psychopathological states such as paranoia or to specific communication resistances as seen with overly loquacious patients.

This chapter began with an historical perspective, and it seems appropriate to end on an historical note as well. In 1806 the psychiatrist Philippe Pinel became renowned for his humanistic treatment of patients in the French institution known as the Asylum de Bicetre. In his book *A Treatise on Insanity* he wrote, "Few subjects in medicine are so intimately connected with the history and philosophy of the human mind as insanity. There are still fewer, where there are so many errors to rectify, and so many prejudices to remove." (p. 3). His point is as penetrating today as it was at the beginning of the 19th century. In the past, part of the prejudice blocking our understanding of human nature was created by a stubborn battle over turf among the various mental health traditions. Contemporary interviewing represents an area in which the disciplines can at last join forces to further our understanding of human nature, both as a function of psychopathology and as a function of health.

REFERENCES

Adler, L., Ware, J. E., & Enelow, A. J. (1968). *Evaluation of programmed instruction in medical interviewing.* Los Angeles: University of Southern California Postgraduate Division.

Alexander, L. B., & Luborsky, L. (1986). The Penn Helping Alliance Scales. In L. S. Greenberg & W.M. Pinsof (Eds.), *The psychotherapeutic process-A research handbook.* New York: Guilford Press.

American Psychiatric Association. (1980). *Diagnostic and statistical manual of mental disorders* (3rd ed.). Washington, DC: Author.

American Psychiatric Association. (1987). *Diagnostic and statistical manual of mental disorders* (3rd ed., Rev. ed.). Washington, DC: Author.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Aronson, M. (1953). A study of the relationships between certain counselor and client characteristics in client-centered therapy. In W. U. Snyder (Ed.), *Pennsylvania State College Psychotherapy Research Groups: Group report of a program of research in psychotherapy.*

Atkinson, D., Matshushita, R., & Yashiko, J. (1991). Japanese-American acculturation, counseling style, counselor ethnicity, and perceived counselor credibility. *Journal of Counseling Psychology*, 38, 473-478.

Barak, A., & LaCrosse, M. B. (1975). Multidimensional perception of counselor behavior. *Journal of Counseling Psychology, 22,* 417-476.

Barkham, M., & Shapiro, D. A. (1986). Counselor verbal response modes and experienced empathy. *Journal of Counseling Psychology, 33,* 3-10.

Barrett-Lennard, G. T. (1981). The empathy cycle: Refinement of a nuclear concept. *Journal of Counseling Psychology, 28,* 91-100.

Benjamin, A. (1969). *The helping interview.* Boston: Houghton-Mifflin.

Berger, D. M. (1987). *Clinical empathy.* Northvale, NJ: Jason Aronson.

Birdwhistell, R. L. (1952). *Introduction to Kinesics: An annotation system for analysis of body motion and gesture.* Louisville: University of Kentucky.

Bloom, B. S. (1954). The thought process of students in discussion. In S. J. French (Ed.), *Accent on teaching: Experiments in general education.* York: Harper.

Brockway, B. S., (1978). Evaluating physician competency: What difference does it make? *Evaluation and Program Planning, 1,* 211.

Canada, R. M. (1973). Immediate reinforcement versus delayed reinforcement in teaching a basic interview technique. *Journal of Counseling Psychology, 20,* 395-398.

Carlat, D. J. (1999). *The psychiatric interview: A practical guide.* New York: Lippincott Williams & Wilkins.

Cash, T. F., Begley, P. J., McCown, D. Q., Weise, B. C. (1975). When counselors are heard but not seen: Initial impact of physical attractiveness. *Journal of Counseling Psychology, 22,* 273-279.

Chang, P. (1994). Effects of interviewer questions and response type on compliance: An analogue study. *Journal of Counseling Psychology, 41,* 74-82.

Cline, D. W., & Garrard, J. N. (1973). A medical interviewing course: Objectives, techniques, and assessment. *American Journal of Psychiatry, 130,* 574-578.

Cochrane, C. T. (1974). Development of a measure of empathic communication. *Psychotherapy: Theory, Research and Practice, 11,* 41-47.

Cooper, J. E., Copeland, J. R. M., Brown, G. W., Harris, T., & Gourlay, A. J. (1977). Further studies on interviewer training and inter-rater reliability of the Present State Examination (PSE). *Psychological Medicine, 7,* 517-523.

Cormier, W. H., & Cormier, L. S. (1979) Interview*ing strategies for helpers—A guide to assessment, treatment, and evaluation.* Montery, CA: Brooks/Cole.

Corrigan, J. D., & Schmidt, L. D. (1983). Development and validation of revisions in the counselor rating form. *Journal of Counseling Psychology, 30,* 64-75.

Costello, A. J., Edlebrock, C., Kalas, R., & Dulcan, M. K. (1984). *The NIMH Diagnostic Interview Schedule for Children (DISC): Development, reliability, and comparisons between clinical and lay interviews.*

Cummings, A. L., Martin, J., Hallberg, E., & Slemon, A. (1992). Memory for therapeutic events, session effectivieness, and working alliance in short-term counseling. *Journal of Counseling psychoogy, 39,* 306-312.

Cummings, A. L., Slemon, A. G., & Hallberg, E. T. (1993). Session evaluation and recall of important events as a function of counselor experience. *Journal of Counseling Psychology, 40,* 156-165.

Danish, S. J., & D'Augelli, A. R. (1976). Rationale and implementation of a training program for paraprofessionals. *Professional Psychology, 7,* 38-46.

Deutsch, F., & Murphy, W. F. (1955a). *The clinical interview, Vol. 1: Diagnosis.* New York: International Universities Press.

Deutsch, F., & Murphy, W. F. (1955b). *The clinical interview. Vol. 2: Therapy.* New York: International Universities Press.

Digiacomo, J. N. (1982). Three-way interviews and psychiatric training. *Hospital and Community Psychiatry, 33,* 287-291.

Dohrenwend, B. P., Shrout, P. E., Egri, G., & Mendelsohn, F. S. (1980). Nonspecific psychological distress and other dimensions of psychopathology. *Archives of General Psychiatry, 37,* 1229-1236.

Donnelly, J., Rosenberg, M., & Fleeson, W. P. (1970). The evolution of the mental status—Past and future. *American Journal of Psychiatry, 126,* 997-1002.

Dowrick, P. W., & Biggs, S. J. (Eds.). (1983). *Using video: Psychological and social applications.* New York: Wiley.

Edelbrock, C., Costello, A. J., Dulcan, M. K., Kalas, R., & Conover, N. C. (1985). Age differences in the reliability of the psychiatric interview of the child. *Child Development, 56,* 265-275.

Edinger, J. A., & Patterson, M. L. (1983). Nonverbal involvement and social control. *Psychological Bulletin, 93,* 30-56.

Egan, G. (1975). *The skilled helper: A model for systematic helping and interpersonal relating.* Belmont, CA: Brooks/Cole.

Eisenthal, S., Koopman, C., & Lazare, A. (1983). Process analysis of two dimensions of the negotiated approach in relation to satisfaction in the initial interview. *Journal of Nervous and Mental Disease, 171,* 49-54.

Eisenthal, S., & Lazare, A. (1977a). Evaluation of the initial interview in a walk-in clinic. *Journal of Nervous and Mental Disease, 164,* 30-35.

Eisenthal, S., & Lazare, A. (1977b). Expression of patient's requests in the initial interview. *Psychological Reports, 40,* 131-138.

Ekman, P. (1985). *Telling lies—Clues to deceit in the marketplace, politics, and marriage.* New York: W.W. Norton & Company.

Ekman, P., & Friesen, W. V. (1974). Detecting deception from the body or face. *Journal of Personality and Social Psychology, 29,* 288-298.

Ekman, P. & Rosenberg, E. (1998). *What the face reveals: Basic and applied studies of spontaneous expression using the facial action coding system.* New York: Oxford University Press.

Elliott, R. (1986). Interpersonal Process Recall (IPR) as a psychotherapy process research method. In L. S. Greenberg (Ed.), *The psychotherapeutic process: A research handbook.* New York: Guilford Press.

Elliott, R., Filipovich, H., Harrigan, L., Gaynor, J., Reimschuessel, C., & Zapadka, J. K. (1982). Measuring response empathy: The development of a multicomponent rating scale. *Journal of Counseling Psychology, 29,* 379-387.

Elliott, R., Stiles, W. B., Mahrer, A. R., Hill, C. E., Friedlander, M. L., & Margison, F. R. (1987). Primary therapist response modes: Comparison of six rating systems. *Journal of Consulting and Clinical Psychology, 55,* 218-223.

Endicott, J., & Spitzer, R. L. (1978). A diagnostic interview: The schedule for affective disorders and schizophrenia. *Archives of General Psychiatry, 35,* 837-844.

Epperson, D. L., Bushway, D. J., & Warman, R. E. (1983). Client self-terminations after one counseling session: Effects of problem recognition, counselor gender, and counselor experience. *Journal of Counseling Psychology, 30,* 307-315.

Errek, H. K., & Randolpf, D. L. (1982). Effects of discussion and role-play activities in the acquisition of consultant interview skills. *Journal of Counseling Psychology, 29,* 304-308.

Feighner, J. P., Robins, E., Guze, S. B., Woodruff, R. A., Winokur, G., & Munoz, R. (1972). Diagnostic criteria for use in psychiatric research. *Archives of General Psychiatry, 26,* 57-63.

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "Mini mental state:" A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research, 12,* 189-198.

Frances, A., First, M. B., & Ross, R. (1995). Getting up to speed on DSM-IV. *Journal of Practical Psychiatry and Behavioral Health,* 1, 2-9.

Friedlander, M. L. (1982). Counseling discourse as a speech event: Revision and extension of the Hill Counselor Verbal Response Category System. *Journal of Counseling Psychology, 29,* 425-429.

Frontman, K. C., & Kunkel, M. A. (1994). A grounded theory of counselors' construal of success in the initial session. *Journal of Counseling Psychology,* 41, 492-499.

Gill, M., Newman, R., & Redlich, F. C. (1954). *The initial interview in psychiatric practice.* New York: International Universities Press.

Gladstein, G. A. (1983). Understanding empathy: Integrating counseling, developmental, and social psychology perspectives. *Journal of Counseling Psychology, 30,* 467-482.

Goodman, G., & Dooley, D. A. (1976). A framework for help-intended communication. *Psychotherapy, Theory, Research, and Practice, 13,* 106-117.

Grinder, J., & Bandler, R. (1975). *The structure of magic I.* Palo Alto, CA: Science & Behavior Books.

Hackney, H., & Nye, S. (1973). *Counseling strategies and outcomes.* Englewood Cliffs, NJ: Prentice-Hall.

Hall, C. S., & Lindzey, G. (1978). *Theories of personality.* New York: Wiley.

Hall, E. T. (1966). *The hidden dimension.* New York: Doubleday.

Hannay, D. R. (1980). Teaching interviewing with simulated patients. *Medical Education, 12,* 246-248.

Harmon, J. I. (1986). Relations among components of the empathic process. *Journal of Counseling Psychology, 33,* 371-376.

Harper, R. G., Wiens, A. N., & Matarazzo, J. D. (1978). *Nonverbal communication: The state of the art.* New York: Wiley.

Havens, L. (1978). Explorations in the uses of language in psychotherapy: Simple empathic statements. *Psychiatry, 41,* 336-345.

Havens, L. (1979). Explorations in the uses of language in psychotherapy: Complex empathic statements. *Psychiatry, 42,* 40-48.

Havens, L. (1980). Experience in the uses of language in psychotherapy: Counterprojective statements. *Contemporary Psychoanalysis,* 16, 53-67.

Hedlund, J. L., & Vieweg, B. W. (1981). Structured psychiatric interviews: A comparative review. *Journal of Operational Psychiatry, 12,* 3 9-67.

Heilbrun, A. B. (1974). Interviewer style, client satisfaction, and premature termination following the initial counseling contact. *Journal of Counseling Psychology, 21,* 346-350.

Helmchen, H. (1975). Schizophrenia: Diagnostic concepts in the ICD-8. In M. H. Lader (Ed.), Studies in schizophrenia. *British Journal of Psychiatry, Special Publication, 10,* 10-18.

Helms, J., Carter, E., & Robert, T. (1991). Relationships of white and black racial identity attitudes and demographic similarity to counselor preference. *Journal of Counseling Psychology, 38,* 446-45.

Helzer, J. E. (1983). Standardized interviews in psychiatry. *Psychiatric Developments, 2,* 161-178.

Herjanic, B., & Campbell, W. (1977). Differentiating psychiatrically disturbed children on the basis of a structured interview. *Journal of Abnormal Child Psychology, 5,* 127-134.

Herjanic, B., & Reich, W. (1982). Development of a structured psychiatric interview for children: Agreement between child and parent on individual symp-

toms. *Journal of Abnormal Child Psychology, 10,* 307-324.

Hersen, M., & Turner, S. M. (1985). *Diagnostic interviewing.* New York: Plenum Press.

Hill, C. E. (1975). Sex of client and sex experience level of counselor. *Journal of Counseling Psychology, 22,* 6-11.

Hill, C. E. (1978). Development of a counselor verbal response category system. *Journal of Counseling Psychology, 25,* 461-468.

Hill, C. E. (1990). Exploratory in-session process research in individual therapy: A review. *Journal of Consulting and Clinical Psychology,* 58, 288-294.

Hill, C. E., & Corbett, M. M. (1993). A perspective on the history of process and outcome research in counseling psychology. *Journal of Counseling Psychology, 40,* 3-24.

Hill, C. E. & Stephany, A. (1990). Relation of nonverbal behavior to client reactions. *Journal of Counseling Psychology, 37,* 22-26.

Hutter, M. J., Dungy, C. I., Zakus, G. E., Moore, V. J., Ott, J. E., & Favret, A. (1977). Interviewing skills: A comprehensive approach to teaching and evaluation. *Journal of Medical Education,* 52, 328-333.

Ivey, A. E. (1971). *Microcounseling: Innovations in interviewing training.* Springfield, IL: Charles C Thomas.

Jackson, M. G., & Pinkerton, R. R. (1983). Videotape teaching in family practice residencies. Jour*nal of Medical Education, 58,* 434-435.

Jarrett, F. J., Waldron, J. J., Burra, P., & Handforth, J. R. (1972). Measuring interviewing skill-The Queen's University Interviewer Rating Scale (QUIRS). *Canadian Psychiatric Association Journal, 17,* 183-188.

Jaynes, S., Charles, E., Kass, F., & Holzman, S. (1979). Clinical supervision of the initial interview: Effects on patient care. *American Journal of Psychiatry, 136,* 1454-1457.

Kagan, N. (1975). *Interpersonal process recall: A method of influencing human interaction.* (Available from N. Kagan, 434 Erickson Hall, College of Education, MSU, East Lansing, Michigan 48824.)

Kaplan, H., Freedman, A., & Sadock, B. (Eds.). (1980). *Comprehensive textbook of psychiatry* (3rd ed.). Baltimore: Williams & Wilkins.

Kato, M. (1977). *Multiaxial diagnosis in adult psychiatry.* Paper presented at the Sixth World Congress of Psychiatry, Honolulu, HI.

Keating, A., & Fretz, B. R. (1990). Christians' anticipations about counselors in response to counselor's descriptions. *Journal of Counseling Psychology, 37,* 293-296.

Kivlighan, Jr., D. M. (1989). Changes in counselor intentions and response modes and in client reactions and session evaluation after training. *Journal of Counseling Psychology, 36,* 471-476.

Kovacs, M. (1983). *The Interview Schedule for Children (ISC): Interrater and parent-child agreement.*

LaCrosse, M. B. (1975). Nonverbal behavior and perceived counselor attractiveness and persuasiveness. *Journal of Counseling Psychology, 6,* 563-566.

Lambert, M. J., DeJulio, S. J., & Stein, D. M. (1978). Therapist interpersonal skills: Process, outcome, methodological considerations, and recommendations for future research. *Psychological Bulletin, 85,* 467-489.

Langsley, D. G., & Hollender, M. H. (1982). The definition of a psychiatrist. *American Journal of Psychiatry, 139,* 81-85.

Langsley, D. C., & Yager, J. (1988). The definition of a psychiatrist: Eight years later. *American Journal of Psychiatry,* 145, 469-475.

Levinson, W., & Roter, D. (1993). The effects of two continuing medical education programs on communication skills of practicing primary care physicians. *Journal of General Internal Medicine, 8,* 318-324.

Liston, E. H., & Yager, J. (1982). Assessment of clinical skills in psychiatry. In J. Yager (Ed.), *Teaching psychiatry and behavioral science.* New York: Grune & Stratton.

Liston, E. H., Yager, J., & Strauss, G. D. (1981). Assessment of psychotherapy skills: The problem of interrater agreement. *American Journal of Psychiatry, 138,* 1069-1074.

Lonborg, S. D., Daniels, J. A., Hammond, S. G., Houghton-Wenger, B., & Brace, L. J. (1991). Counselor and client verbal response mode changes during initial counseling sessions. *Journal of Counseling Psychology, 38,* 394-400.

MacKinnon, R. A., & Michels, R. (1971). *The psychiatric interview in clinical practice.* Philadelphia: W.B. Saunders.

Maguire, P., Roe, P., Goldberg, D., Jones, S., Hyde, C., & O'Dowd, T. (1978). The value of feedback in teaching interviewing skills to medical students. *Psychological Medicine, 8,* 695-704.

Mahalik, J. R. (1994). Development of the client resistance scale. *Journal of Counseling Psychology,* 41, 58-68.

Mallinckrodt, B. (1993). Session impact, working alliance, and treatment outcome in brief counseling. *Journal of Counseling Psychology, 40,* 25-32.

Margulies, A. (1984). Toward empathy: The uses of wonder. *American Journal of Psychiatry, 141,* 1025-1033.

Margulies, A., & Havens, L. (1981). The initial encounter: What to do first? *American Journal of Psychiatry, 138,* 421-428.

Marmar, C. R., Horowitz, M. J., Weiss, D. S., & Marziali, E. (1986). The development of the Therapeutic Alliance Rating System. In L. S. Greenberg & W. M. Pinsof (Eds.), *The psychotherapeutic process—A research handbook.* New York: Guilford Press.

Matarazzo, J. D., & Wiens, A. N. (1972). *The interview: Research on its anatomy and structure.* Chicago: Aldine-Atherton.

McKee, K., & Smouse, A. D. (1983). Clients' perceptions of counselor expertness, attractiveness, and trustworthiness: Initial impact of counselor status and weight. *Journal of Counseling Psychology, 30,* 332-338.

Meyer, A. (1951). *The collected papers of Adolf Meyer* (Vol. 3). E.E. Winters (Ed.). Baltimore: John Hopkins Press.

Mezzich, J. E. (1985). Multiaxial diagnostic systems in psychiatry. In H. I. Kaplan & B. J. Sadock (Eds.), *The comprehensive textbook of psychiatry* (4th ed.). Baltimore: Williams & Wilkins.

Mezzich, J. E., Dow, J. T., Ganguli, R., Munetz, J. R., & Zettler-Segal, M. (1986). Computerized initial and discharge evaluations. In J. E. Mezzich (Ed.), *Clinical care and information systems in psychiatry.* Washington, DC: American Psychiatric Press.

Mezzich, J. E., Dow, J. T., Rich, C. L., Costello, A. J., & Himmelhoch, J. M. (1981). Developing an efficient clinical information system for a comprehensive psychiatric institute. II. Initial evaluation form. *Behavior Research Methods & Instrumentation, 13,* 464-478.

Morrison, J. (1993). *The first interview: A guide for clinicians.* New York: Guilford Press.

Othmer, E., & Othmer, S. C. (1989). *The clinical interview using DSM-III-R.* Washington, DC: American Psychiatric Press, Inc.

Othmer, E., & Othmer, S. C. (1994). *The clinical interview using DSM-IV, Vol. 1: Fundamentals.* Washington, DC: American Psychiatric Press, Inc.

Othmer, E., & Othmer, S. C. (1994). *The clinical interview using the DSM-IV, Vol. 2: The difficult patient.* Washington, DC: American Psychiatric Press, Inc.

Ottosson, J. O., & Perris, C. (1973). Multidimensional classification of mental disorders. *Psychological Medicine, 3,* 238-243.

Ovadia, A. B., Yager, J., & Heinrich, R. L. (1982). Assessment of medical interview skills. In J. Yager (Ed.), *Teaching psychiatry and behavioral science.* New York: Grune & Stratton.

Paradise, L. V., Conway, B. S., & Zweig, J. (1986). Effects of expert and referent influence, physical attractiveness, and gender on perceptions of counselor attributes. *Journal of Counseling Psychology, 33,* 16-22.

Pascal, G. R. (1983). *The practical art of diagnostic interviewing.* Homewood, IL: Dow Jones-Irwin.

Paurohit, N., Dowd, E. T., & Cottingham, H. F. (1982). The role of verbal and nonverbal cues in the formation of first impressions of black and white counselors. *Journal of Counseling Psychology, 4,* 371-378.

Pinel, P. (1988). A *treatise on insanity.* Birmingham, AL: Gryphon Editions. (Original work published 1806)

Puig-Antich, J., & Chambers, W. (1978). *The schedule for affective disorders and schizophrenia for school-aged children.* Unpublished manuscript, New York State Psychiatric Institute, New York, NY.

Reik, T. (1952). *Listening with the third ear.* New York: Farrar, Strauss.

Richardson, S. A., Dohrenwend, B. S., & Klein, D. (1965). *Interviewing: Its forms and functions.* New York: Basic Books.

Robins, L. N., Helzer, J. E., Croughan, J., & Ratcliff, K. S. (1981). National Institute of Mental Health Diagnostic Interview Schedule—Its history, characteristics, and validity. *Archives of General Psychiatry, 10,* 41-61.

Robinson, D. J., & Chapman, B. (1997). *Brain calipers: A guide to a successful mental status exam. Gratiot, MI: Rapid Psychler Press.*

Robinson, F. R. (1950). *Principles and procedures in student counseling.* New York: Harper.

Rogers, C. R. (1951). *Client-centered therapy.* Boston: Houghton-Mifflin.

Rogers, C. R. (1959). A theory of therapy, personality and interpersonal relationships as developed in the client-centered framework. In S. Koch (Ed.), *Psychology: A study of a science: Formulations of the person and the social context* (Vol. 3). New York: McGraw-Hill.

Rosenthal, R., Hall, J. A., DiMatteo, M. R., Rogers, P. L., & Archer, D. (1979). *Sensitivity to nonverbal communication: The PONS test.* Baltimore: Johns Hopkins University Press.

Rutter, M., Cox, A., Egert, S., Holbrook, D., & Everitt, B. (1981). Psychiatric interviewing techniques IV. Experimental study: Four contrasting styles. *British Journal of Psychiatry, 138,* 456-465.

Rutter, M., & Graham, P. (1968). The reliability and validity of the psychiatric assessment of the child: Interview with the child. *British Journal of Psychiatry, 11,* 563-579.

Rutter, M., Shaffer, D., & Shepherd, M. (1975). A *multiaxial classification of child psychiatric disorders.* Geneva, Switzerland: World Health Organization.

Sanson-Fisher, R., Fairbairn, S., & Maguire, P. (1981). Teaching skills in communication to medical student—A critical review of the methodology. *Medical Education, 15*, 33-37.

Sanson-Fisher, R. W., & Martin, C. J. (1981). Standardized interviews in psychiatry: Issues of reliability. *British Journal of Psychiatry, 139*, 138-143.

Scheflen, A. E. (1972). *Body, language and social order.* Englewood Cliffs, NJ: Prentice-Hall.

Shea, S. C. (1988). *Psychiatric interviewing: The art of understanding.* Philadelphia: W.B. Saunders.

Shea, S. C. (1995). Psychiatric interviewing. In M. H. Sachs, W. H. Sledge, & C. Warren (Eds.), *Core readings in psychiatry.* Washington, DC: American Psychiatric Press, Inc.

Shea, S. C. (1998). The chronological assessment of suicide events: A practical interviewing strategy for the elicitation of suicidal ideation. *Journal of Clinical Psychiatry, 59,* (Suppl.), 58-72.

Shea, S. C. (1998). *Psychiatric interviewing: The art of understanding* (2nd ed.). Philadelphia: W.B. Saunders Company.

Shea, S. C. (1999). *The practical art of suicide assessment.* New York: John Wiley & Sons, Inc.

Shea, S. C., & Mezzich, J. E. (1988). Contemporary psychiatric interviewing: New directions for training. *Psychiatry: Interpersonal and Biological Processes, 51*, 385-397.

Shea, S. C., Mezzich, J. E., Bohon, S., & Zeiders, A. (1989). A comprehensive and individualized psychiatric interviewing training program. *Academic Psychiatry, 13*, 61-72.

Sheehan, D., & Lecrubier, Y., et al. (1999). Appendix 1, The Mini-International Neuropsychiatric Interview (M.I.N.I.). *Journal of Clinical Psychiatry, 60,* (Suppl. 18) 39-62.

Siassi, I. (1984). Psychiatric interview and mental status examination. In G. Goldstein & M. Hersen (Eds.). *Handbook of psychological assessment.* New York: Pergamon Press.

Siegel, J. C., & Sell, J. M. (1978). Effects of objective evidence of expertness and nonverbal behavior on client-perceived expertness. *Journal of Counseling Psychology, 25*, 188-192.

Smith, C. K., Hadac, R. R., & Leversee, J. H. (1980). Evaluating the effects of a medical interviewing course taught at multiple locations. *Journal of Medical Education, 55,* 792-794.

Smith-Hanen, S. S. (1977). Effects of nonverbal behaviors on judged levels of counselor warmth and empathy. *Journal of Counseling Psychology, 24*, 87-91.

Snyder, W. U. (1945). An investigation of the nature of nondirective psychotherapy. *Journal of General Psychology, 33*, 193-223.

Snyder, W. U. (1963). *Dependency in psychotherapy: A casebook.* New York: Macmillan.

Sommers-Flanagan, R., & Sommers-Flanagan, J. (1999). *Clinical interviewing* (2nd ed.). New York: John Wiley & Sons, Inc.

Spitzer, R. L., Endicott, J., & Robins, E. (1978). Research diagnostic criteria. *Archives of General Psychiatry, 35,* 773-782.

Spitzer, R. L., & Williams, J. B. W. (1983). *DSM-III SCID Manual.* New York: New York State Psychiatric Institute, Biometrics Research Department.

Spooner, S. E., & Stone, S. C. (1977). Maintenance of specific counseling skills over time. *Journal of Counseling Psychology, 24,* 66-7 1.

Srole, L., Langer, T. S., Michael, S. T., Opler, M. K., & Rennie, T. A. C. (1962). *Mental health in the metropolis: The Midtown Manhattan Study* (Vol. 1). New York: McGraw-Hill.

Stein, S. P., Karasu, T. B., Charles, E. S., & Buckley, P. J. (1975). Supervision of the initial interview. *Archives of General Psychiatry, 32*, 265-268.

Stiles, W. B. (1978). Verbal response modes and dimensions of interpersonal roles: A method of discourse analysis. *Journal of Personality and Social Psychology, 7*, 693-703.

Stiles, W. B. (1984). Measurement of the impact of psychotherapy sessions. *Journal of Consulting and Clinical Psychology, 48*, 176-185.

Strong, S. R. (1968). Counseling: An interpersonal influence process. *Journal of Counseling Psychology, 15*, 215-224.

Strong, S. R., & Schmidt, L. D. (1970). Expertness and influence in counseling. *Journal of Counseling Psychology, 17*, 81-87.

Strong, S. R., Taylor, R. G., Bratton, J. C., & Loper, R. (1971). Nonverbal behavior and perceived counselor characteristics. *Journal of Counseling Psychology, 18,* 554-561.

Strub, R. L., & Black, W. W. (1979). *The mental status examination in neurology.* Philadelphia: F.A. Davis.

Strupp, H. H. (1960). *Psychotherapists in action: Explorations of the therapist's contribution to the treatment process.* New York: Grune & Stratton.

Sue, D. W. (1981). *Counseling the culturally different.* New York: Wiley.

Sue, D. W., & Sue, D. (1977). Barriers to effective cross-cultural counseling. *Journal of Counseling Psychology, 24,* 420-429.

Sullivan, H. S. (1970). *The psychiatric interview.* New York: W. W. Norton Company.

Tepper, D. T., Jr., & Haase, R. F. (1978). Verbal and nonverbal communication of facilitative conditions. *Journal of Counseling Psychology, 25*, 35-44.

Tomlinson-Clarke, S. & Cheatham, H. E. (1993). Counselor and client ethnicity and counselor intake judgments. *Journal of Counseling Psychology, 40*, 267-270.

Truax, C. B., & Carkhuff, R. R. (1967). *Toward effective counseling and psychotherapy: Training and practice.* Chicago: Aldine.

Tryon, G. S. (1985). The engagement quotient: One index of a basic counseling task. *Journal of College Student Personnel, 26*, 351-354.

Tryon, G. S. (1990). Session depth and smoothness in relation to the concept of engagement in counseling. *Journal of Counseling Psychology, 37*, 248-253.

Trzepacz, P. T. & Bakev, R. W. (1993). *The psychiatric mental status.* New York: Oxford University Press.

Vargas, A. M., & Borkowski, J. G. (1982). Physical attractiveness and counseling skills. *Journal of Counseling Psychology, 29*, 246-255.

von Cranach, M. (1977). *Categorical vs. multiaxial classification.* Paper presented at the Seventh World Congress of Psychiatry, Honolulu, HI.

Waldron, J. (1973). Teaching communication skills in medical school. *American Journal of Psychiatry, 130*, 579-591.

Ward, N. G., & Stein, L. (1975). Reducing emotional distance: A new method of teaching interviewing skills. *Journal of Medical Education, 50*, 605-614.

Wenegrat, A. (1974). A factor analytic study of the Truax Accurate Empathy Scale. *Psychotherapy: Theory, Research and Practice, 11*, 48-5 1.

Whalen, C. K., & Flowers, J. V. (1977). Effects of role and gender mix on verbal communication modes. *Journal of Counseling Psychology, 24*, 281-287.

Whitehorn, J. C. (1944). Guide to interviewing and clinical personality study. *Archives of Neurology and Psychiatry, 52*, 197-216.

Wiens, A. N. (1983). The assessment interview. In I. B. Weiner (Ed.), *Clinical methods in psychology*. New York: Wiley.

Wing, J. K., Cooper, J. E., & Sartorius, N. (1974). *Measurement and classification of psychiatric symptoms.* Cambridge, MA: Cambridge University Press.

Wing, J. K., Nixon, J. M., Mann, S. A., & Leff, J. P. (1977). Reliability of the PSE (ninth edition) used in a population study. *Psychological Medicine, 7*, 505-516.

Wing, L. (1970). Observations on the psychiatric section of the International Classification of Diseases and the British Glossary of Mental Disorders. *Psychological Medicine, 1*, 79-85.

Zimmer, J. M., & Anderson, S. (1968). Dimensions of positive regard and empathy. *Journal of Counseling Psychology, 15*, 417-426.

CHAPTER 14

STRUCTURED INTERVIEWS FOR CHILDREN AND ADOLESCENTS

Craig Edelbrock
Amy Bohnert

INTRODUCTION

Interviewing is a universal method of assessment in all areas of mental-health research and clinical practice. Face-to-face interviewing of people is a natural, and arguably indispensable, means of gaining information about emotional and behavioral functioning, physical health, and social relationships—both past and present. Part of the appeal of interviewing is that it is a "low tech" assessment method that is adaptable to many different purposes. It is highly flexible and can be quickly adapted to a broad range of target phenomena, or alternatively to probe in-depth in a specific area. Interviewing provides unparalleled ability to insure that respondents understand questions, to evoke rich and detailed examples, and to document chronicity of events.

Compared to other assessment methods, such as psychological testing and direct observation, interviewing can be efficient and cost-effective in terms of professional time and training. Interviewing is also usually readily accepted by both research subjects and clinical clients and is typically expected to be the "default" assessment technique. Interviewing is, of course, ubiquitous as a means not only of obtaining assessment information, but of "breaking the ice" and establishing rapport between the interviewer and interviewee. It also represents a potential way of obtaining information from children, including those too young to complete paper-and-pencil questionnaires.

As universal as interviewing is, it is perhaps one of the least rigorous and most fallible assessment procedures. The flexibility of most interviews is a double-edged sword, allowing us to adapt assessments to individual respondents, but opening the door to numerous uncontrolled sources of variation in the assessment process. Simply put, interviewers differ widely in *what they ask* and *how they ask it*. Given free reign, interviewers choose different lines and styles of questioning. They cover different material, in different ways. They project different verbal and nonverbal cues to the respondent, not to mention the fact that interviewers differ in how they rate, record, interpret, and combine interviewees' responses. Such broad variations in content, style, level of detail, and coverage make interviewing highly suspect from a measurement point of view. In the language of measurement, interviewing is prone to high "information variance"—variability in what information is *sought* and *elicited* from respondents—which is blamed as a major cause of low reliability in the assessment and diagnostic process (see Matarazzo, 1983).

A simple experiment effectively illustrates the problem: Suppose there was a pool of subjects who were absolutely identical in every way. Clinical interviews would not elicit identical information from such clones. Different interviewers would ask

different questions in different ways, and would rate and record the responses idiosyncratically. Moreover, interviews conducted by the *same interviewer* might yield quite different information due to variations in interviewing style and content from one clone to the next. If this hypothetical example were a study of diagnostic reliability, the *subject variance* would eliminated, since all subjects are identical. The *criterion variance*—variability due to use of different diagnostic criteria—could also be eliminated if one diagnostic system were used. But *information variance* would remain as a major threat to reliability. Given the freedom of unstructured interviews, differences in the information obtained would undoubtedly arise and the reliability of diagnoses would be less than perfect.

How can the advantages of interviewing be maintained while making it more scientifically rigorous as a measurement technique? The answer, at least in part, involves standardizing the interview process. Standardizing in this sense means imposing some structure—literally limiting variability in the question-answer interactions between interviewer and respondent. This is accomplished in three ways. The first is by *defining* the phenomena to be assessed. Differences between interviewers can be reduced considerably by establishing what the interview does (and does not) cover. Second is by *limiting* to some degree the order and wording of questions to be asked. Individual differences between interviewers are thus further reduced by restricting *how* the target phenomena are covered. Third is by *standardizing* how responses are rated, recorded, combined, and interpreted. Structuring the interview process in these ways addresses both the criterion variance and information variance inherent in any assessment process. There is less criterion variance in structured versus unstructured interviews because the range and coverage of the interview is set, and in the case of diagnostic interviews, the diagnostic system and specific diagnostic criteria are specified. There is less information variance, as well, because the order and wording of items is predetermined and there is a standard format for translating interviewee's responses into objective data.

Numerous interview schedules have been developed, beginning with those designed for adults to report about themselves. Researchers in the child areas were quick to follow suit and develop interview schedules for child and adolescent populations. Many of these interviews were spin-offs or downward extrapolations of adult interviews. Many interview schedules have parallel formats for interviewing adults (usually parents) about children, and a separate format for direct interview of children themselves. Viewing the child as a valuable source of information about themselves was a revolutionary change in assessment theory and practice—and one that created numerous challenges.

The assessment of child psychopathology has traditionally depended upon reports and ratings by adults, particularly parents. This makes sense because parents are the most common instigators of child mental-health referrals and they are almost always involved in the assessment process. Parents' perceptions are often crucial in the implementation of child interventions and the evaluation of child outcomes. For many decades, direct interview of the child was not considered a useful endeavor. Psychodynamic theories postulated that children lack insight into their own problems. Child developmentalists argued that young children are not cognitively mature enough to understand life history or symptom-oriented interviews. These assumptions have been increasingly questioned, and numerous interview schedules have been developed for directly interviewing the child—not always with successful results. The challenges of interviewing children, however, do create theoretical and practical problems—many of which remain to be solved.

Historical Foundations

The historical development of structured interviews for children and adolescents owes much to precedents set in the adult area—especially the epoch-making development of the Diagnostic Interview Schedule (DIS) and the Schedule for Affective Disorders and Schizophrenia (SADS). In fact, two early interviews for children—the Diagnostic Interview for Children and Adolescents (DICA) and the Kiddie-SADS (K-SADS)—inherited much of their format, style, and mode of administration to their respective adult forerunners. But interviewing children has a history of its own and appears to trace out two distinct lines of influence: one *diagnostic* and the other *descriptive*.

The diagnostic line of development corresponds to the emergence of more differentiated taxonomies of childhood disorders. Prior to 1980 and the publication of the third edition of the Diagnostic and Statistical Manual (DSM-III) of the American

Psychiatric Association (APA) (1980) there was little need for diagnostic interview schedules that provided precise, detailed, and reliable assessments of child psychopathology. During the era of the first edition of the DSM (APA, 1956) there were only two diagnostic categories for children: Adjustment Reaction and Childhood Schizophrenia. Adult diagnoses could be applied to children, but the vast majority of children seen in psychiatric clinics were either undiagnosed or were labeled adjustment reactions (Rosen, Bahn, & Kramer, 1964). More differentiated taxonomies of childhood disorders were provided by the Group for the Advancement of Psychiatry (GAP) (1966) and the second edition of the DSM (APA, 1968), but both systems lacked explicit diagnostic criteria and operational assessment procedures. Not surprisingly, the reliability of both systems was mediocre (Freeman, 1971; Sandifer, Pettus, & Quade, 1964; Tarter, Templer, & Hardy, 1975).

In 1980, however, the DSM-III provided a differentiated taxonomy of "Disorders Usually First Evident in Infancy, Childhood, or Adolescence" that had more explicit diagnostic criteria. The need for more reliable and valid ways of assessing diagnostic criteria was a primary stimulus for the development of structured interview schedules for children and adolescents. More impetus was gained from the successes in the adult area. Although adult psychiatric disorders had explicit diagnostic criteria, refined through decades of trial-and-error tinkering, reliability of adult diagnoses was too low for research purposes, such as epidemiologic surveys and clinical trials. This prompted the development of structured interview schedules, such as the Diagnostic Interview Schedule (Robins, Helzer, Croughan, & Ratcliff, 1981) and the Schedule for Affective Disorders and Schizophrenia (Endicott & Spitzer, 1978), which substantially reduced information variance and boosted diagnostic reliability (see Matarazzo, 1983). Researchers interested in child and adolescent psychopathology were quick to follow suit and, in fact, many interview schedules for children are downward extrapolations of adult interviews.

Apart from diagnostic purposes, there had long been a need for obtaining descriptive data on children's emotional, behavioral, and social problems, but standardized assessment procedures were lacking. In the spring of 1955 Lapouse and Monk (1958) undertook a survey to determine the prevalence and patterning of problem behaviors in a community sample. A standard format was used for interviewing mothers about their children's behavior. This had the obvious advantage of yielding more objective data than an unstructured clinical interview and it insured that direct comparisons could be made between subjects assessed by different interviewers. Moreover, the goal was to *describe* children's behavioral problems, rather than to detect prespecified syndromes and disorders. The unresolved questions about the existence and definition of specific childhood disorders were thus circumvented.

Interviews were conducted with 482 mothers and 193 children ages 6 to 12 years. The interview comprised 200 questions and took about 90 minutes to complete. Most items had a yes/no response format, but some involved rating the frequency or intensity of the target behavior.

Several findings from this landmark study were replicated by later researchers. Reinterviews with mothers, for example, indicated high test-retest reliability for items such as thumb sucking, bed wetting, and stuttering. But reliability was low for items such as fears and worries and for items requiring precise estimates of frequency (e.g., number of temper tantrums). Mother-child agreement was low for most behaviors, but was higher for behaviors such as bed-wetting, temper tantrums, and biting fingernails. Mothers tended to report more behavior problems that are irksome to adults (e.g., bed-wetting, restlessness, overactivity), whereas children tended to report more problems that are distressing to themselves (e.g., fears, worries, nightmares). These findings have been replicated many times over the years.

In another pioneering effort, Rutter and Graham (1968) developed structured procedures for directly interviewing the *child.* This was a major departure from the prevailing thought and clinical practice of the time. In clinical settings, direct interview of the child was used primarily as a *therapeutic* rather than as an assessment technique. Moreover, the assessment uses of the interview were largely restricted to uncovering unconscious wishes, fears, conflicts, and fantasies (see Group for the Advancement of Psychiatry, 1957). In contrast, the interview procedures developed by Rutter and Graham were aimed at descriptive assessment of the child's emotional, behavioral, and social functioning and were based on direct questioning of the child (and parent).

The parent and child versions of this interview schedule differ somewhat, but parallel one another in content and rating procedure. Both are

semi-structured interviews designed for clinically trained interviewers. The exact order and wording of questions is not prescribed. Instead, areas of functioning, such as school performance, activities, and friendships are listed, along with symptom areas such as antisocial behavior, anxiety, and depression. The parent version has more detail as to duration, severity, action taken, presumed cause, and expected course of problems reported. The rating of many items requires clinical judgment. Parent and child are interviewed separately. After each interview, the interviewer rates the child's mental status and determines if the child has *no psychiatric impairment, some impairment,* or *definite or marked impairment.*

Two findings from this early work have been replicated by later studies. First, higher reliabilities were obtained for ratings of global psychiatric status than for ratings of specific symptoms, syndromes, and disorders. Rutter and Graham (1968), for example, found high interrater reliability (r=.84) for the overall ratings of psychiatric impairment based on separate interviews of the child by different interviewers. But reliabilities were mediocre for items pertaining to attention and hyperactivity (r=.61), social relations (r=.64), and anxiety and depression (r=.30). Second, as illustrated by these results, reliabilities were generally higher for problems, such as hyperactivity and antisocial behavior, than for problems such as fears, anxiety, and depression.

Recent Trends

Structured interview schedules for children and adolescents have proliferated in the last 20 years as the need for descriptive and diagnostic assessment tools increased. There are now many well-developed interview schedules that are widely used in research and to a lesser extent in clinical practice. A major trend has been toward increasing specialization of interview schedules. Specialization of *purpose*, for example, has resulted in different interview schedules for screening nonreferred populations versus differential diagnosis of identified cases. Specialization in *age range* has resulted in different interview schedules for preschool-aged children, grade schoolers, and adolescents. Interview schedules have also become more specialized in *coverage and focus*. Most cover a broad range of symptoms and behaviors, but some are focused on specific syndromes and disorders, such as childhood depression. Lastly, there has been increasing specialization in *interviewer training and qualifications*. Some are designed for clinically sophisticated interviewers, whereas others are designed for lay interviewers having only interview-specific training.

Summary

Development of structured clinical interviews for children and adolescents can be traced along two historical lines. First, emergence of differentiated taxonomies of childhood disorders with more explicit diagnostic criteria necessitated more accurate, precise, comprehensive, and reliable diagnostic interviewing procedures. Diagnostic interview schedules were therefore developed for purposes of differential diagnosis of children already identified as cases. Second, standard interview procedures for assessing children's emotional, behavioral, and social functioning were needed for descriptive, developmental, and epidemiological studies. Interview schedules aimed at obtaining descriptive information about children's functioning were developed primarily for use with nonreferred populations.

The pioneering studies by Lapouse and Monk (1958) and Rutter and Graham (1968) broke new ground and introduced several innovations in interviewing, including (a) structuring the content of the interview around specific target phenomena, (b) providing prespecified formats for rating and recording responses, (c) focusing on the child's functioning rather than psychodynamic states, (d) directly interviewing the child, and (e) using parallel interview schedules for parent and child.

Research in the past 20 years has amplified and improved upon these methodological innovations. A broad range of interview schedules for children and adolescents is now available and these schedules are widely used in research. These interview schedules have become more specialized in purpose, age range, coverage, and training requirements. Additionally, the child is now viewed as a potentially important source of information, so interview schedules have been developed specifically for interviewing children about their own functioning.

THEORETICAL UNDERPINNINGS

Theory has played little part in the development of descriptive interviews. The lack of a consensual theory of child psychopathology leaves researchers with little guidance about what phenomena are important to assess. Not surprisingly, interview items are selected primarily on the basis of face validity—not theoretical importance.

Any taxonomy is an implicit theory about what disorders exist and how to diagnose them. As such, diagnostic interviews are operationalizations of the prevailing taxonomic theory—which in the United States has been the DSM. The DSM is evolving—or perhaps (as some critics would assert) just *changing*—rapidly in the child area. Recent revisions (DSM-III-R, DSM IV) barely resemble the epoch-making edition of 1980 (DSM-III). At one time, low diagnostic reliability could have been blamed on inadequate assessment procedures, but now such inadequacy appears at least equally due to limitations of the underlying taxonomy itself. No diagnostic procedure can be expected to yield more valid diagnoses than the taxonomy will allow. There has been considerable taxonomic progress in the child area, but the validity of many diagnostic categories has been questioned, and it is not yet clear if the criteria and diagnostic thresholds proposed in the ever-changing DSM are correct. It is not clear that such changes are really taxonomic *improvements* as opposed to mere permutations and preferences of the DSM committees.

In a broader historical view, each version of the DSM must be seen as provisional, subject to revision and refinement. It is ironic, however, that about the time research results can address the validity of diagnostic categories, the diagnostic categories and criteria are revised. The changes have not been subtle. Some child diagnoses have disappeared completely, many new diagnoses have appeared, and many others have been radically reformulated. Researchers have been in a seemingly endless race of trying to "catch up" with such rapid revisions and interject empirical research results into the process of revision.

The design and use of structured interviews is not without assumptions of a theoretical nature. It is a major assumption, of course, that informants can provide valid information about children's emotional, behavioral, and social functioning. That parents can report on their own children's overt behavioral and social functioning is rarely questioned. It is less clear, however, that parents can provide reliable and valid information about covert behaviors that may be intentionally hidden from adults, such as truancy, alcohol and drug abuse, stealing, and vandalism; or about private phenomena such as fears, worries, and anxiety. Conversely, children and adolescents seem unimpeachable as sources of information about their own feelings and covert behaviors—even though a minimum level of cognitive maturity and degree of insight may be required. Whether children can see certain behaviors such as disobedience, inattentiveness, and stubbornness, as *symptoms* and report them during an interview remains controversial.

Lastly, the age and developmental level of the child being interviewed has created thorny problems—both theoretical and practical. Theoretically, issues involve how best to adapt interview procedures to abilities of the child—abilities that vary widely by age and developmental level. How to make interviews more developmentally attuned has been a major source of debate and empirical trial-and-error. The practical offshoot is obvious: to what age range can various interview schedules be used? This is often thought of as: "What is the lower age limit, or youngest-aged child, to which a given interview can be administered?" But the opposite is of concern as well: "Is there an upper age limit for which an interview is deemed appropriate?"

Many developmentalists have expressed caution about administering structured interviews to young children, on the grounds that they do not have the cognitive skills or language abilities to understand or respond correctly to complex and abstract questions—especially those about psychological phenomena. Indeed, questions designed to operationalize DSM diagnostic criteria are necessarily complex, if only because the diagnostic criteria themselves are complex.

Procedures for making interviews more amenable to young children have been advocated. Bierman (1984) was particularly articulate in stressing the importance of reducing task complexity, by using familiar vocabulary, simple sentences, and clear response options. Questions pertaining to time and frequency of past events have proven particularly vexing. Fallon and Schwab-Stone (1994) found that reliability was lower for the Diagnostic Interview Schedule for Children (DISC) questions requiring children to delineate time, compared to questions that either specified a time frame or did

not refer to time at all. Some interviews employ a visual time line as an aid to children's recall, and/or try to anchor recall to significant events (e.g., before the child's last birthday, after school started, during last summer). But the value of such procedures is not well established.

Empirical results can also address these issues. With a highly structured diagnostic interview, Edelbrock, Costello, Dulcan, Kalas, & Conover (1985) found that reliability of child reports increased rapidly over the age range from 6 to 18 years, and was low for children ages 6 to 9 years (average=.43), moderate for those ages 10 to 13 years (average = .60), and moderately high for those ages 14 to 18 years (average = .71). For many symptom areas, reliabilities for the younger group were unacceptably low, prompting the suggestion of ages 10 to 11 years as a practical lower limit for interviews of this type. Interview data from parents proved quite reliable across the age range from 6 to 18 years in this study. Fallon and Schwab-Stone (1994) also found that reliability of child reports increase with age, whereas parents are more highly reliable regardless of the child's age. Schwab-Stone and colleagues have also found that children were particularly unreliable in reporting about time factors such as symptom duration and onset (Schwab-Stone, Fallon, Briggs, & Crowther, 1994). These findings have supported the notion that only parents should be interviewed for children below age 10 or so; after ages 10 to 11 years-both parents and children should be interviewed. As reasonable as this sounds, it is a disappointing compromise, since a primary motivation for developing structured interviews was to provide a means of obtaining data from children themselves—especially younger children.

Results of a recent study are also quite disconcerting. This study involved interviewing children ages nine, 10, and 11, then debriefing them afterwards to determine their level of understanding of the interview items. The findings were dramatic and discouraging. The majority of children (more than 60 percent) did not understand the interview items. Questions involving time and frequency judgments were even more poorly understood (20-30 percent correctly understood). Unfortunately, their poor comprehension of the questions did not prevent these children from answering. Almost without exception, they responded to questions: It is the meaning or potential value of such responses that must now be seriously questioned.

DESCRIPTION

A structured interview is a list of target behaviors, symptoms, and events to be covered, guidelines for conducting the interview, and procedures for recording the data. Interview schedules differ widely in degree of structure. A crude, but useful, distinction can be made between *highly structured* and *semi-structured* interviews. Highly structured interviews specify the exact order and wording of questions and provide explicit rules for rating and recording the subject's responses. The interviewer is given very little leeway in conducting the interview and the role of clinical judgment in eliciting and recording responses is minimized. In fact, the interviewer is seen as an interchangeable part of the assessment machinery. Different interviewers should ask exactly the same questions, in exactly the same order, and rate and record responses in exactly the same way. Semistructured interviews, on the other hand, are less restrictive and permit the interviewer some flexibility in conducting the interview. The interviewer plays more of a role in determining what is asked, how questions are phrased, and how responses are recorded. Different interviewers should cover the same target phenomena when using a semistructured interview, but they may do so in different ways.

A high degree of structure does not necessarily yield consistently better data. Each type of interview has its advantages. Highly structured interviews minimize the role of clinical judgment and typically yield more objective and reliable data. But they are rigid and mechanical, which results in a stilted interview style that cannot be adapted to the individual respondent. Alternatively, semistructured interviews try to capitalize on expert clinical judgment and permit a more spontaneous interview style that can be adapted to the respondent. Of course, such flexibility allows more information variance to creep into the assessment process, which compromises reliability to some degree. The key unresolved issues are how highly structured clinical interviews should be and how much they should depend on clinical judgments by interviewers. These are complex issues, of course, and there may be no simple answer. The more appropriate questions may be: When would it be best to minimize clinical judgment by highly structuring the interview, and when would it be best to capitalize upon the expertise of clinically trained interviewers by providing less structure?

Structured clinical interviews for children and adolescents differ in other ways besides degree of structure. Most interview schedules have been developed for interviewing parents about their children, but parallel versions for directly questioning the child are becoming more common. Interview schedules also differ in length, organization, time requirements, age appropriateness, amount and type of interviewer training, and diagnostic coverage.

Semi-Structured Interviews

In the following section, we will briefly review several semi-structured interviews including, the Kiddie-SADS, the Child Assessment Schedule, and the Child and Adolescent Psychiatric Assessment.

The Kiddie-SADS

The Kiddie-SADS or K-SADS (Puig-Antich & Chambers, 1978) is a semi-structured diagnostic interview schedule for children ages 6 to 17 years, modeled after the Schedule for Affective Disorders and Schizophrenia (SADS), an interview schedule for adults developed by Endicott and Spitzer (1978). The K-SADS is designed to assess current psychopathology. It is focused on affective disorders but also covers conduct disorder, separation anxiety, phobias, attention deficits, and obsessions-compulsions. The K-SADS is administered by clinically sophisticated interviewers having intensive training using the interview schedule and expert knowledge about the DSM diagnostic criteria.

The parent is usually interviewed first about the child. Then the child is interviewed and any discrepancies between parent and child reports are addressed. The interviewer may confront the child about discrepancies and attempt to resolve them before making final ratings. The interviews begin with an unstructured section aimed at establishing rapport, obtaining a history of the present problems, and surveying current symptoms. Onset and duration of the current disorder and type of treatment received are then recorded. The interviewer then moves on to more structured sections covering specific symptoms. Each section includes an item (e.g., depressed mood) to be rated by the interviewer on a seven-point scale ranging from *not at all* to *very extreme*. Each section has a series of model questions (e.g., Have you felt sad? Have you cried?) that serve as guidelines for the interviewer. Interviewers are free, however, to ask as many questions as necessary to substantiate their symptoms ratings.

The K-SADS also embodies a *skip structure* whereby sections can be omitted if initial screening questions or "probes" are negative. If depressed mood is not evident, for example, subsequent questions in that section can be skipped. This reduces interviewing time substantially, but little information is lost.

Following the section on psychiatric symptoms, the interviewer rates 11 observational items (e.g., appearance, affect, attention, motor behavior) and rates the reliability and completeness of the entire interview. Finally, the interviewer completes a global-assessment scale reflecting overall degree of psychiatric impairment.

The K-SADS yields information on presence and severity of about 50 symptom areas (depending on the version of the interview). Most of the core areas concern depressive disorder, but somatization, anxiety, conduct disorder, and psychosis are also tapped. Additionally, there are 12 summary scales: four hierarchically related depression scales, five depression-related scales (e.g., suicidal ideation), and scales reflecting somatization, emotional disorder, and conduct disorder. The K-SADS data can also be translated into Research Diagnostic Criteria (RDC) and DSM diagnostic criteria for major depressive disorder, conduct disorder, and neurotic disorder. Diagnoses are based on the clinician's overview of the interview responses, rather than computer algorithms applied directly to the K-SADS data.

An epidemiological version of the K-SADS (K-SADS-E) is also available for assessing lifetime psychopathology (Orvaschel, Puig-Antich, Chambers, Tabrizi, & Johnson, 1982). It parallels the K-SADS, but most questions are phrased as "Have you *ever* done or had X?" As a preliminary test of validity, 17 subjects having previous depressive episodes were reinterviewed six months to two years later. For all but one subject, the K-SADS-E detected the same diagnosis that was made previously, suggesting accurate retrospective recall of previous psychiatric disturbances.

The K-SADS is widely used in clinical research and there is a growing body of findings supporting its reliability and validity. Short-term test-retest reliability has been evaluated on 52 disturbed chil-

dren and their parents (Chambers et al., 1985). Reliabilities averaged .55 (intraclass correlation, range: .09 to .89) for individual items, and averaged .68 (range: .41 to .81) for the 12 summary scales. Internal consistency for the 12 summary scales has averaged .66 (alpha statistic, range: .25 to .86). For diagnoses, agreement over time ranged from .24 to .70 (kappa statistic). Parent-child agreement has averaged .53 (intraclass correlation, range: .08 to .96) for individual items.

The K-SADS was developed primarily to identify children with major affective disorders. Since it is designed to assess diagnostic criteria, the validity of the K-SADS depends upon the validity of the diagnostic system (currently the DSM-III-R). In a sense, the K-SADS has strong content validity because it directly operationalizes DSM criteria. On the other hand, the DSM is evolving rapidly in the child area and the validity of many child psychiatric diagnoses is questionable—so the validity of the K-SADS is necessarily limited. Nevertheless, the K-SADS serves it intended purpose well. It has proven to be very useful in selecting homogeneous subgroups of depressed children from heterogeneous clinic populations (e.g., Puig-Antich, Blau, Marx, Greenhill, & Chambers, 1978). Preliminary investigations also suggest that the K-SADS is useful in research aimed at elucidating the biological correlates of childhood depression (Puig-Antich, Chambers, Halpern, Hanlon, & Sachar, 1979) and some core depression items are sensitive to treatment effects (Puig-Antich, Perel, Lupatkin, Chambers, Shea, Tabrizi, & Stiller, 1979).

The Child Assessment Schedule

The Child Assessment Schedule (CAS) is a semi-structured interview for children and adolescents ages 7 to 12 years (Hodges, McKnew, Cytryn, Stern, & Kline, 1982; Hodges, Kline, Stern, Cytryn, & McKnew, 1982). It was originally designed for directly interviewing the child only, but a parallel version for interviewing parents has been developed. The CAS is designed for clinically trained interviewers and requires about 45 to 60 minutes to administer to each informant (parent and child). It comprises 75 questions about school, friends, family, self-image, behavior, mood, and thought disorder. Most item responses are coded Yes/No. The interview is organized thematically beginning with questions about family and friends, followed by feelings and behaviors, and ending with items about delusions, hallucinations, and other psychotic symptoms. After interviewing the child, the interviewer rates 53 items (e.g., insight, grooming, motor behavior, activity level, speech).

The CAS was intended to facilitate evaluation of child functioning in various areas and to aid in the formulation of diagnostic impressions. It is less structured than other interview schedules, providing a simple outline of target phenomena to be assessed, suggested questions, and a simple format for recording the presence/absence of symptoms. The CAS yields scores in 11 content areas (e.g., school, friends, activities, family) and nine symptom areas (e.g., attention deficits, conduct disorder, overanxious, oppositional). A total score reflecting total number of symptoms is also obtained.

Clinical interpretation of the CAS is also flexible and requires considerable expertise. The interview was not originally designed to yield DSM diagnosis, although many items correspond to DSM criteria. A diagnostic index has been developed indicating the correspondence between CAS items and DSM criteria. To address DSM criteria more fully, a separate addendum to the interview has been developed for assessing symptom onset and duration. This complicates the interview somewhat, but provides more adequate coverage of DSM criteria for diagnosis of attention deficit disorder, conduct disorder, anxiety disorders, oppositional disorder, enuresis, encopresis, and affective disorders. Diagnosis are based on clinical overview of CAS responses, rather than explicit algorithms.

Interrater reliability based on independent ratings of 53 videotaped child interviews was r=.90 for total symptom score, and averaged r=.73 for content areas, and r=.69 for symptom areas. Reliabilities were somewhat higher for hyperactivity and aggression (average r=.80) than for fears, worries, and anxiety (average r=.60). Interrater reliabilities averaged kappa= .57 for individual items. Test-retest reliability has been high for quantitative scores in both diagnostic areas and content areas for an inpatient sample (Hodges, Cools, & McKnew, 1989). Diagnostic reliability has been moderately high for that inpatient sample for many diagnoses (range: kappa=.56–.1.00), but lower for ADHD (kappa=.43) and Overanxious Disorder (kappa–.38).

The validity of the CAS has been supported by several findings. Total symptom score discriminated significantly between inpatient, outpatient,

and normal children and correlated significantly (r=.53, p<.001) with total behavior problem score derived from the Child Behavior Checklist. Using referral for either inpatient or outpatient services as the criterion for psychopathology, the CAS achieved a sensitivity of 78 percent and a specificity of 84 percent, based on discriminant analysis (Hodges, Kline, et al., 1982). Combining CAS scores and CBCL scores in one discriminant analysis boosted sensitivity to 93 percent and specificity to 100 percent (no false positives). This suggests that combining parent and child data (or alternatively, interview and rating-scale data) may yield better discriminative power. Scores on the CAS overanxious scale have correlated significantly (r=.54, p<.001) with scores on the State-Trait Anxiety Scale for Children. CAS depression scores have also correlated significantly (r=.53, p<.001) with scores on the Child Depression Inventory.

Concordance between the CAS and the K-SADS has also been explored (Hodges, McKnew, Burbach, & Roebuck, 1987). Thirty clinically referred children ages 6 to 17 years and their parents were interviewed separately using either the CAS and the K-SADS, then reinterviewed the next day with the other interview schedule. Order of interviewing was counterbalanced so about half of the subjects were interviewed first with the CAS, whereas the other half were interviewed first with the K-SADS. Concordance between the two interview schedules was determined in four DSM-III diagnostic areas: ADD, conduct disorders, anxiety disorders, and affective disorders. Diagnoses were also made based on (a) the child only, (b) the parent only, (c) parent or child, and (d) parent and child consensus. Concordance between the CAS and K-SADS was moderately high for interviews with parents (average kappa =.62, range: .51 to .75), but lower for interviews with children (average kappa = .44, range: .36 to .52). Concordance was lower for anxiety disorders than other areas (kappa = .37 for the child interviews and .51 for the parent interviews). Taking all diagnoses from parent or child interviews reduced concordance slightly (average kappa = .54). Requiring parent-child consensus on diagnoses reduced concordance even more (average kappa = .46). Nevertheless, these results suggest moderately high concordance between the CAS and the K-SADS, particularly for parent interviews.

Overall, the CAS is a useful descriptive tool and diagnostic aid. It can be used with children as young as seven years of age and it has a very simple format. Development of a parallel form for parents, a diagnostic index, and an addendum covering symptom onset and duration are useful additions even though they complicate the interview and extend the interviewing time required. The CAS depends upon clinical inferences to a large extent, but is relatively easy for interviewers to learn.

The Child and Adolescent Psychiatric Assessment

A relatively new structured interview is the Child and Adolescent Psychiatric Assessment (CAPA) (Angold, Prendergast, Cox, Harrington, Simonoff, & Rutter, 1995). The CAPA is designed to assess both DSM and International Classification of Diseases (ICD) criteria for a range of core diagnoses including affective and anxiety disorders, and disruptive behavior disorders. The CAPA is moderately structured, with questions designed to be administered exactly as written, and a series of follow-up questions designed to determine if the respondent meets clinical criteria for a specific symptom. Test-retest reliability has been evaluated with 77 psychiatric patients ages 10 to 18 years (Angold & Costello, 1995). Surprisingly, higher reliabilities were obtained for affective and anxiety disorders (kappa=.74–.90) than for disruptive behavior disorders (kappa=.55–.64). Further research following up on these promising findings is warranted.

Highly Structured Interviews

The following section reviews two highly structured interviews: the Diagnostic Interview for Children and Adolescents and the Diagnostic Interview Schedule for Children.

The Diagnostic Interview for Children and Adolescents

The Diagnostic Interview for Children and Adolescents (DICA) was one of the first structured interviews for children and it has been widely used in clinical and epidemiological research. The original version, developed in 1969, was patterned after the Renard Diagnostic Interview and keyed to

the ICD and Feighner diagnostic criteria (see Welner, Reich, Herjanic, Jung, & Amado, 1987 for a review). The DICA was revised in 1981 along the lines of the NIMH Diagnostic Interview Schedule (Robins, Helzer, Croughan, & Ratcliff, 1981) and was keyed to then new DSM-III criteria. Research using the earlier version (e.g., Herjanic, Herjanic, Brown, & Wheatt, 1975; Herjanic & Campbell, 1977) was pioneering in many ways, but is probably obsolete, at least with respect to the reliability and validity of the later DSM-III and DSM-III-R versions.

The revised DICA is highly structured and provides the interviewer with specific wording of questions and explicit categories for response coding. Most symptom items are coded 1(No), 2(Yes), or 3(Uncertain). Responses coded "Uncertain" can be clarified by subsequent sub-questions and recoded either "Yes" or "No." The role of clinical inference in conducting the interview and making symptoms ratings has been minimized, so the DICA can be administered by clinicians or lay interviewers. A moderate amount of instrument-specific training is required, however.

Parallel interview schedules have been developed for interviewing the child (DICA-C) and parent (DICA-P) about the child. The parent version covers demographic-background information, pregnancy and childbirth, and medical and developmental history. A long section covers specific symptoms organized by diagnostic area (e.g., Attention Deficit Disorder, Conduct Disorder, Separation Anxiety Disorder). For each diagnosis, one or more questions have been written to cover each diagnostic criterion. The interview also includes questions about possible disorders in siblings and a brief family medical and psychiatric history. The child interview parallels the symptoms portion of the parent interview. Although the symptoms sections of the interviews are quite long, a skip structure is employed to reduce interviewing time if few symptoms are present.

The DICA yields information on the presence/absence of more than 150 specific symptoms, as well as their severity, onset, duration, and associated impairments (see Herjanic & Reich, 1982). Diagnoses are made by directly comparing item responses to DSM criteria for symptoms, severity, onset, and duration. All DSM diagnoses applied to children and adolescents are covered. Unlike the K-SADS where parent and child responses are first reconciled, DICA diagnoses are formulated separately from the parent interview and the child interview.

To test interrater reliability, 10 interviewers independently coded two videotaped interviews with children. Agreement on symptom items averaged 85 percent (Herjanic & Reich, 1982). Test-retest reliability has been determined by having five psychiatrists code the same videotaped interview twice over a two-to-three-month interval. Agreement over time averaged 89 percent (range: 80 percent – 95 percent) for individual symptom items. In another study, 27 children admitted to an inpatient psychiatric unit were interviewed twice by two different interviewers, one to seven days apart (Welner, Reich, Herjanic, Jung, & Amado, 1987). Inter-interviewer agreement on the presence/absence of specific diagnoses was quite high (kappas ranged from .76 to 1.00). Mother-father agreement was tested for a sample of 74 children (Sylvester, Hyde, & Reichler, 1987). Agreement regarding the presence of any diagnosis was moderately high (kappa = .54). Agreement was higher for Oppositional/Conduct Disorder and Attention Deficits (range: .54–.61) than for Anxiety Disorders and Depression (range: .33–.39). Parent-child agreement has also been determined using a sample of 84 children referred for outpatient services and their parents (Welner et al., 1987). For five diagnostic groupings (ADD, conduct disorders, affective disorders, enuresis, oppositional disorder), parent-child agreement on the presence/absence of the diagnosis averaged .62 (kappa statistic, range: .49 – .80). This represents much higher parent-child agreement than has been found in previous studies (e.g., Reich, Herjanic, Welner, & Gandhy, 1982).

Validity of the original DICA was supported by its ability to discriminate significantly between matched samples of pediatric and psychiatric referrals (Herjanic & Campbell, 1977). Validity of the DICA-C was tested for 27 inpatients by comparing DICA diagnoses with independent discharge diagnoses formulated by clinicians (Welner et al., 1987). Agreement was moderate for three diagnostic groupings: attention deficit disorders (kappa=.50), conduct disorders (.43), and affective disorders (.52), but was low for anxiety/phobic disorders (.03) and adjustment disorders (–.18).

Two other recent studies have addressed validity of the DICA. In one study, agreement between the DICA and best-estimate clinical diagnoses were determined for a sample of 30 children receiving inpatient services (Carlson, Kashani, Thomas,

Vaidya, & Daniel, 1987). For six diagnostic areas (ADD, conduct disorder, oppositional disorder, affective disorder, overanxious disorder, and separation anxiety), agreement with the best-estimate diagnoses was low-moderate for the DICA-C (average kappa=.38, range: .15 – .75) and the DICA-P (average kappa=.40, range: .05 – .66). In the other study, the DICA was compared with scores on the Personality Inventory for Children (PIC), a measure of child personality completed by parents (Sylvester, Hyde, & Reichler, 1987). Scores greater than T=65 on certain PIC scales (e.g., hyperactivity) were used to categorize children, then these categorizations were compared to their corresponding diagnosis (e.g., attention deficit disorder) derived from the DICA-C. This is a very stringent test of convergence, since it involves comparing a parent-completed personality inventory with the child-completed diagnostic interview. There were significant relationships between the two instruments in many areas, although the degree of convergence was fairly low (average kappa=.28, range: .11 to .48).

In sum, the DICA has broad diagnostic coverage and its moderate training requirements make it suitable for large-scale epidemiological surveys involving many nonprofessional interviewers. Recent studies have generally supported the reliability and validity of the DICA.

The Diagnostic Interview Schedule for Children

NIMH has sponsored the decade-long development of the Diagnostic Interview Schedule for Children (DISC) for use in epidemiological studies of child and adolescent psychopathology (see Costello, Edelbrock, Kalas, Kessler, & Klarie, 1982). The DISC is similar in design and purpose to the Diagnostic Interview Schedule (DIS) used in epidemiological research on adult disorders (Robins et al., 1981) and its offspring the DICA (see description above). The DISC is a highly structured diagnostic interview in which the order, wording, and coding of all items is specified. Like its predecessors, the DISC employs a skip structure to reduce interviewing time with children having few symptoms. Since it was designed for large-scale epidemiologic studies, the DISC can be administered by lay interviewers having two to three days of instrument-specific training. Parallel versions have been developed for separately interviewing children (DISC-C) and parents about their children (DISC-P). The child interview takes about 40 to 60 minutes to complete with clinically referred children, whereas the parent version takes about 60 to 70 minutes. A time frame of the last year is used for most items and specific information about onset and duration is sought for many symptom items.

The DISC covers a broad range of symptoms as well as their severity and chronicity. Most items are coded 0-1-2 where 0 corresponds to *no or never*, 1 corresponds to *somewhat, sometimes, or a little*, and 2 corresponds to *yes, often, or a lot.* Descriptions and examples offered by the respondent are recorded verbatim for later editing. The DISC was originally keyed to the DSM-III and covers most psychiatric diagnoses applicable to children and adolescents. Some diagnoses (e.g., pica, autism) are derived from the parent interview alone. Diagnoses are generated by computer algorithms applied to edited DISC data. Diagnoses are derived separately from the DISC-C and DISC-P. Both interviews also yield quantitative symptoms scores in symptoms areas (e.g., overanxious, conduct disorder, attention deficits).

Interrater reliability has been tested by having three lay interviewers independently code videotaped interviews of 10 children (Costello, Edelbrock, Dulcan, Kalas, & Klarie, 1984). Reliabilities averaged .98 for symptom scores (range: .94 to 1.00), indicating very little rater disagreement in how responses are coded. Test-retest reliability has been determined on 242 clinically referred children and their parents (Edelbrock et al., 1985). Parents and children were interviewed twice at a median interval of nine days. For parent interviews, test-retest reliability was .90 (intraclass correlation) for total symptom score and averaged .76 for symptom scales (range: .44 to .86). For child interviews, reliability was strongly related to age. For total symptom scores, reliabilities were .39, .55, and .81 for children ages 6-9, 10-13 and 14-18, respectively. For symptom scores, reliabilities also increased with age and averaged .43, .60, and .71 for children ages 6-9, 10-13, and 14-18 years, respectively. For 21 DSM-III diagnoses having sufficient prevalence, test-retest reliability for the parent interview averaged *kappa* = .56 (range: .35 to .81). Reliabilities of diagnoses derived from the child interviews averaged .36 (range: .12 to .71).

Parent-child agreement on child symptom scores has also been examined for 299 parent-child dyads (Edelbrock, Costello, Dulcan, Conover, & Kalas,

1986). Only a moderate degree of agreement was found overall (average $r = .27$), but agreement was higher for behavior/conduct symptoms than affective/neurotic symptoms and was higher among older than younger children. Regardless of the child's age, parents reported significantly more behavior/conduct problems than their children reported about themselves. Children reported significantly more affective/neurotic problems and drug and alcohol abuse than their parents reported about their children. Similar results were obtained using a Spanish version of the DISC with a community sample in Puerto Rico (Rubio-Stipec, Canino, Shrout, Dulcan, Freeman, & Bravo, 1994); and with the CAS interview with an inpatient sample (Hodges, Gordon, & Lennon, 1990). So despite generally low *levels* of agreement, *pattern* of disagreement seem consistent.

Validity of the DISC interviews has been supported by several lines of evidence. Costello, Edelbrock, and Costello (1985) compared matched samples of pediatric and psychiatric referrals and found that symptoms scores computed from both the DISC-P and DISC-C discriminated significantly between these criterion groups. Total symptom score derived from the DISC-P provided the best discrimination ($p<.001$). In a multiple discriminant analysis, symptoms scores derived from both parent and child interviews contributed significantly to the equation which correctly classified 77 of the 80 children. Based on the DISC-P, the psychiatric referrals obtained 51 diagnoses of severe disorders, compared to only two diagnoses in the pediatric group.

Symptoms scores derived from the DISC-C and DISC-P have also been shown to correlate significantly with other measures of child psychopathology, such as the parent and teacher versions of the Child Behavior Checklist (CBCL). Total symptom score derived from the DISC-P, for example, has correlated $r=.70$ with total behavior-problem score from the CBCL (Costello et al., 1984). The DISC-C has shown weaker, but significant relations to the CBCL ($r=.30$). Costello, Edelbrock, and Costello (1985, p. 591) also have found significant convergence between severe diagnoses from the DISC-P and CBCL scores above the normative range.

Edelbrock and Costello (1988) also have explored the relationship between DISC diagnoses and specific scales of the Child Behavior Profile. They found considerable convergence between diagnoses of attention deficit disorders, conduct disorder, and depression/dysthymia and scales labeled hyperactive, delinquent, and depressed, respectively. These relations were generally linear. An increasing score on the scale corresponded to an incrementally higher probability of obtaining the diagnoses. No "diagnostic threshold" was apparent. However, children scoring above the normative range ($T > 70$) on the scales were much more likely to receive the diagnosis than children scoring within the normative range. This suggests substantial convergence between two different ways of assessing child psychopathology.

In addition, relations between the DISC and the K-SADS has been determined in a community sample of children (Cohen, O'Conner, Lewis, Velez, & Malachowski, 1987). One hundred children ages 9 to 12 years who had been interviewed with the DISC were reinterviewed with the K-SADS three to four months later. Significant, although moderate, levels of agreement were obtained for many diagnoses. DISC diagnoses, however, have been shown to agree very poorly (kappa=.03–.17) with independent diagnoses by clinicians (Weinstein, Stone, Noam, Grimes, & Schwab-Stone, 1989). This is probably more an indictment of the reliability of the clinicians than the validity of the DISC, but sources of disagreement between the two sources are worth further investigation.

In the largest study to date, NIMH sponsored a multisite evaluation of the test-retest reliability of the DISC v2.1 on 97 clinically referred subjects and 278 non-referred community subjects (Jensen, Roper, Fisher, & Piacentini, 1995). This study produced many important findings. First of all, there was wide variability in reliability across the three participating sites. For ADHD, for example, test-retest reliability ranged from a high of kappa=.72 to a low of kappa=.38 for clinic cases. For conduct disorder, reliabilities ranged from a high of kappa=.90 to a low of kappa=–.11! Second, for most diagnoses, reliabilities were higher for parent interviews than child interviews, and this was true for both clinic and community subjects. Third, for all diagnostic categories, reliabilities were *lower* for the community sample, compared to the clinic sample (Jensen, Roper, Fisher, & Piacentini, 1995, p. 66). For example, for the category "Depression and/or Dysthymia," kappa=.70 for the clinic sample, but only kappa=.26 for the community sample. This is consistent with generally higher reliabilities found with clinic samples (e.g., Schwab-Stone et al.,

1993). This is very disconcerting, however, given that the DISC was designed for use in epidemiological studies of non-referred community samples.

Other Interviews

The range of interview schedules available for use has been expanded in several ways. The Teacher Interview for Psychiatric Symptoms (TIPS) (Kerr & Schaeffer, 1987), for example, is a semi-structured interview designed to obtain diagnostic information from teachers. Modeled after the K-SADS and ISC, the TIPS comprises 46 questions about psychiatric symptoms that might be evident in school (e.g., general anxiety, attention deficits). The TIPS takes about 45 minutes to complete and can be administered over the telephone. The interview begins with questions about the teacher's own teaching experience and style, then moves on to specific child symptoms which parallel the Interview Schedule for Children (ISC) (Kovacs, 1982) in format. The teacher is then asked 11 questions about the child's grooming, social popularity, school performance, and family problems.

Most research has focused on diagnostic interviews, but several interview schedules have been developed to assess non-diagnostic aspects of children's adaptation, social adjustment, and utilization of mental health services. The Social Adjustment Inventory for Children and Adolescents (SAICA) is a new interview schedule for assessing children's adaptive functioning in several domains (John, Gammon, Prusoff, & Werner, 1987). The SAICA is a semi-structured interview covering children's social and adaptive functioning in school, in their spare time, and with peers, siblings, and parents. It can be used to interview children and adolescents directly or to interview parents about their children. The SAICA is a very useful supplement to diagnostic assessment procedures.

The Child and Adolescent Functional Assessment Scale (CAFAS) was designed to determine the extent to which a psychiatric disorder is disruptive to the child's normal functioning (Hodges, 1994). The CAFAS covers five areas: role performance, thinking, behavior towards others, moods/self harm, and substance abuse. Interview questions also assess degree to which the child's caretaker can provide for the child's physical and emotional needs.

An interview called the Adolescent Adaptive Process Scales have also been developed to assess competence and adaptive behavior in adolescence (Beardslee, Jacobson, Hauser, Noam, & Powers, 1985). This clinical interview covers areas such as relationships with others, performance of age-appropriate tasks, thinking, and impulse control. Scores derived for the interview data appear to have good reliability and correlate with an independent measure of ego development (Beardslee, Jacobson, Hauser, Noam, Powers, Houlihan, & Rider, 1986).

Lastly, the Child and Adolescent Services Assessment (CASA) is an interview-based measure of amount and type of mental health services received by children and adolescents ages 8-18 years (Farmer, Angold, Burns, & Costello, 1994).

RESEARCH FINDINGS

Taken as a whole, the last 20 years of research on structured diagnostic interviews for children and adolescents indicate some very sobering conclusions. First of all, reliability of most interview schedules has been mediocre. Across all diagnoses, test-retest reliabilities have averaged about .40 – .50 (kappa statistic). Reliabilities have typically been somewhat higher (.50 – .60) for Disruptive Behavior Disorders, and lower (.20 – .40) for affective/anxiety disorders. It has been rare for any study to achieve test-retest reliability above kappa=.75 for any diagnosis—a commonly cited cutoff point for "excellent" reliability (Landis & Koch, 1977). Most studies have achieved "poor" to "fair" reliabilities for most diagnoses.

Second, most interviews have achieved about the same (mediocre) level of reliability. This is somewhat surprising, given the structural and procedural differences between interview schedules. It appears that wide variations in administration procedure, item structure and wording, level of interview training, and so on., have little net impact on diagnostic reliability. One exception to this general pattern of findings is the higher-than-typical reliability for Internalizing diagnoses obtained by the CAPA interview (Angold & Costello, 1995). These findings need to be replicated, but they might represent one benefit of trying to determine the *clinical significance* of reported symptoms during the interview process.

Third, there has been little improvement in reliability in the last 15 years, despite extensive efforts to refine the interview schedules themselves. This point is controversial, with some investigators seeing more improvement than others (see Costello, Burns, Angold, & Leaf, 1993; Hodges, 1994; Shaffer, 1994). But the weight of evidence suggests that increases in reliability, if there are any, are minimal to say the least.

Fourth, the "information yield"—literally what one gets out of the assessment process—has been disappointingly low for most interviews. Many investigators have found that interviews require lengthy interviewer training, they are very time-consuming, require extensive data management and exceedingly complex algorithmic scoring. This effort is typically not repaid with a rich database, but rather disappointingly crude diagnoses of general categories (e.g., "Any Anxiety Disorder," or "Disruptive Behavior Disorders," or alternatively, symptom scores for global syndromes such as "Internalizing" and "Externalizing." Only at the end of a study does an investigator realize that they could have obtained equivalent information using simple, quick, and inexpensive symptom checklists and behavioral rating scales. Structured interviews are often selected for use with the implicit hope and expectation of yielding more richness, detail, and contextual depth, than expedient paper-and-pencil measures. But most interview schedules yield only crude symptom counts and diagnoses—not rich descriptive data. To use Raymond Cattell's term, structured interviews have a very high "dross rate." Most of the information collected is not used.

Finally, there have been few innovations in structured interviewing in the modern era. Most child interviews represent downward extrapolations of adult interviews, which themselves have not changed much over the decades. Much of the effort to revise and refine child interviews has been at the level of exact item-wording or in structural details such as skip structure, and so on. It seems safe to say from our historical vantage point, that this has been misplaced precision.

Taking these general conclusions together, it appears that the limiting factor in interviewing lies not in the details of the interview schedules themselves but in the human respondent. Perhaps almost any interview schedule will quickly hit the maximum yield, reliability, and validity of information provided orally by another human being. Perhaps, as it appears, that level is disappointing low. If that is the case, then further tinkering with interview schedules is a waste of time, and more creative work on getting more and better information from human respondents will be needed.

SUMMARY

In the last fifteen years, more work has been done on the development and testing of structured interviews for children and adolescents than all previous years. Nevertheless, structured interviewing with children is relatively new and most interview schedules are still evolving. Research on the reliability and validity of structured interviews is still needed. Many interview schedules are reliable enough for making global distinctions among groups. Whether or not they are reliable enough for idiographic description and diagnosis remains to be seen.

Validation efforts have increased dramatically in the past few years. The most common approaches to testing validity have been: (a) comparing criterion groups such as clinically referred and nonreferred samples; (b) determining convergent relations with other indices of child psychopathology, particularly child-behavior rating scales; and (c) determining convergence between different interview schedules. Overall, most interview schedules have performed quite well, certainly well enough to warrant continued development and testing.

A range of applications has also been explored, including screening, description, and diagnosis. As screening tools, structured interviews are more costly and time-consuming than checklists and rating scales. Their screening performance is also usually no better and often much worse than much cheaper paper-and-pencil assessment techniques. As descriptive tools, structured interviews are roughly comparable to checklists and rating scales in terms of reliability and information yield. However, they lack the psychometric development and normative standardization of many rating scales and are probably not the best choice if the goal is description only. The advantage of interviews lies primarily in their diagnostic applications. Unlike most checklists and rating scales, many interview schedules are keyed to specific diagnostic criteria and cover not only symptom presence and severity but also the onset, duration, and associated impairments necessary for formal diagnoses.

Even so, no single interview schedule can be recommended for diagnostic assessments. Rather,

different types of interviews seem suited to different purposes. Both the K-SADS and ISC were developed to select subjects for research on childhood depression and they serve that purpose very well. Both are focused on affective symptoms and provide precise and detailed information about symptom severity and chronicity. These interviews are semi-structured and are intended for clinically sophisticated interviewers having extensive instrument-specific training. To the extent that they tap symptoms in other areas, the ISC and the K-SADS can also be recommended more generally for purposes of differential diagnosis among clinically referred samples.

The DISC and the DICA are at variance with the ISC and K-SADS. They are highly structured interviews in which the role of clinical judgment has been minimized. Both cover a broad range of symptoms and disorders and are suitable for large-scale studies employing lay interviewers. For these reasons, they seem more useful for describing symptom prevalence and distribution among non-referred populations, rather than for purposes of differential diagnosis of identified cases.

Future Directions

Research on the reliability and validity of structured interviews for children will undoubtedly continue for many years. It seems unlikely that many new interview schedules will be developed, but rather that research will concentrate on the handful of interviews already available. The more highly developed and tested interview schedules, such as the K-SADS, CAS, DICA, and DISC, will become standard assessment and diagnostic tools in clinical and epidemiologic research. Although such studies will not be directed at testing the interviews themselves, their results will certainly contribute to the evaluation and ultimate refinement of the interview schedules.

Researchers using structured interviews will also have to face many unsolved problems and issues. A key question is whether children are reliable and valid reporters of their own social, emotional, and behavioral functioning. The ability to directly question children about themselves is a major strength of interviewing and was one of the major stimuli to the development of structured interviews. However, many studies have obtained disappointingly low reliability from interviews with children. One study (Edelbrock et al., 1985) found that reliability of child reports was low for children below the age of 10, but increased to moderately high levels through middle childhood and adolescence. Parent-child agreement has also been low in most studies, although this depends upon many factors, such as the area being assessed, the age of the children, and the clinical status of the respondents. Low parent-child agreement is not necessarily an indictment of the interview schedules, since they may be accurately reflecting true differences in the way parents and children view child functioning. However, low agreement does raise the complex issue of how to deal with disparate data from different informants. Researchers have begun to try different strategies for integrating data from parent and child interviews, particularly when trying to formulate diagnoses (see Young, O'Brien, Gutterman, & Cohen, 1987).

A final issue involves taxonomic progress within child psychiatry. The diagnostic interviews are tied to the prevailing taxonomy of child disorders (i.e., the DSM). Validity of the interviews is simultaneously *built upon* and *limited by* the validity of the taxonomic system. Wholesale changes in diagnostic interviews were mandated by the advent of the DSM-III-R and DSM-IV (American Psychiatric Association, 1987, 1994), which embodies many substantive changes in the categories and criteria applied to children and youth. Some diagnostic interviews (e.g., DISC, DICA, K-SADS) were rekeyed to the DSM-III-R, thento the DSM-IV, but research on their performance is not yet available.

REFERENCES

Angold, A., & Costello, E. J. (1995). A test-retest reliability study of child-reported psychiatric symptoms and diagnoses using the Child and Adolescent Psychiatric Assessment (CAPA-C). *Psychological Medicine, 25*, 755–762.

Angold, A., Prendergast, M., Cox, A., Harrington, R., Simonoff, E., & Rutter, M. (1995). The Child and Adolescent Psychiatric Assessment (CAPA). *Psychological Medicine, 25*, 739–753.

American Psychiatric Association. (1956). *Diagnostic and statistical manual of mental disorders* (1st ed.). Washington, DC: Author.

American Psychiatric Association. (1968). *Diagnostic and statistical manual of mental disorders* (2nd ed.). Washington, DC: Author.

American Psychiatric Association. (1980). *Diagnostic and statistical manual of mental disorders* (3rd ed.). Washington, DC: Author.

American Psychiatric Association. (1987). *Diagnostic and statistical manual of mental disorders* (3rd ed. rev.). Washington, DC: Author.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington :D.C.:Author.

Beardslee, W., Jacobson, A., Hauser, S., Noam, G., & Powers, S. (1985). An approach to evaluating adolescent adaptive processes: Scale development and reliability. *Journal of the American Academy of Child Psychiatry, 24*, 637–642.

Beardslee, W., Jacobson, A., Hauser, S., Noam, G., & Powers, S., Houlihan, J., & Rider, E. (1986). An approach to evaluating adolescent adaptive processes: Validity of an interview-based measure. *Journal of Youth and Adolescence, 15*, 355–375.

Bierman, K. (1984). Cognitive development and clinical interviews with children. In B. Lahey and A. Kazdin (Eds.), *Advances in clinical child psychology* (Vol. 6, pp. 217–250). New York: Plenum.

Carlson, G. A., Kashani, J., Thomas, M., Vaidya, A., & Daniel, A. E. (1987). Comparison of two structured interviews on a psychiatrically hospitalized population of children. *Journal of the American Academy of Child Psychiatry, 26*, 645–648.

Chambers, W., Puig-Antich, J., Hirsch, M., Paez, P., Ambrosini, P., Tabrizi, M. A., & Davies, M. (1985). The assessment of affective disorders in children and adolescents by semi-structured interview: Test-retest reliability of the K-SADS. *Archives of General Psychiatry, 42*, 696–702.

Cohen, P., O'Connor, P., Lewis, S., Velez, N., Malachowski, B. (1987). Comparison of DISC and K-SADS interviews of an epidemiologic sample of children. *Journal of the American Academy of Child Psychiatry, 26*, 662–667.

Costello, E. J., Burns, B., Angold, A., & Leaf, P. (1993). How can epidemiology improve mental health services for children and Adolescents? *Journal of the American Academy of Child Psychiatry, 32*, 1106–1113.

Costello, E. J., Edelbrock, C., & Costello, A. J., (1985). The validity of the NIMH Diagnostic Interview Schedule for Children. *Journal of Abnormal Child Psychology, 13*, 579–595.

Costello, A. J., Edelbrock, C., Dulcan, M. K., Kalas, R., & Klaric, S. H. (1984). *Development and testing of the NIMH Diagnostic Interview Schedule for Children in a clinic population.* Final report (Contract # RFP-DB-81-0027). Rockville, MD: Center for Epidemiologic Studies. National Institute of Mental Health.

Costello, A. J., Edelbrock, C., Kalas, R., Kessler, M. D., & Klaric, S. H. (1982). *The NIMH Diagnostic Interview Schedule for Children (DISC).* Unpublished interview schedule. Pittsburgh: Department of Psychiatry, University of Pittsburgh.

Edelbrock, C., & Costello, A. J. (1988). Convergence between statistically derived behavior problem syndromes and child psychiatric diagnoses. *Journal of Abnormal Child Psychology, 16*, 219–231.

Edelbrock, C., Costello, A. J., Dulcan, M. K., Conover, N. C., & Kalas, R. (1986). Parent-child agreement on child psychiatric symptoms assessed via structured interview. *Journal of Child Psychology and Psychiatry, 27*, 181–190.

Edelbrock, C., Costello, A. J., Dulcan, M. K., Kalas, R., & Conover, N. C. (1985). Age differences in the reliability of the psychiatric interview of the child. *Child Development, 56*, 265–275.

Endicott, J., & Spitzer, R. (1978). A diagnostic interview: The Schedule for Affective Disorders and Schizophrenia. *Archives of General Psychiatry, 35*, 837–844.

Fallon, T., & Schwab-Stone, M. (1994). Determinants of reliability in psychiatric surveys of children aged 6–12. *Journal of Child Psychology and Psychiatry, 35*, 1391–1408.

Farmer, E., Angold, A., Burns, B., & Costello, E. J. (1994). Reliability of self-reported service use: Test-retest consistency of children's responses to the child and adolescent services assessment (CASA). *Journal of Child and Family Studies, 3*, 307–325.

Freeman, M. (1971). A reliability study of psychiatric diagnosis in childhood and adolescence. *Journal of Child Psychology and Psychiatry, 12*, 43–54.

Group for the Advancement of Psychiatry. (1957). *The diagnostic process in child psychiatry.* New York: Author.

Group for the Advancement of Psychiatry. (1966). *Psychopathological disorders in childhood: Theoretical considerations and a proposed classification.* New York: Author.

Herjanic, B., & Campbell, W. (1977). Differentiating psychiatrically disturbed children on the basis of a structured interview. *Journal of Abnormal Child Psychology, 5*, 127–134.

Herjanic, B., Herjanic, M., Brown, F., & Wheatt, T. (1975). Are children reliable reporters? *Journal of Abnormal Child Psychology, 3*, 41–48.

Herjanic, B., & Reich, W. (1982). Development of a structured interview for children: Agreement

between child and parent on individual symptoms. *Journal of Abnormal Child Psychology, 10*, 307–324.

Hodges, K. (1994). *The Child and Adolescent Functional Assessment Scale: Instructions for scoring.* Test manual available from the author, Eastern Michigan University, Psychology Dept., Ypsilanti, MI 48197.

Hodges, K. (1994). Debate and argument: Reply to David Shaffer: Structured interviews for assessing children. *Journal of Child Psychology and Psychiatry, 35*, 785–787.

Hodges, K., Cools, J., & McKnew, D. (1989). Test-retest reliability of a clinical research interview for children: The Child Assessment Schedule. *Psychological Assessment, 1*, 317–322.

Hodges, K., Gordon, Y., & Lennon, M. P. (1990). Parent-child agreement on symptoms assessed via a clinical research interview for children: The Child Assessment Schedule (CAS). *Journal of Child Psychology and Psychiatry, 31*, 427–436.

Hodges, K., Kline, J., Stern, L., Cytryn, L., & McKnew, D. (1982). The development of a child assessment interview for research and clinical use. *Journal of Abnormal Child Psychology, 10*, 173–189.

Hodges, K., McKnew, D., Burbach, D. J., & Roebuck, L. (1987). Diagnostic concordance between the Child Assessment Schedule (CAS) and the Schedule for Affective Disorders and Schizophrenia for School-Age Children (K-SADS) in an outpatient sample using lay interviewers. Journal of the *American Academy of Child Psychiatry, 26*, 654–661.

Hodges, K., McKnew, D., Cytryn, L., Stern, L., & Kline, J. (1982). The Child Assessment Schedule (CAS) diagnostic interview: A report on reliability and validity. *Journal of the American Academy of Child Psychiatry, 21*, 468–473.

Jensen, P., Roper, M., Fisher, P., & Piacentini, J. (1995). Test-retest reliability of the Diagnostic Interview Schedule for Children (DISC 2.1). *Archives of General Psychiatry, 52*, 61–71.

John, K., Gammon, G.D., Prusoff, B., & Werner, V. (1987). The Social Adjustment Inventory for Children and Adolescents (SAICA): Testing of a new semistructured interview. *Journal of the American Academy of Child and Adolescent Psychiatry, 26*, 916–921.

Kerr, M.M., & Schaeffer, A.L. (1987). *Teacher Interview for Psychiatric Symptoms (TIPS).* Pittsburgh, PA: Author.

Kestenbaum, C.J., & Bird, H. (1978). A reliability study of the Mental Health Assessment Form for school age children. *Journal of the American Academy of Child Psychiatry, 7*, 338–347.

Kovacs, M. (1982). *The Interview Schedule for Children (ISC).* Unpublished interview schedule. Department of Psychiatry, University of Pittsburgh.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174.

Lapouse, R., & Monk, M. A. (1958). An epidemiologic study of behavior characteristics of children. *American Journal of Public Health, 48*, 1134–1144.

Matarazzo, J. D. (1983). The reliability of psychiatric and psychologic diagnosis. *Clinical Psychology Review, 3*, 103–145.

Orvaschel, H., Puig-Antich, J., Chambers, W., Tabrizi, M. A., & Johnson, R. (1982). Retrospective assessment of prepubertal major depression with the Kiddie-SADS-E. *Journal of the American Academy of Child Psychiatry, 21*, 392–397.

Puig-Antich, J., Blau, S., Marx, N., Greenhill, L., & Chambers, W. (1978). Pre-pubertal major depressive disorder: A pilot study. *Journal of the American Academy of Child Psychiatry, 17*, 695–707.

Puig-Antich, J., & Chambers, W. (1978). *The Schedule for Affective Disorders and Schizophrenia for school-aged children.* Unpublished interview schedule. New York State Psychiatric Institute, New York, NY.

Puig-Antich, J., Chambers, W., Halpern, F., Hanlon, C., & Sacher, E. (1979). Cortisol hypersecretion in prepubertal depressive illness: A preliminary study. *Psychoneuroendocrinology, 4*, 191–197.

Puig-Antich, J., Perel, J. M., Lupatkin, W., Chambers, W., Shea, C., Tabrizi, M. A., & Stiller, R. L. (1979). Plasma levels of imipramine (IMI) and desmethylimipramine (DSI) and clinical response in prepubertal major depressive disorder: A preliminary report. *Journal of the American Academy of Child Psychiatry, 18*, 616–627.

Reich, W., Herjanic, B., Welner, Z., & Gandhy, P. R. (1982). Development of a structured psychiatric interview for children: Agreement on diagnosis comparing parent and child. *Journal of Abnormal Child Psychology, 10*, 325–336.

Robins, L., Helzer, J. E., Croughan, J., & Ratcliff, K. S. (1981). National Institute of Mental Health Diagnostic Interview Schedule: Its history, characteristics, and validity. *Archives of General Psychiatry, 38*, 381–389.

Rosen, B. M., Bahn, A. K., & Kramer, M. (1964). Demographic and diagnostic characteristics of psychiatric clinic patients in the USA. *American Journal of Orthopsychiatry, 34*, 455–468.

Rubio-Stipec, M., Canino, G., Shrout, P., Dulcan, M., Freeman, D., & Bravo, M. (1994). Psychometric properties of parents and children as informants in child psychiatry epidemiology with the Spanish Diagnostic Interview Schedule for Children (DISC.2). *Journal of Abnormal Child Psychology, 22*, 703–720.

Rutter, M., & Graham, P. (1968). The reliability and validity of the psychiatric assessment of the child: Interview with the child. *British Journal of Psychiatry, 11*, 563–579.

Sandifer, M. G., Pettus, C. M., & Quade, D. (1964). A study of psychiatric diagnosis. *Journal of Nervous and Mental Disease, 139*, 350–356.

Schwab-Stone, M., Fallon, T., Briggs, M., & Crowther, B. (1994). Reliability of diagnostic reporting for children aged 6-11 years: A test-retest study of the Diagnostic Interview Schedule for Children-Revised. *American Journal of Psychiatry, 151*, 1048–1054.

Schwab-Stone, M., Fisher, P., Piacentini, J., Shaffer, D., Davies, M., & Briggs, M. (1993). *Journal of the American Academy of Child and Adolescent Psychiatry, 32*, 651–657.

Shaffer, D. (1994). Debate and argument: Structured interviews for assessing children. *Journal of Child Psychology and Psychiatry, 35*, 783–784.

Sylvester, C. E., Hyde, T. S., & Reichler, R. J. (1987). The Diagnostic Interview for Children and Personality Inventory for Children in studies of children at risk for anxiety disorders or depression. *Journal of the American Academy of Child and Adolescent Psychiatry, 26*, 668–675.

Tarter, R., Templer, D., & Hardy, C. (1975). Reliability of the psychiatric diagnosis. *Diseases of the Nervous System, 36*, 30–31.

Weinstein, S., Stone, K., Noam, G., Grimes, K., & Schwab-Stone, M. (1989). Comparison of DISC with clinicians' DSM-III diagnoses in psychiatric inpatients. *Journal of the American Academy of Child and Adolescent Psychiatry, 28*, 53–60.

Welner, Z., Reich, W., Herjanic, B., Jung, K., & Amado, H. (1987). Reliability, validity, and parent-child agreement studies of the Diagnostic Interview for Children and Adolescents (DICA). *Journal of the American Academy of Child Psychiatry, 26*, 649–653.

Young, G., O'Brien, J., Gutterman, E., & Cohen, P. (1987). Research on the clinical interview. *Journal of the American Academy of Child and Adolescent Psychiatry, 26*, 613–620.

CHAPTER 15

STRUCTURED CLINICAL INTERVIEWS FOR ADULTS

Arthur N. Wiens
Patricia J. Brazil

INTRODUCTION

Since the earlier publication of this chapter (Wiens, 1990) there has been a remarkable evolution in health care including mental-health care. From an earlier era in which many clinical psychologists practiced in private with their patients, we have moved to an era of managed care, defined broadly as any patient care that is not determined solely by the provider (Goodman, Brown, & Dietz, 1996). That definition covers virtually all clinical practice today and may range from management of the immediate expense of care to the total management of the patient's care, from wellness through chronic illness. Goodman, and colleagues (1996) point out that organizations managing mental-health benefits may utilize psychiatrists, clinical psychologists, clinical social workers, and marriage, family, and child counselors to provide the same-coded treatment service. Each of these service providers is subject to preauthorization and each is asked to justify the necessity for, and demonstrate the effectiveness of, their services. Standardized assessment and treatment documentation is being required more and more often.

Now, in the context of careful time management, the clinical psychologist must seek to establish patient rapport quickly and attend to the unique needs and concerns of patients while amassing a large amount of critical information about a patient's assessment and treatment needs, any immediate decisions that have to be made, issues of suicide potential or other untoward behavior, and a differential diagnosis based on the DSM-IV (*Diagnostic and Statistical Manual of Mental Disorders-Fourth Edition*, American Psychiatric Association, 1994).

The DSM-IV-PC (*Diagnostic and Statistical Manual of Mental Disorders-Fourth Edition-Primary Care Version*, American Psychiatric Association, 1995) is an offshoot of the DSM-IV and was developed as a collaborative effort of representatives from primary care specialties in medicine and psychiatry. Of particular interest in this document is the presentation of a general algorithm for use with the DSM-IV-PC, that is, each algorithm presents a series of steps that indicate in sequence which disorders to consider. What these definitions have in common with the clinical interview process is that a structured clinical interview proceeds through a series of steps, each step meant to indicate a particular segment of the diagnostic picture with the ultimate conclusion, or problem to be solved, being the diagnosis. The clinician's first task is to determine which of the DSM-IV-PC algorithms should be reviewed. Once the correct algorithm has been selected, step-by-step instructions are provided for considering those disorders that may account for the presenting symptom. Algorithms provide a structured, systematic framework that allows the clinician to move from the

beginning of the interview to the end of an interview, without becoming lost as the patient decides to take a side road or detours from the question at hand. It enables the interviewer to feel confident when approaching a specific patient problem knowing that there is a guide through the interview process.

Many of the DSM-IV-PC diagnostic algorithms share certain steps in common, including:

- considering the role of a general medical condition, that is, determining whether the presenting symptom is due to the direct physiological effects of a general medical condition;
- considering the role of substance abuse, that is, evaluating whether the presenting symptom is a direct physiological effect of a drug of abuse, a medication side effect, or toxin exposure;
- considering whether symptoms are better accounted for by another mental disorder since some symptoms are associated with a number of conditions;
- considering whether the presenting symptomatology is attributable to a mental disorder at all (APA, 1995).

For example, in the Depressed Mood Algorithm there are considerations for epidemiology, primary care presentation, and differential diagnoses. The algorithm is organized so that diagnoses with immediate-treatment implications are addressed first, chronic-depressive conditions are addressed next, and the algorithm ends with three conditions seen frequently by primary-care physicians. Again, the point is that there is a step-wise progression in gathering data.

A third major development has been the information superhighway and automated clinical information systems. Today people communicate with each other in many ways. For example, with the technological advances of the past decade individuals can now "chat" with each other on their computers. Patients and clinicians can send each other questions and replies via e-mail. A patient may feel that with e-mail contact is possible with his or her clinician at any time day or night, and in some instances obviate the necessity for an office visit. With simultaneous video/computer capability, it is possible for a clinician-specialist in another city to have essentially face-to-face observation and communication with a patient. To us, there is little doubt that a transformation in health-care delivery is under way, that computers are the instruments of change, and that communication between patients and medical databases and between patients and clinicians promises to replace a substantial amount of care now delivered in person. We discuss this issue later in this chapter under the heading of telehealth and health-care informatics. Clinician-to-clinician communication can also be enhanced; as is the case when rural practitioners can go on-line with a national network of specialists. Present-day clinicians are part of a transition generation of practitioners who rely on the historical and traditional uses of the interview but also are using newer technologies.

INTERVIEWING

Interviews take place hundreds of thousands of times a day throughout the world in offices, government agencies, cafeterias, law enforcement, and so on, and are used by journalists, teachers, politicians, clinicians, and others. The clinical interview is but one of many types of interviews; to think about it in this broader context may make it easier to adapt some of the newer innovations in interviewing to clinical information gathering.

An interview has several traditional characteristics. One characteristic is that it has a serious general or specific purpose, for example, wanting to provide or obtain information for problem solving or more specific decision-making. The interview is usually planned before meeting with the interviewee so that the serious purpose of the interview can be achieved. Clinicians review what is already known about a patient and may structure the planned interview quite specifically to acquire further information.

Clinicians have also always thought that there is both objective and subjective information to be obtained in an interview, for example, feelings and beliefs in addition to objective responses to questions and nonverbal communications. For such reasons it has been assumed that interviews are face-to-face verbal exchanges in which one person, the interviewer, attempts to elicit information or expressions of opinion or belief from another person, the interviewee. As we have already suggested, this traditional definition of the interview is being modified.

During an interview the participants assume differential roles of "interviewer" and "interviewee" and expectations are established for each role. Yet, both participants maintain some degree of control;

for example, the interviewer is expected to control the direction of the interview but the interviewee possesses informational control. In some interviews the participants may alternate roles as in an unstructured interview when the patient may control both the direction and information of the interview. Finally, the interviewer has, or should have, some criteria as to whether the interview was successful. We take a look next, in a bit more detail, as to how an interview might be distinguished from a conversation.

INTERVIEW VERSUS CONVERSATION

Both an interview and a conversation typically involve a face-to-face verbal exchange of information, ideas, attitudes, or feelings and contain messages exchanged through non-verbal as well as verbal modes of expression (Wiens, 1983). However, a crucial characteristic that distinguishes an interview from a conversation is that the interview is designed to achieve a consciously selected purpose. There may be no central theme in a conversation, but in an interview the content is directed toward a specific purpose and is likely to have unity, progression, and thematic continuity. If the purpose of the interview is to be achieved, one participant must assume and maintain responsibility for directing the interaction (asking questions) toward the goal, and the other participant must facilitate achievement of the purpose by following the direction of the interaction (answering questions).

The nonreciprocal roles of the two participants in an interview result from the fact that, in one form or another, the purpose of the interview is to give some benefit to the interviewee. Furthermore, whereas in conversation a person may behave in a spontaneous and unplanned manner, an interviewer deliberately and consciously plans actions to further the purpose of the interview. A conversation may be started and terminated at will; however, an interview, once initiated, ordinarily is continued until its purpose has been achieved or until it is clear that the purpose cannot be achieved. Stated in another way, the immediate purpose of most interviews is to encourage the interviewee to engage in some kind of self-exploration to satisfy a purpose explicitly or implicitly agreed upon by the interviewer and interviewee. Most interviews involve interviewer and interviewee agreeing on the objective to be reached, deciding what topics need to be discussed, establishing a relationship of trust that allows the interviewee to talk freely, and keeping the discussion focused on relevant information. The interviewee usually does most of the talking: usually using about 80 percent of the talk-time. If the percentage of talk-time varies greatly from this it will probably mean that the interviewer is enjoying talking about himself or herself but is not learning much about the interviewee. A relatively unstructured interviewing style allows interviewees to discuss experiences largely on their own terms. In the case of selection interviews this could make it difficult to compare candidates, and in the case of clinical interviews this style could make it difficult to ascertain whether diagnostic criteria have been met. An additional consideration in either a selection or clinical interview is the importance of details. A detailed inquiry or a persistent follow-up on an initially general question can bring to the foreground critical information about the interviewee.

SELECTED INTERVIEWING APPLICATIONS

Information Interview

Interviewing can be observed daily on national and local television, and one might profit from paying close attention to the skills of television interviewers as they try to elicit verbal responses from a variety of interviewees, some of whom are presumably motivated to be frank, open, and expressive, while others are quite reticent. One such television interviewer has published her thoughts and experiences on how to talk with practically anybody about practically anything (Walters, 1970). Ted Koppel has been quoted as saying: "I listen. Most people don't. Something interesting comes along—and whoosh!—it goes right past them" (Koppel, 1987). It might be assumed that the information being sought exists only in the interviewee's mind, which may also be true for other types of interview applications. It is interesting to contemplate how much in common this interview may have with many other types.

Research Interview

The research interview may be among the most carefully planned interview situations in that it is designed to gather the same information from a large number of individuals. For example, standardized, or structured, interviews have been used for many purposes such as determining public opinion on a wide range of issues, for example, foreign policy, the popularity of candidates in elections, consumer preferences, human sexual behavior, leisure activities, epidemiology of psychiatric symptoms. A standardized interview may be used once or it may be used on a series of occasions to determine the stability of certain attitudes, values, and practices.

Standardized research interviews undergo pilot testing and development and this pretesting provides a final set of questions. Often the interviewer reads questions verbatim and records responses into a set of precoded choices. There is a common vocabulary so that it is possible to formulate questions which have the same meaning for each interviewee. If the meaning of each question is to be identical for each respondent, its context must be identical and, since all preceding questions constitute part of the context, the sequence of the questions must be identical. The experience with research interviewing also has contributed much to the development of structured clinical interviewing.

Employment Interview

Millions of employment interviews take place annually in the United States with the explicit purpose of determining an individual's suitability for a particular job. Eder & Ferris (1989) have suggested that research on the employment interview has advanced knowledge in understanding interviewer cognitive processes. They support the research efforts to examine interviewer judgment from alternative theoretical perspectives by focusing on preinterview impression effects, structured question formats, interviewer-applicant process dynamics, applicant-impression management strategies, situation variables and their combined effects on the interviewer's information processing system.

Eder & Ferris (1989) also believed that further application of computer technology within a decision support-system context could be used to enhance the interviewer's information-processing capability. They suggested that software packages might call up the latest job description, draft job-relevant rating criteria, recommend questions, and glean relevant information from a candidate's completed application form five minutes prior to the scheduled interview. Post-interview ratings could then be entered immediately to enhance recall, analyze reliability across multiple interviewers, highlight rating inconsistencies for subsequent discussion, and calculate likely incumbent productivity. Parallel software to aid the applicant's job-search efforts could also be developed. They suggested that researchers focus on both interview process and outcome measures.

Treatment Interview

The suggestion for a focus on both process and outcome has long been an interest for psychotherapists as well. The therapeutic interview has often been considered the most sensitive of interview applications because it is commonly used to deal with sensitive subjective information such as feelings, attitudes, and personal problems. Alcoholism, drug abuse, mental health, marital strife, and sexual inadequacies are all common topics for therapeutic interviews. The goal of such interviews is to get interviewees to gain insight into their own problems and attempt to solve them themselves. Thus, the focus is often on (the means to that end) interviewee insight especially as this might be contrasted with the goal of the diagnosis of the problems.

Interviewer attitudes that facilitate such interviewee problem-solving efforts are many; we suggest three. The interviewer quality of *acceptance* involves a basic regard for the worth of human individuals and particularly for the interviewee sitting in the office. The accepting interviewer does not view the interviewee with cynicism or contempt. In fact, we suggest that if the interviewer has not discovered something to like about the interviewee by the end of an initial interview, the session has not been successful in establishing a relationship. We remind the reader of the old saying that "people don't care what you know until they know that you care." The ability to *understand* emotionally (or empathize) requires the effort to grasp clearly and completely the meaning the interviewee is trying to convey. The attempt to understand is a sharing process in which the inter-

viewer tries to assume the client's place and tries to see the circumstances as they appear to the interviewee. The quality of *sincerity*, which has been called congruence, refers to interviewer consistency, or the harmony that must exist between what an interviewer says and does and what he or she really is or feels. You can "accept" or "understand" somebody, but you cannot "sincere" somebody. The interviewer can only be "sincere," and although this quality is hard to define, it might be considered an interviewer characteristic to the extent to which the interviewer communicates a valid and reliable picture of what he or she is like inside.

The interviewer characteristics just described are desirable no matter what type of interview application one is engaged in. They are certainly desirable for the clinical interview as well as the therapeutic interview.

THE CLINICAL INTERVIEW

Many clinicians have been heavily influenced by earlier psychoanalytic thought that placed considerable emphasis on the *indirect* techniques of interviewing, and a free-flowing exchange between the clinician and patient. Generally, such unstructured interviews allow the clinician freedom to reword questions, to introduce new questions, or to modify question order, and to follow patients' spontaneous sequence of ideas. It is often assumed that such spontaneous discussion allows patients to follow more nearly their natural train of thought and may allow them to bring out interview material that is more predictive of what they would say or do in real-life situations. The flexibility of the unstructured interview may allow clinicians to adapt their techniques to patients' particular situations. In some cases the interviewer may omit topics that do not seem applicable, and in other cases he or she may introduce related topics not originally planned. Many readers may have watched skilled clinician-interviewers elicit previously hidden facts, using attention to conflicts, dysphoric affects, defenses used by the patient, and symptom origins. The sophisticated data-reduction techniques and hypothesis testing carried out consciously or preconsciously in interviews by skilled clinicians have such practical value that development of these skills became the primary pursuit of many clinicians (Young, O'Brien, Gutterman, & Cohen, 1987).

Experienced clinicians often assume that they can maintain best rapport with patients by formulating questions in words that are familiar to patients and habitually used by them, and by pursuing topics when patients indicate a readiness and willingness to discuss them. It is usually assumed that the unstructured clinical interview gives more discretion to the clinician in formulating the wording and sequence of questions in this way, and accordingly it requires a higher level of experience, skill, and training than is required in following a more standardized interview format. Required in particular is an overall conceptual grasp of theoretical context and considerable prior knowledge of the subject matter of the interview.

Semi-Standardized Interviewing

While clinicians may have espoused a spontaneous interview style, it seems that, actually, most experienced clinicians have adopted a semi-standardized interviewing style or format. If one listens to a clinician interviewing a series of patients, one soon discerns topic areas that are routinely introduced, and questions that are asked in almost the same way of every patient.

Furthermore, the topics to be covered in an initial clinical interview are relatively consistent from one clinician to the next. The general objective is to obtain a careful history that can be the foundation for diagnosis and treatment of the patient's illness (Kaplan & Sadock, 1988). More specific objectives are to understand the individual patient's personality characteristics, including both strengths and weaknesses; to obtain insight into the nature of their relationships with those closest to them, both past and present; and to obtain a reasonably comprehensive picture of the patient's development from the formative years until the present.

In preparing a written record of a clinical interview most clinicians begin by presenting *identifying information* such as the patient's name, age, marital status, sex, occupation, race, place of residence and circumstances of living, history of prior clinical contacts, and referral and information sources. The *chief complaint*, or the problem for which the patient seeks professional help, is usually reviewed next and is stated in the patient's own words or in the words of the person supplying this information. The intensity and duration of the presenting problem is noted, specifically the length of time each symptom has

existed and whether there have been changes in quality and quantity from a previous state. It is also useful to include a description of the patient's appearance and behavior. In reviewing *present illness* the clinician looks for the earliest and most disabling symptoms and whether there were any precipitating factors leading to the chief complaint. Often the precipitating or stress factors associated with the onset of symptoms may be subtle and require the clinician to draw on knowledge of behavior and psychopathology to help with inquiry regarding relevant life- change events. The clinician should also report on how the patient's symptoms have affected his or her life activities. It is important to review *past health history* for both physical and psychological problems; for example, are there physical illnesses that might be impacting the patient's emotional state? Prior episodes of emotional and mental disturbances should be described. The clinician also needs to inquire about and report prescribed medication and alcohol and drug use. Possible organic mental syndromes must be noted. *Personal history* may include information about the patient's parents and other family members and any history of psychological or physical problems. The account of the patient's own childhood and developmental experiences may be quite detailed. Educational and occupational history is noted as well as social, marital, military, legal, and other experiences. The personal history should provide a comprehensive portrait of the patient independent of his or her illness (Siassi, 1984). The *mental status examination* is reviewed under the following headings: general appearance and behavior; mood, feelings, and affect; perception; speech and thought; sensorium and cognition; judgment; insight; and reliability. Finally, *recommendations* are presented about what kind of treatment the patient should receive for what problems and target symptoms.

Topic areas to be covered are also relatively consistent among clinicians with different theoretical approaches. The interested reader may note commonalities between the description of the clinical interview and the assessment schema that many behavioral interviewers refer back to (Kanfer & Saslow, 1969). These authors suggest examination of the following areas: analysis of the problem situation (including behavioral excesses, deficits and assets); clarification of the problem situation that maintains the targeted behaviors; a motivational analysis; a developmental analysis (including biological, sociological, and behavioral spheres); a self-control analysis; an analysis of social relationships; and an analysis of the socio-cultural-physical environment.

SELECTED PURPOSES OF CLINICAL INTERVIEWS

Diagnosis

The act of classification is basic to all science and to every other aspect of living. Accurate and reliable description that differentiates and predicts is the basis of hypothesis formation and testing in science (Wiens & Matarazzo, 1983). Diagnosis in clinical practice introduces order into the clinician's observations, with an attendant increase in meaningfulness and, ultimately, control (prevention and amelioration). Placing an object or organism or a set of behaviors into a certain class allows us to infer certain characteristics without needing to demonstrate each characteristic *de novo*. Classification can also help to put individual observations into a different perspective or context, and stimulate new questions for better treatment, prevention, control, and future research.

The purposes of clinical interviews and mental status examinations are to arrive at a diagnostic formulation and a rational treatment plan. Several decades ago, it probably made little difference what specific diagnosis was assigned, since the available treatments were highly limited and, by necessity, more or less applied to all patients (Siassi, 1984). However, as diagnostic criteria have become more detailed, and some treatment procedures applied more selectively, specific treatment implications have become attached to such diagnoses as unipolar depression, acute schizophrenic episode, or elevator phobia. Careful diagnostic delineation is also critical for researchers who wish to study a homogeneous group of patients or who wish to define a group of patients who are comparable to those being studied by a researcher in another setting. Prevention or control must be based on understanding the development and maintenance of a given diagnostic condition. Reliable diagnosis enhances the search for commonalities across individual observations and allows for the development of abstractions not possible in the single case.

A caveat regarding psychiatric diagnosis must be kept in mind. Diagnoses are conventions to be adopted or discarded depending on whether they contribute usefully to functions of administration, treatment, research, or prevention. Like the term *disease*, a given diagnosis may not actually correspond to anything in nature at all and, just as diseases have come and gone, the diagnoses that we presently use may not survive; more useful ones may emerge. Diagnostic nomenclatures represent a way of thinking and communicating with each other. They should not be thought of as defining physical "reality," which will continue to be increasingly differentiated with advances in scientific understanding in the future.

There is a further scientific and pragmatic reason to establish and refine diagnostic procedures and diagnostic-specific treatments. Managers for third-party payers are accumulating data on which treatment interventions work with which diagnostic group of patients, in which treatment setting, in how many treatment sessions, and so on. It seems likely that future treatment authorization will be tied to patient diagnosis and the documented effective treatment for that diagnostic condition. That is, different treatment protocols will be established for different patient diagnoses with corresponding pay schedules. Clinicians will need to know the needs of their patients and what science/research indicates is the preferred treatment. The obvious questions to the clinician will be: (1) What is your diagnosis? (2) What is your planned course of treatment? The opportunity exists now for the professional who has the combination of clinical skills and researcher's logic to help establish the diagnosis and treatment protocols.

Problem Lists

As an alternative and perhaps an extension of the DSM-III and DSM-IV diagnostic schema, Goodman, Brown, & Deitz (1992) developed a *patient impairment profile* in which impairments, behaviorally defined, are rated for severity from imminently dangerous to non-pathological. Goodman and colleagues (1992) recognized that the same or similar treatment procedures might be used by clinicians educated and trained in different disciplines under the guise of different terminology. They attempted to develop a common behavioral language of treatment that would communicate why a particular type of treatment is needed at a given point in time. They defined a list of impairments and concomitant severity levels for each. They further developed treatment suggestions and outcome objectives for each impairment, consistent with both the severity of the impairments and the patient's strengths and limitations. Treatments and outcomes can be measured and reported in terms of the behavioral patient-outcome objectives.

Based on clinical interviewing, the listing of impairments and severity is done by the therapist. Examples of impairments include: altered sleep, dysphoric mood, educational performance deficit, family dysfunction, social withdrawal, suicidal thought/behavior, and others. It is of interest to note that "schizophrenia" is not listed as an impairment because it does not communicate why a particular type of treatment is needed at a given point in time. What would be listed, for example, is "hallucinations" or "social withdrawal" and the severity level for each.

Goal Lists

Based on the premise that clear, therapeutic goals are crucial to psychological counseling Maple (1994) has described and has participated in the development of "the goal-focused therapy approach." Since most problems experienced by individuals who seek psychological counseling occur in their interactions with other people, goal-focused interviewing focuses on formulating therapeutic goals developed from the interaction between therapist (or interviewer) and client(s) (or interviewees). Maple (1994) suggests that clients tend to make fairly specific goal statements while talking about their problems, and that the occurrence of such goal statements increases in response to specific kinds of interviewer inquiries. He has identified a set of specific interviewer skills which encourage client goal statements, and which enhance the interviewer's ability to recognize problematic patterns of interaction. Among the suggested interviewer skills is "formulating goal statements." This involves eliciting client goals and transforming the client's often vaguely stated goal into a well-formed goal statement. A well-formed goal statement is thought to include: (1) an action verb to state what the client wants, and(2) an outcome the client hopes to achieve as a result of the action (e.g., "I want to own a car so I can feel like an adult.").

Additionally, Maple (1994) describes the development of goal-focused interactive videodiscs to improve training for mental-health workers in individual, group, and family treatment-session interviewing. The videodiscs are actual videotaped therapy sessions that also use interactive question-and-answer scripts. The student interviewer can be scored on effectiveness, leadership, and nonverbal-cues exercises. The prime importance for trainers of these structured exercises is the consistent feedback that is provided for student learners.

HISTORICAL FOUNDATION

For many clinicians the most used methods of diagnostic study in the past have been relatively open-ended history taking and the mental status examination, which does introduce some organization into the diagnostic interview and into classification and reporting of the information that is offered by the patient. Although certain information was to be obtained, the clinician was not expected to follow a rigid interview outline.

As will be seen below, it seems clear that reliability in diagnosis would be enhanced by using more structured interviews than has often been the case in the past. Open-ended history taking is likely to omit important questions and leave significant aspects of patient functioning without review. Furthermore, the specific biases of individual clinicians are likely to lead them to over- or underemphasize certain aspects of history taking. Related to this is the fact that an initial impression may lead one to miss diagnostic cues that are contrary to the expectations established on the basis of that first impression. All of us must be aware of how likely we are to see and observe what we are looking for in a clinical interview or any other situation. Subjective impressions have powerful effects. Try as one might to conduct an objective evaluation, an interview is essentially an interpersonal event. Therefore, subjective emotional reactions, whether conscious or unconscious, are inevitable. The clinician who takes a strong like or dislike to a patient must be particularly concerned about this reaction and needs to ask whether it is because the patient is very similar to oneself or just different: not like me.

Ash (1949) showed 40 years ago that the open-ended diagnostic interview was not a reliable instrument across interviewers. Three psychiatrists participated in one interview and made separate diagnoses. There was only 45 percent agreement for major diagnostic categories and 20 percent for specific subcategories. However, as became abundantly clear in hindsight, a major aspect of the disagreement was that the interviewers often did not agree on the symptoms or behaviors diagnostic of a given psychiatric category. Thus, it was necessary to establish agreed-upon criteria before greater reliability in diagnosis could be achieved.

Diagnostic Criteria: DSM-III and DSM-IV

There is little argument about the need for a common vocabulary of psychopathological behaviors and disorders (Siassi, 1984). The general consensus that clinicians need a rational, uniform, and systematic vocabulary led psychiatrists for the past several decades to develop successive versions of the Diagnostic and Statistical Manual, culminating in the DSM-III (APA, 1980) and DSM-IV (APA, 1994), which provide multiaxial diagnostic formulations and are based on a psychobiosocial theory/model of behavior and psychopathology.

The diagnostic criteria in DSM-III were not general descriptions but specific, denotable features designed to assist clinicians in making a diagnosis. DSM-III attempted comprehensively to describe the specifiable features of each of the mental disorders and only rarely attempted to account for how the disturbance came about, unless the mechanism was included in the definition of the disorder. The text in the DSM-III manual began with a clinical description for each psychiatric or psychological disorder, including its essential features, age at onset, course, typical level of impairment, complications, predisposing factors, prevalence, sex ratio, and family pattern. The discussion of each disorder ended with a box summary of the operationally denotable diagnostic criteria for that disorder.

Thus, with the publication and use of the DSM-III, diagnosis was based on specific criteria for each disorder so that when a given diagnosis was used we could know quite exactly what was meant because we knew the precise criteria that guided the diagnostician. Since each diagnostic entity was based on specific information, the interviewer had to proceed in a way that would allow those details to be obtained. Generally, this meant that interviewing had to be more focused. Furthermore, the interviewer usually had to obtain longitudinal as well as cross-sectional data. Duration of symptoms is a diagnostic criterion for a number of mental dis-

orders in DSM-III. Of course, this longitudinal focus also allowed the interviewer to search for associations between life events (stressors) and symptoms.

Diagnostic Interview Schedules

In thinking about clinical criteria for diagnosis and sources of unreliability in diagnostic formulations, Spitzer, Endicott, & Robins (1975) noted five sources of unreliability and then determined that two of these contributed most heavily to diagnostic unreliability. The first source of unreliability they noted was *subject variance*, which occurs when patients actually have different conditions at different times. They gave the example of the patient who may show alcohol intoxication on admission to a hospital but develop delirium tremens several days later. A second source of unreliability is *occasion variance*, which occurs when patients are in different stages of the same condition at different times. An example of this would be a patient with a bipolar disorder who is depressed during one period of illness and manic during another. A third source of unreliability is *information variance*, which occurs when clinicians have different sources of information about their patients. Examples here include clinicians who talk with patients' families and those who do not, or interviewers who question patients about areas of functioning and symptoms about which other interviewers do not. A fourth area of unreliability is *observation variance*, which occurs when clinicians notice different things although presumably observing the same patient behavior. Clinicians may disagree on whether a patient was tearful, hard to follow, or hallucinating. A fifth source of unreliability is *criterion variance*, which occurs when clinicians use varying diagnostic criteria (e.g., whether a formal thought disorder is necessary for the diagnosis of schizophrenia or precludes a diagnosis of affective disorder). Spitzer and collaborators (1975) concluded that the largest source of diagnostic variability by far was criterion variance. Their efforts on behalf of the development of DSM-III diagnostic criteria obviously reflected their confidence in this conclusion.

Their research efforts to reduce information variance (the second most important source of unreliability) led to the development of structured clinical interviews that reduce that portion of the unreliability variance based on different interviewing styles and coverage. The Research Diagnostic Criteria, or RDC, (Spitzer, Endicott, & Robins, 1978) provide sets of specific inclusion and exclusion criteria for a large number of functional disorders, with particular emphasis on various ways of subtyping affective disorders. In following RDC, the clinician is required to use these criteria regardless of his or her own personal concept of the disorder. With this approach, the clinician's task is: (1) to determine the presence or absence of specific clinical phenomena, and (2) to apply the comprehensive rules provided for making the diagnosis. A single patient can be categorized in different ways, such as by the presence or absence of endogenous psychopathology, situational stresses, psychotic features, and the like. The kappa values for the RDC were usually above .70 and frequently above .80, and represent impressive levels of agreement (Spitzer, Forman, & Nee, 1979). The RDC is an excellent tool available also to the researcher who wishes to study homogeneous patient groups. It is one of about ten structured interview guides distributed through Biometric Research of Columbia University.

Siassi (1984) concludes that the structured psychiatric interview has already become the foundation of much modern clinical research and that the clinical psychiatric interview and mental status examinations, as used in the past, will likely be replaced in the future with the use of structured interview schedules for routine psychiatric examinations. This shift is supported by trends toward the use of operational criteria for diagnosis, well-defined taxonomies, almost exclusive use of structured examinations in research settings, and the growing influence of clinician-researchers. Further, the demand for accountability has also forced a problem-oriented type of record-keeping system in most institutions, with emphasis on branch-logic systems of psychiatric decision making, and progress notes that reflect resolution of symptom-syndromes and changes in problem status, rather than changes in psychodynamics. Finally, the impact of computers appears decisive in that they allow for efficient retrieval of information, unlike the narrative psychiatric interviews. Computers can also be used to apply an algorithm to yield highly reliable diagnoses from raw data (Siassi, 1984, p. 272).

The nature of a structured clinical interview is discussed by Edelbrock & Costello (1984) who note that it is essentially a list of target behaviors, symptoms, and events to be covered, and some

guidelines or rules for conducting the interview and recording the data. Interview schedules vary in that some offer only general and flexible guidelines and others have strict and detailed rules, that is, some are *semi-structured* and others are *highly structured*. With the latter, wording and sequence of questions, recording responses, and rating responses are all specified and defined. The interviewer may be regarded as an interchangeable piece of the assessment machinery. Clinical judgment in eliciting and recording information is minimized and, given the same patient, different interviewers should obtain the same information. Clinical judgment may play more of a role in the semi-structured interview with more latitude about what is asked, how it is asked, and how it is recorded. Edelbrock & Costello (1984) suggest that both types of interviews have some advantages. Highly structured interviews reduce the role of clinical inference and interpretation in the assessment and diagnostic process, and they typically yield more objective and quantifiable raw data. Alternatively, semi-structured interviews are less stilted and permit a more spontaneous interview that can be tailored to the patient.

Edelbrock & Costello (1984) also conclude that structured interviews are here to stay, that they will become the standard assessment and diagnostic tools in clinical research and epidemiology and, that they will become more closely integrated into the training of mental health professionals and the delivery of service. They also predicted that the interview would continue to evolve along the lines of increasing specialization of purpose, coverage, age, range, degree of structure, and interviewer qualifications. As diagnostic taxonomies evolve and become more differentiated, structured interviews will necessarily change in terms of their content. We can also expect results obtained via structured interviews to precipitate change in the diagnostic systems. They noted another significant development: namely, the synergistic combination of structured interview data with data derived from other assessment methods such as check lists, rating scales, and self-report inventories. They expect multi-method assessment to yield a more comprehensive, reliable, and valid picture of the patient. Finally, they saw a significant trend toward computer-assisted diagnosis, especially the use of the computer to sift through numerous bits of data relevant to diagnostic decision making.

Computer Development

Computers have long played a significant role in assessment. Much modern test construction has been dependent on the availability of computing resources. As test administration itself became more feasible with the advent of microcomputers, one of the questions raised was the comparability of data obtained with traditional paper-and-pencil administration and computerized administration. Lukin, Dowd, Plake, & Kraft (1985) obtained no significant differences between scores on measures of anxiety, depression, and psychological reactance across administration format. Most important, while producing results comparable to the pencil-and-paper assessment, the computerized administration was preferred over the pencil-and-paper administration by 85 percent of the subjects.

Ferriter (1993) compared three interview methods for collecting information for psychiatric social histories: human unstructured interviewing, human structured interviewing, and the same structured interview delivered by computer. He concluded that structured interviewing collected significantly more information than unstructured interviewing. A comparison of structured human and computer interviews showed greater extremes of response with fewer discrepancies of fact in the computer condition, indicating greater candidness of subjects in that group, and therefore, greater validity of data collected by computer.

Choca & Morris (1992) compared a computerized version of the Halstead Category Test to the standard projector version of the test using neurologically impaired adult patients. Every patient was tested with both versions and the order of administration was alternated. Results indicated that difference in mean number of errors made between the two versions of the test was not significant. The scores obtained with the two versions were seen as similar to what would be expected from a test-retest administration of the same instrument. The authors note that one advantage of the computerized version is that it assures an error-free administration of the test. Secondly, the computer version allows the collection of additional data as the test is administered, such as the reaction time and the number of perseverations when a previous rule is inappropriately used. Finally, it may be eventually possible to show that promptings from the examiner do not make a significant difference in terms of the eventual outcome. If this were the case, the computer version would have the added

advantage of requiring a considerably smaller time commitment by the examiner (Choca & Morris, 1992, p. 11-12).

These studies, and others not reviewed here, support the contention that computerized testing techniques provide results comparable to traditional assessment techniques when using individual tests.

While psychological software has not kept pace with hardware development, the availability of new programs of interest to psychologists and other clinicians has been dramatic. Samuel E. Krug (1987) compiled a product listing that included more than 300 programs designed to assess or modify human behavior. Of these listings, eight percent were categorized as structured interviews and were most likely to be described as intake interviews. Krug (1993) subsequently has compiled a similar product listing that now includes more than 500 programs designed to assess or modify human behavior. Of these listings about 11 percent are categorized as structured interviews. The products in this category almost always are designed to be self-administered.

One of the earlier proponents of automated computer interviewing, John H. Greist (1984), observed that clinician training, recent experience, immediate distractions, and foibles of memory are among the factors that may compromise our competence as diagnosticians. Further, he stated that in virtually every instance in which computer interviews and clinicians have been compared, the computer outperforms the clinician in terms of completeness and accuracy. Erdman, Klein, & Greist (1985) suggest that one appeal of computer interviewing is the ability of the computer to imitate, even if only to a limited degree, the intelligence of a human interviewer. Like a human interviewer, the computer can be programmed to ask follow-up questions for problems that the respondent reports, and to skip follow-up questions in those areas of no problem. This branching capability leads to an interaction between computer and human, that is, what happens in the interview depends on what the subject says. Thus, a computer interview can be tailored to the person using the program, for example, not to ask a male subject about pregnancy. Of more interest though is the capacity to ask follow-up questions in the subject's own words and to compare responses from different points in the interview. While it has been asserted that computers cannot detect flat affect, Erdman, Klein, & Greist (1985) do note that it is possible to record response latency and heart rate simultaneously and to use these variables to branch into questions/comments regarding emotional arousal. It does seem clear, however, that to date it has not been possible for the computer to report the many nonverbal cues to which a human interviewer could observe and respond. To be candid, however, it must be acknowledged that a human interviewer also remains oblivious to a great deal of information available in a two-person interaction.

DESCRIPTION OF ASSESSMENT METHODS

There are some general guidelines that should be considered when evaluating structured interviews. As noted by Spiker and Ehler (1984), these include the following: (1) sources of information should be specified in an effort to reduce information variance; (2) terms should be defined so that interviewers are consistent in their usage of them; (3) guidelines for determining the presence or absence of specified signs and symptoms should be given; (4) questions should be specified to ensure that necessary information is obtained to determine whether criteria for a given diagnosis have been met; and, (5) information gathered should be in such a format that a given set of data will consistently lead to a given diagnosis.

Young and coworkers (1987) point out that efforts to reduce rater variability have formally consolidated into joint training, testing, and calibration of interviewers using standard procedures. The elements of the training programs will vary, particularly according to whether or not the interviewer is expected to formulate diagnoses. All training procedures attempt to ensure that interviewers have the skills to elicit and record the required information accurately and efficiently. This training involves monitoring interviews. If, in addition, the clinician/interviewer is to produce a diagnosis, it is necessary to know the diagnostic criteria thoroughly, have experience with differential diagnosis and have a well-developed understanding of the clinical manifestations that determine severity and clinical significance (Young et al., 1987, p. 617). A training program could include progressive steps such as studying sample cases, videotaped and live interviews by trainees, and providing continual monitoring to maintain reliability. In the case of computer-

administered interview schedules, a different criterion applies: namely, that the computer program has been sufficiently de-bugged so that it runs without error. This assumes, of course, that well-conceptualized questions went into the program in the first place.

Behavior checklists and patient questionnaires provide useful information to answer diagnostic questions, are often easy for different clinicians or even technicians to use, and often present data in such a way that it can be easily computer coded. Used by themselves, behavior checklists typically allow for only fairly general observations of an interviewee, so that additional procedures or clinician input is needed to arrive at a specific diagnosis.

The assessment methods described below were selected to be illustrative of a number of structured interview applications. They are by no means an exhaustive listing or review. Included are general diagnostic interviews, psychosocial history interviews, specific-purpose interview schedules, and several behavior checklists and patient questionnaires.

Diagnostic Interview Schedule (DIS)

The DIS (Robins, Helzer, Croughan, & Ratcliff, 1981) is a fully structured interview schedule designed to enable clinicians to make consistent and accurate DSM-III psychiatric diagnoses. A more recent version, the Diagnostic Interview Schedule, Version III-Revised (DIS-III-R) (APA, 1987) incorporated DSM-III-R diagnoses into its structure. The DIS was designed to be administered by persons not professionally trained in clinical psychiatry and all of the questions and the probes to be used are fully explained. It reminds interviewers not to omit critical questions, and presents well-tested phrasing for symptoms that are difficult to explain or potentially embarrassing to patients. Questions about symptoms cover both their presence or absence and severity, for example, taking medication for the symptom, seeing a professional about the symptom, and having the symptom significantly interfere with one's life. In addition, the interview ascertains whether the symptom was explained entirely by physical illness or injury, or as a complication of the use of medication, illicit drugs, or alcohol. The age at which a given diagnostic symptom first appeared is also determined along with the most recent experience of the symptom. These questions are designed to help determine whether a disorder is current, that is, the last two weeks, the last month, the last six months, or the last year. Demographic information including age, sex, occupation, race, education, marital status, and history of treatment is also determined. Current functioning is evaluated by ability within the last 12 months to work or attend school, maintain an active social life, act as head or co-head of a household, and get along without professional care for physical or emotional problems.

Aside from a few open-ended questions at the start of the interview to allow the interviewee the opportunity to voice the chief complaint and to give the interviewer some background for understanding answers to closed-ended questions, the interview is completely precoded. Symptoms assessed by the computer are precoded at five levels: (1) negative, the problem has never occurred; (2) present, but so minimal as to be of no diagnostic significance; (3) present and meets criteria for severity, but not relevant to the psychiatric diagnosis in question because every occurrence resulted from direct or side effects of prescribed, over-the-counter, or illicit drugs, or alcohol; (4) present and meets criteria for severity, but not relevant to the psychiatric diagnosis in question because every occurrence resulted from medical illness or injury; (5) present, meets criteria for severity, and is relevant to the psychiatric diagnosis under consideration.

Richard Rogers in his book, *Diagnostic and Structured Interviewing* (1995) devotes an entire chapter to the DIS. He describes it in considerable detail and discusses its rationale and development. He also reviews data from studies on its reliability, criterion-reliability, validity of specific disorders, and validity of alternate versions.

The DIS has been translated into different languages and its use is now underway, or planned, in about 20 different countries. Cross-national comparisons in psychiatric epidemiology are possible due to the growing number of population surveys in various countries that have used the DIS. For example, to determine the prevalence of psychiatric disorders within a large Australian community sample, Clayer, McFarlane, & Wright (1992), used a computerized version of the Diagnostic Interview Schedule Screening Interview. The study of 1,009 Australians found that this technique provided prevalence estimates similar to those obtained from the more established DIS, with the

added benefit that all interviews were successfully carried out in the examinee's home.

In another study, Hollifield, Katon, & Morojele (1994) examined anxiety and depression in an outpatient clinic in Lesotho, Africa using a translated version of the DIS. They randomly selected 126 outpatients (response rate=77 percent) attending a general hospital clinic and found 29 (23 percent) of the patients with depression, 30 (24 percent) with panic disorder, 36 (29 percent) with generalized anxiety disorder. Patients with depression or panic disorder presented with a significantly higher number of physical symptoms. A primary conclusion was that there is significant psychiatric morbidity in outpatient medical clinics in Lesotho, and patients present primarily with somatic symptoms, as in developed countries.

Other studies using the DIS have examined Vietnamese-speaking medical patients in the United States (Lee & Chan, 1986), major depression in Cuban Americans and Puerto Ricans (Cho, et al., 1993), and alcohol drinking and alcoholism in Shanghai, China (Wang, Liu, Zhang, Yu, Xia, Fernandez, Lung, Xu, & Qu, 1992). These studies have primarily focused on the assessment of adults. Other studies have focused on the assessment of children. Rubio-Stipec, Canino, Shrout, Dulcan, Freeman, and Bravo (1994), using a Spanish version of the DIS for Children, showed that parents and children provided unique information when interviewed with a structured psychiatric interview about child psychopathology.

Since the early development of clinical structured interviews some of the more commonly used instruments have gone through several revisions, expansions, and translations. For example, the Structured Clinical Interview for the DSM-III-R (SCID) has been translated into Chinese to assess differences in diagnostic practices between Western and Chinese psychiatry (Wilson & Young, 1988). Similarly, a Dutch version of the SCID for Dissociative disorders (SCID-D) has been used to improve assessment and diagnosis of dissociative symptoms and disorders in The Netherlands (Boon & Draijer, 1991). The SCID-D is now available in a revised version (SCID-D-Revised, 1994).

The Composite International Diagnostic Interview (CIDI) (WHO, 1993) has been designed for use in a variety of cultures (the core version is currently available in 16 languages). It is intended primarily for use in epidemiological studies of mental disorders but it can also be used for other clinical and research purposes. It is available in different versions and can be supplemented by modules for diagnoses not covered in the core version (Robins et al., 1989). (Both the SCID-D and the CIDI are available through the American Psychiatric Press, Inc., 1400 K Street, N.W., Washington, DC 20005.)

One of the principal instruments to come out of a pilot study in 1976 by the World Health Organization (WHO, 1984) was the WHO Psychiatric Disability Assessment Schedule (WHO/DAS). This schedule was used to record information about the patients' functioning and some of the factors that might influence it. The collaborating investigators finalized a version in 1984, after the completion of the field studies. Since that time, this instrument has been used in many other studies, both within and outside the framework of the WHO mental health program and in over 20 countries. In addition to English, the schedule is available in Arabic, Bulgarian, Chinese, Danish, French, German, Hindi, Japanese, Russian, Serbo-Croat, Spanish, and Urdu. (Anyone wishing further information on the use of the schedule, including details of training material, could contact the Division of Mental Health, World Health Organization, 1211 Geneva 27, Switzerland.)

Other well-known structured diagnostic interviews include the *Present State Examination* (Wing, Birley, & Cooper, 1967), the *Renard Diagnostic Interview* (Helzer, Robins, Croughan, & Welner, 1981), and *The Schedule for Affective Disorders and Schizophrenia* (Endicott & Spitzer, 1978).

Psychosocial History Interviews

While psychological testing and laboratory studies may be elective, there is scarcely a patient we see on whom we fail to gather considerable background information. The psychosocial history is seen as indispensable in the proper evaluation of a patient and as having a central role in clinical practice (Giannetti, 1987). The psychosocial history is seen as providing a biographical-historical perspective of the personality, the stresses and realities within which a person lives, and the nature of the relationships with those closest to the individual, that is, the patient's individuality. It is an effort to get to know the particular individual who is presenting with a given problem. Psychological tests provide standardized estimates on a set of variables (personality, intellectual, symptomatic) against

normative standards. The psychosocial history, on the other hand, provides information on the long series of external stimuli, events, and individuals with which the person has interacted, including the consequences of those interactions (Giannetti, 1987, p. 125).

The *Giannetti On-Line Psychosocial History (GOLPH)* (Giannetti, 1985) was written after reviewing clinical history outlines obtained from different service settings. A content analysis of the items from these outlines suggested that they could be sorted into 10 categories that were reasonably mutually exclusive and exhaustive (Giannetti, 1987):

1. Identifying demographic data/current living arrangements
2. Family of origin
3. Client development
4. Educational history
5. Marital history/current family functioning
6. Occupational history/current financial circumstances
7. Military history
8. Legal/criminal history
9. Physical illnesses/current somatic symptoms
10. Psychological symptoms/treatments

After these general categories were identified, items that could be reasonably obtained by self-report were written. For example, items in 14 content areas were written for the general category of developmental history. The response alternatives to each item were researched so that they would be exhaustive, and the patient, and later the computer printout, would have a minimum of "other" responses. Giannetti (1987) noted that they were able to limit "other" responses to approximately 2.5 percent of 2,400 response alternatives.

The GOLPH was designed to be administered after the clinician had met with the patient initially to discuss the reasons for seeking treatment. Knowing that a reasonably comprehensive historical review will be obtained using the structured psychosocial interview, the clinician is free to focus on the chief complaint and on establishing a therapeutic relationship. During the initial contact the clinician can also determine whether the patient would be unable to provide a valid self-report because of attention or memory deficits, psychosis, extreme anxiety, psychomotor retardation, or some other reason. The GOLPH is intended for individuals at least 16 years old, having a sixth-grade reading level, and seeking mental-health services. It is estimated that most patients can complete the GOLPH in about the time it takes to complete an MMPI; administration can be interrupted and resumed. The program prints out a 3 to-12-page report; the first section is a narrative history under the general category headings noted above. It selects the appropriate personal pronouns and develops a sequence of response statements that turn into a narrative that is quite easy to read. The second section of the report is the follow-up summary, in which the data are arranged into a format that allows for easy follow-up of positive symptoms and for consideration of the differential diagnostic implications of these data. The general history items in the first section of the printout makes the GOLPH pertinent to any situation requiring a history. The second section of the report is particularly apt for a general psychiatric population, but it could be modified to specific clinical situations (e.g., behavioral medicine or neuropsychology) or to nonclinical evaluative contexts, such as in organizational or industrial settings (Giannetti, 1987, p. 142). (The Giannetti On-Line Psychosocial History (GOLPH) is classified as a clinical assessment/structured interview and is distributed by National Computer Systems, P.O. Box 1416, Minneapolis, MN 55440.)

The *Psychological/Social History Report 4.0* (PSH) is a structured intake interview that the patient can complete either at a computer keyboard or on a questionnaire. The patient responses, if recorded on a questionnaire, are entered into the computer by the clinician. Basic information is obtained on the patient's presenting problem(s), emotional status, developmental history, education, finances, employment history, alcohol/drug use, health, diet/exercise, stressors, and interpersonal relations. An important advantage of the PSH is that the patient is queried systematically in all these areas, thus reducing the possibility of clinician errors of omission in data gathering.

The PSH generates a narrative report for each of the topic headings. Shown below are two computer-generated narrative reports for "Family and Developmental History."

> Mr. Doe was raised primarily by his father. In retrospect he describes his childhood as being happy, and secure. Mother was characterized as warm, and distant. He describes his father as warm. Characteristics of his parental caretakers' relationship were given as follows: reserved, and happy. There were no other

children in the family. As a child, Mr. Doe was characteristically outgoing, happy, friendly, and emotional. He was an only child. The following problems occurred during childhood: having feelings hurt, and fear of failure. As a child Mr. Doe's father worked primarily in government service and his mother worked primarily as an executive. Mother's method of discipline is described as lenient and father's as fairly strict. Childhood fears included: none. Sexual experiences are reported to have been pleasant.

Ms. Doe was raised primarily by her natural parents. In retrospect she describes her childhood as being frightening, unhappy, and painful. Mother was characterized as distant, uncaring, strict, unpleasant, domineering, abusive, and faultfinding. She describes her father as warm, and strict. Characteristics of her parental caretakers' relationship were given as follows: cold, violent, indifferent, full of conflict, and hostile. There was one other child in the family. As a child, Ms. Doe was characteristically shy, active, emotional, nervous, and unhappy. She was the oldest child. The following problems occurred during childhood: mother, sibling(s), excessive fears or worries, academic, feeling a burden to parents, and having feelings hurt. Parental caretakers argued about drinking. As a child Ms. Doe's father was disabled and her mother worked primarily as an office worker. Mother's method of discipline is described as strict and father's as strict. Childhood fears included: being laughed at, and other children. Sexual experiences are reported to have been unpleasant.

The PSH can be administered on the computer or pencil-and-paper format and takes from 30 to 45 minutes to complete. It is for ages 17 years and older. (It is distributed by MHS, 908 Niagara Falls Blvd., North Tonawanda, New York, 14120-2060.)

The *Psychological/Psychiatric Status Interview* (PPSI) is designed for on-line computer administration of an initial psychological/psychiatric interview. The program interviews the patient with respect to presenting problems, current living situation, mental status, biological/medical status, interpersonal relations, and socialization. This program is designed specifically for computer administration of the interview and cannot be completed in a paper-and-pencil format. The PPSI provides the clinician with an organized database on the client that can be reviewed prior to a personal interview. The three-to-five-page report may be printed or written to a text file. This structured interview essentially assumes that the respondent has come to a clinical service setting for help with a psychological/psychiatric problem. As noted above, a section of the narrative printout is devoted to a discussion of "presenting problem." With this particular interview, the respondent is also allowed to type into the program a narrative description of the problem that is at issue. The PPSI is also a combination of psychosocial history and diagnostic interview.

A unique feature of this structured interview is how it goes about asking questions to complete a "mental status examination." A sample "mental status" section is reproduced below; the reader will be able to infer the questions that are the basis of the descriptive narrative.

John stated that he is oriented to person, place, and time. John indicated subjective experiences suggestive of attention/concentration difficulties. It is difficult for him to concentrate on what's going on around him. He indicated his attention often wanders. Recent difficulty thinking and concentrating was indicated. Perceived memory difficulties may be present. He endorsed the statement: "I have problems remembering things." He indicated he can remember things that happened in the past better than things that happened recently. John reported an inability to remember what happened for a period of several hours/days of his life.

When visually presented three unrelated words (snake-city-priest) for two seconds each, John required only one presentation of the words before he could correctly identify all three words. When asked to identify the same words later (after approximately 15-20 minutes with no re-presentation of the words) John was not able to correctly select the word-triad from a set of six word-triads. This may suggest deficits in short term verbal memory.

John rated his intellectual ability as average. He indicated that he believes his current intellectual functioning is below his level of functioning in the past. When asked what the saying, "People who live in glass houses shouldn't throw stones" means, John endorsed the response "Stay out of arguments." His response to how a dog and a cat are alike was: "They are both animals". The endorsed answer for how reward and punishment are alike was: "I have received both." Some difficulty was encountered performing simple addition. His endorsed responses to the addition problems were: 7 + 4 = "11;" 29 + 14 = "42." Some difficulty was encountered with simple subtraction. Answers given to subtraction problems were: 9 - 3 = "6;" 45 - 27 = "don't know." John was unable to perform simple multiplication (6 x 7 = "don't know"). He correctly solved a simple division problem (28 / 7 = "4"). Possible hallucinatory experiences may be present. He reportedly sometimes hears voices or noises that other people can't hear. He indicated that his imagination often plays tricks on him. Further exploration of these experiences to ascertain if they represent hallucinations is indicated.

The possibility of delusional thinking is suggested. He indicated that he holds certain beliefs that everyone else tells him are not true. He reported that he is not who people think he is. These beliefs need to be

further discussed with John to find out if there is a delusional component to his thought processes.

Unusual thought content may be present. He indicated that recently he has had unusual or strange ideas. He has had repetitive thoughts that he can't get out of his mind. He indicated the presence of suicidal ideation. He often believes things are not real. He indicated that other people believe he has a strange and unusual way of thinking. John denied the presence of any extreme and unrealistic fears.

John described his current emotional state as "depressed." The intensity of this feeling was described as extreme and he experiences this feeling throughout the day. He indicated that the intensity level of the feeling does not vary as a function of the time of day. Specific events that intensify the feeling include being alone. He has had this feeling for three to six months. No additional current emotions were reported. John denied that he has had periods in which he felt very anxious and fearful for no apparent reason. An episode of depression for no apparent reason was reported. The depression reportedly lasted for six to twelve months. He indicated that the most recent or current depressive emotional state include feeling "slowed down," doesn't enjoy doing anything, feeling tired most of the time, feelings of worthlessness, very pessimistic about the future, and he thinks a lot about death. He reported that he recently has had thoughts about killing himself. He admitted that he has threatened/attempted suicide in the past. He did not endorse symptoms suggestive of a current or past period of mania or hypomania. Reported subjective experiences possibly indicating an affective disturbance include trouble thinking and concentrating and feeling very irritable.

Problematic behavior patterns may be present. He admitted that he has periods when he feels compelled to spend money. He indicated that he has on at least one occasion spent so much money at one time that it created serious financial problems. He reported that he gambles frequently. He admitted that gambling has created problems in his life. He indicated that he often does things without thinking of the consequences.

The Psychological/Psychiatric Status Interview is distributed by Psychologistics, Inc., P.O. Box 033896, Indialantic, Florida 32903. This company also distributes several other structured interviews and report formats that are of interest to clinicians. One of these is the *Intake Evaluation Report-Clinician Version 3.0*. The clinician is provided with a comprehensive checklist to use as a guide when evaluating the patient with respect to presenting problem, current situation, physical presentation, mental status, biological/medical status, interpersonal relations and socialization, diagnostic impressions, and recommendations. The checklist data is then typed into the computer program that organizes the obtained information in a manner useful for case conceptualization and treatment planning. The report may be printed out or written to a text file where, using a word processor, it may be reformatted or revised to meet specific needs of the clinician.

Another program of interest is *Session Summary,* designed to aid the clinician in completion of case notes and documentation of treatment. It may be completed by the clinician directly on the computer or by paper-and-pencil checklist. The checklist is included on the program disk and may be printed out by the user. The program generates a one-page narrative summary of the session.

The *Termination/Discharge Summary* was designed to assist the clinician in developing a concise, yet comprehensive, summary of the patient's evaluation/treatment. The program summarizes information in the areas of presenting problems, initial mental and physical status, evaluation results, goals of treatment, outcomes of treatment, and termination or discharge recommendations. Changes in problem focus and/or intervention strategies are also documented. The summary may be completed directly on the computer or by paper-and-pencil checklist. The program generates a two-page narrative report providing documentation of the patient's treatment.

In his psychware sourcebook, Krug (1993) has listed 62 instruments under the heading of structured interviews. We have noted only a few in this chapter. Other psychosocial history interviews in addition to the ones noted above are included in his listing. He also includes a number of specific-purpose interviews.

Specific Purpose Interview Schedules

Thus far, the discussion in this chapter has focused on patients most likely seen in mental-health practices. The assessment of such patients has general significance for all patients of course, as one can obtain information about overall level of emotional health and the possible presence of such universal and potentially debilitating symptoms as anxiety and depression. Fortunately, assessment procedures have been developed with other patient populations in mind as well.

The *Comprehensive Drinker Profile* (CDP) was first developed in 1971 as a structured intake-interview procedure for assessing alcoholism in male inpatients. The CDP is appropriate for use with

both male and female clients entering any of a wide variety of treatment modalities in either inpatient or outpatient settings (Miller & Marlatt, 1984). The CDP provides an intensive and comprehensive history and status of the individual client with regard to his or her use and abuse of alcohol. It is intended to be administered as a structured clinical interview and normally requires one to two hours for completion. If necessary, however, the interview can be completed over more than one session. The authors note that for some clinical and research purposes it may be desirable to corroborate client self-report by interviewing collateral sources such as family and friends.

The *Type A Structured Interview* was developed by Friedman and Roseman (1974) to elicit characteristics of the Type A syndrome. It consists of 22 questions and takes about 10 minutes to administer. Supervised training in administration is suggested because the interviewer must assess not only the specific content of answers, but also the general stylistics and mannerisms of the individual as he or she answers the questions. How something is said may be more important than what was actually said in assessing Type A characteristics.

The *Psychosocial Pain Inventory* (PSPI) was developed to provide a standardized and reliable method of evaluating a number of psychosocial factors considered to be important in maintaining and exacerbating chronic pain problems (Heaton, Lehman, & Getto, 1980). The PSPI includes evaluation of the following factors: several forms of secondary gain; the effects of pain behavior on interpersonal relationships; the existence of stressful life events that may contribute to subjective distress or promote avoidance learning; and components of past learning history that familiarize the patient with the chronic invalid role and with its personal and social consequences. PSPI ratings also consider the fact that patients differ in the degree to which they are likely to be influenced by potential sources of secondary gain.

One section of the PSPI considers home or family responsibilities the patient discharged prior to the pain problem as compared to now. Comparison is made from "before" to "now" for those activities that were primarily the patient's responsibility; responses are graded from "less now" (i.e., the activity has decreased by no more than 25 percent) to "never now." Areas of activity the patient is asked about include: housecleaning, clothes washing, clothes ironing, shopping, cooking, repair work (home), repair work (car), yard work, errands, caring for children, disciplining children, driving other family member, family finances, family correspondence, and others (specify). Early findings suggested that the PSPI has some value in predicting response to medical treatment for pain.

Behavior Checklists and Patient Questionnaires

As in each of these areas of discussion, the identification of structured information-gathering devices is illustrative and not exhaustive. This is especially true in the area of behavior checklists and patient questionnaires; literally, hundreds of them have been developed. Some devices that have been used to show changes in patients over time, especially as such changes might be related to treatment interventions, include observer-rating scales and self-descriptive inventories. Examples of observer-rating scales for adult patients are the *Brief Psychiatric Rating Scale* (Overall & Gorham, 1962), the *Hamilton Depression Scale* (Hamilton, 1967), the *Wittenborn Psychiatric Rating Scale* (Wittenborn, 1955), and the *Nurses' Observation Scale for Inpatient Evaluation* (Honigfeld & Klett, 1965).

Examples of self-descriptive inventories include the *Beck Depression Inventory* (Beck, Ward, Mendelson, Mock, & Erbaugh, 1961), the *Hopkins Symptom Checklist* (Derogatis, Lipman, Rickels, Uhlenhuth, & Covi, 1973) and the *State-Trait Anxiety Inventory* (Spielberger, Gorsuch, & Lushene, 1970). As the reader may reflect, questionnnaires have been developed for many target-specific adult behavior problems, including: social-skills deficits, obsessive-compulsive behaviors, fears and phobias, anger, marital distress and dysfunction, ingestive disorders, sexual dysfunctions, and so on.

THEORETICAL AND METHODOLOGICAL CONSIDERATIONS

Much clinical research is difficult to do if reliable diagnoses cannot be achieved and clinicians disagree on patient classification. When researchers looked for sources of undependable information, or unreliability in diagnosis, it became apparent that *criterion variance* (application of different rules for assigning a diagnosis) and *information variance* (different interview data) were prime sources of unreliability. Continuing efforts to

develop and improve specific, operationalized criteria for diagnoses have been broadly accepted and applied by clinicians. Paradoxically, procedures to clarify sources of erratic data in the interview by standardizing the form of interviews have not been widely applied (Young et al., 1987).

Issues in Diagnosis

Diagnostic nomenclatures (e.g., DSM-III) represent a way of thinking and communicating with each other. The DSM-III has provided a uniform and systematic vocabulary, and clinicians using its diagnostic criteria have achieved good agreement in diagnosis. Diagnostic agreement is achieved by using operationally defined criteria for the presence of a disorder and essentially substituting operational definitions for the varying and subjective judgments of clinicians. Parenthetically, it must be recognized that subjective judgments are never totally obviated because even in deciding whether operational criteria have been met (e.g., was the patient inattentive?) subjective judgment is exercised. However, the more general issue here is the limitations inherent in operational definitions, namely, that reliance on a single operational definition confines our understanding of a concept (or diagnosis) to those aspects monitored by the particular set of rules, and restricts opportunities to improve the rule system itself. By analogy, many psychologists are aware of the statement that "intelligence is what the intelligence test measures" and how important it has been, in developing concepts of cognitive functioning, to use procedures for study in addition to those included in intelligence tests. Research methodologists warn against the use of single operational definitions in measurement systems. Each operational definition is inherently imperfect. Otherwise the construct measured is self-evident and does not require the development of any measures. The use of multiple measures avoids reliance upon an assumed final, perfect operational definition and provides a framework for illuminating comparisons among measures. Using both structured interviews and rating scales provides a cross check so that errors and biases in each can be identified. In a similar fashion, the use of multiple probes (derived from operational definitions) within a structured interview improves reliability and validity (Young et al., 1987).

Gold (1986) acknowledges that the DSM-III has enormous influence and that its diagnostic codes are used by all mental-health providers whose services are reimbursable by insurance companies. However, he is concerned that it is a reductionistic, symptom-based system that discourages practitioners from looking beyond the obvious. He notes, for example, that without objective verification one cannot be sure of the correct diagnosis if the symptoms of two disorders are similar. The example he uses is that of a grandiose delusion which could be seen in a manic or schizophrenic episode and that the correct differentiation is essential to institute the right treatment. He further observes that patient report regarding symptoms like "appetite" or "pain" depends on whether a patient is grateful for no appetite or has experienced pain so long it no longer seems unusual. Patients of differing ages and backgrounds display differing symptoms, for in sickness and in health, behavior is shaped by generation and culture, even by sex (Gold, 1986, p. 52).

Gold's major point is that the DSM-III encourages its users to believe that behavioral symptoms necessarily mean something psychiatric and leads clinicians not to consider organic conditions that can mimic psychiatric illnesses. Further, he asserts that the DSM-III categories do not parallel the biological subtypes that are being revealed in laboratory research (as in biological differences among depressed people). In the case of depression, it is necessary to differentiate primary and secondary affective disorders and to recognize that systemic medical diseases, CNS disorders, endocrine disorders, drug-induced disorders, and infections are major bases for secondary affective disorder. Gold suggests that at least 75 illnesses or conditions can cause symptoms of apparent mental disorder and, very importantly, that psychiatric symptoms are often the first and only signs of a developing illness. In the diagnosis of cancer, he notes that many types of tumors throughout the body can exhibit mental symptoms, which may be the only symptoms to appear for weeks, months, or years. In fact, he asserts that anyone who has an abrupt personality change, depression without a history of mood disorders, or weight loss of greater than twenty pounds—or who is unresponsive to standard psychiatric treatment—should be evaluated for cancer or other mimickers (Gold, 1986, p. 83). The most common mimickers of psychiatry, according to Gold (1986), are: drug (illicit, prescribed, and over-the-counter) and alcohol reactions; endocrine

disorders; and, diseases of the central nervous system, infectious diseases, cancers, metabolic conditions, and nutritional and toxic disorders. Drugs, for example, must be considered in all psychiatric diagnoses, no matter how classically "psychiatric" the person may appear, because everybody takes them in one form or another. Since the brain is quicker to react than the rest of the body, mental and behavioral symptoms may outweigh organic signs as an indicator of reaction to environmental toxins. The generalization to be drawn for the construct of diagnosis, in the case of depression as an example, is that no one can conclude that a patient is in the midst of a major depressive episode without first ruling out possible organic causes.

Related to this generalization is the obvious truism that symptoms, particularly emotional symptoms, are not specific. First of all, the patient report of something like depressive symptoms is colored by the emotion itself, and, secondly, the depressive emotion may relate to infectious mononucleosis, a bad marriage, an enzyme deficiency, or other etiologies. Clearly, objective measures are needed to verify, or clarify, particular diagnoses. Laboratory tests have come to play a more important role in psychiatry in screening for medical illness, improving diagnostic reliability, monitoring treatment (especially through measurement of the blood levels of psychoactive drugs), and continuing research into mental illness. Kaplan & Sadock (1988) note a number of neuroendocrine tests (used particularly in depression), tests for sexually transmitted diseases, tests to assess plasma levels of psychotropic drugs, electroencephalography, evoked potentials, radioisotope brain scanning, and tests of regional blood flow.

In a paper on brain imaging in psychiatry, Andreasen (1988, p. 1387) concludes that brain imaging offers psychiatry a broad range of investigative techniques that fulfill the popular fantasy of being able to "read the mind," albeit in the form of "seeing the brain" both structurally and functionally. At present, brain imaging provides a modest amount of information that is useful in differential diagnosis, as in distinctions between depression and dementia. It has provided more information about possible pathophysiological mechanisms of major mental illnesses, including structural abnormalities in some forms of schizophrenia. Metabolic abnormalities, such as hypofrontality in schizophrenia or hyperfrontality in obsessional disorder, have also been observed. The long-term promise of brain imaging is substantial. It will permit the mapping of cerebral function in normal individuals so that we can achieve a better understanding of normal brain structure, physiology, chemistry, and functional organization. On the basis of this knowledge, the abnormalities underlying the major mental illnesses can also be mapped.

Andreasen's report gives further support to Wittenborn's (1984) hope that we are now at the threshold of important new knowledge of the relationships between neurochemical changes and behavior changes. The properties of behavior that are included in these new relationships may be different from those that form the content of current assessments.

Issues in Structured Interviewing

Young and colleagues (1987, p. 614) review common sources of interview misinformation and in so doing delineate a number of variables that must be considered in the development of structured interviewing procedures. Regarding the structure of the interview, they note the following sources of misinformation:

Structure of the interview

Lack of specificity in the question

Complex and multidimensional concepts of question

Sequence of questions

Number of questions

Question structure

Unwarranted assumptions in the questions

More than one question embedded in a single question

Sensitive or threatening element in the questions

Wording of the question (inexact terms, ambiguous or vague terms, complex terms and sentences, biased words)

Respondent sources of interview misinformation:

Need to give socially desirable answers

Lack of understanding of the questions

Memory lapses

Experience of questioning as stressful

No true opinion

Differing emotional intensity among respondents

Variable perceptions of the situation and purpose

Timing of interview

Interviewer sources of misinformation:

Interviewer characteristics

Preferences and biases

Variable emotional intensity

Variable verbal facility

Variable understanding of the questions

Recording errors

Test-retest reliability of structured interviews has been satisfactory and usually superior to traditional interview methods. However, the test-retest method, using different interviewers, faces fairly consistent methodological problems. These problems include: clinical change in the patient (e.g., new symptoms or symptom remission); recall involving efforts to give the same response or omitting symptoms mentioned earlier; therapeutic effects of the first interview; greater symptom severity at the first interview; and regression to the mean. Then there is always the question whether the diagnostic criterion that is being used is a valid yardstick.

Young and collaborators (1987) point out that, in the absence of biological markers to designate discontinuities among diagnoses, we resort to symptom grouping and enumeration as the basis for clinical diagnosis. Symptom designation involves more or less arbitrarily establishing cutting points for inclusion or exclusion from a diagnostic category. Location of the cutting point has an important effect on the percentage of correctly identified cases and non-cases. In a study sample with many severe cases, the accuracy of identification will be high, certainly higher than in a general-population sample including the full range of less severe and borderline cases.

Clinicians have been concerned that structured interviews are undergoing constant revision and that the lack of a final measure indicates an essential flaw in the instrument. A more positive view of this continuing modification is that the continuing efforts to improve the instrument attests to its importance and that it is reflecting the essential nature of the research process, that is, gradual unfolding of knowledge and facts. It must again be recognized that diagnostic systems using specific operational criteria evolve over the years, so that an ideal, static nomenclature always remains elusive. Other clinicians may feel that current structured interviews are cumbersome, or that turning pages of an interview guide may interfere with the interviewer-patient relationship, or that the interviews take too long.

Perhaps the question as to whether they are beneficial for patients in routine clinical use still has to be answered. Research has indicated that they increase the number of clinical observations (e.g., number of problem areas), and the amount of relevant patient information that is recorded by a factor of two to one. Clinicians using structured interviews tend not to be limited to the presenting symptoms in their diagnostic formulations and to have higher interrater reliability. Interviewers using structured interviews consider themselves as empathic during the interaction as when using free-flowing interviews. With practice, they are used with increasing efficiency so that there is little time difference from traditional interviews (Young et al., 1987).

There is evidence (Giannetti, 1987) that automated self-reports have advantages for both clinical practice and research. Patients accept and enjoy responding to on-line computerized questionnaires and frequently prefer them over clinical interviews or paper-and-pencil questionnaires. Even chronic and disturbed inpatients can answer computer-presented questions without assistance. There are indications that respondents are more likely honestly to report socially undesirable behavior to a computer, for example, reporting greater alcohol consumption to computers than to interviewers. Self-report and interviewer-collected history data show high agreement. Finally, it is likely to be cost saving to complete interviews by computer rather than by traditional means.

Adams and Heaton (1987) call attention to a further administrative/research role of computers in clinical practice, that is, creating and maintaining an informational database. This database might include information concerning patient demographics, referral sources, historical data, criterion test results (e.g., brain tests), psychological test findings and clinical outcome. Such information is valuable in documenting the sources of patients, their demographic and base-rate profiles, the relationship of neuropsychological tests to other results, and the impact of testing, or other services, on patient outcome. Such data are of importance in quality assurance and in evaluation research. External reviewers and 3rd-party agencies increasingly request data showing the accuracy of diagnosis and relationship to hospital/clinic utilization, more appropriate care, and improved outcome. As Adams and Heaton (1987) note, no amount of pro-

fessional insistence on freedom to practice will substitute for such data, given the current climate in health-services delivery. Once this view is accepted, it follows that the optimal way to gain control of the quality and accuracy of such data is to implement one's own system to generate the data.

Telehealth and Health-Care Informatics

Finally, we want to call attention to developments in telemedicine and medical informatics that are both very recent and in some instances well-advanced. *Behavioral telehealth* is the use of advanced telecommunications to provide access to behavioral health assessment, diagnosis, intervention, consultation, supervision, education, and information for underserved populations and isolated practitioners (DeLeon, 1997). Physicians, over the past several decades, have developed telehealth delivery models, and presently, over 40 states have ongoing "telemedicine" projects. Over 20 medical schools have established departments, graduate programs, and fellowships in "medical informatics" (Maheu, 1997).

Legislation is currently being written and passed at both federal and state levels that will add telemedicine as a normal part of health-care services. A California statute that went into effect in July 1997 stipulates that 3rd-party carriers cannot require face-to-face contact as a condition for reimbursement. Reimbursement and interstate-licensure issues need to and are being addressed. Ethics principles will have to be reviewed for application to this new form of practice and teaching models need to be developed.

Anders (1997) has described the existence of a telephone-triage system that allows shifts of nurses working on a telephone bank around the clock to talk to callers from a thousand miles away about anything from a stuffy nose to crushing chest pains and decide who should rush to an emergency room, who can safely wait for a physician's appointment, and who needs only simple home care. About 35 million Americans now have access to phone-triage lines and it is expected that this service could cover 100 million people in four years. Review of the triage recommendations revealed that about 40 percent of callers are told they don't need a physician at all. Only two percent are steered to an emergency room, around 15 percent to urgent care, and 40 percent to some sort of physician consultation.

It appears that what makes this triage service possible is that many triage questions are phrased almost identically and that the nurses scrutinize PCs for predetermined lines of inquiry. Algorithms exist for about 550 common ailments and, as callers describe their symptoms, the nurses click a few computer buttons and pull up an algorithm for those symptoms, for example, pediatric cough. The computer tells the nurses what to ask, offers new questions depending on patients' answers, and ultimately guides the nurses' advice. Advice-line executives clearly believe that a lot of medical knowledge can be codified and put into a computer. As noted by Anders (1997), the computer algorithms are designed to mimic the way that a good emergency-room physician thinks. It is of interest to note that the triage service is staffed by veteran nurses and not by low-paid clerical workers. The triage service involves an interaction among the nurse, the computer, and the patient, in which the nurse, on average, has about 8.5 minutes to build enough trust to elicit intensely personal health information and to problem-solve with the patient.

DeLeon (1997, p. 26) notes that, at a large university hospital system, clinical psychology professors and their students are analyzing how conducting various modalities of individual and group therapy via two-way interactive video-conferencing between the hospital and rural clinics affects patient and therapist satisfaction, treatment outcomes, and the development of patient-therapist relationships. When the system is not being used to directly supply services, psychology interns based in rural clinics receive supervision; rural practitioners earn continuing-education (CE) credits without having to close their practices and travel; and hospital, clinic, and university administrators conduct monthly meetings without losing hours of work to travel. Other applications include using telehealth to manage mentally ill patients in a prison-community setting and using telehealth to help severely ill patients maintain themselves in remote rural communities.

SUMMARY

For many years, the clinical interview has been used as a primary source of information about patients. Information obtained in the clinical inter-

view has been used to determine diagnosis and select treatment interventions. This has been true despite the observations made in systematic studies of the clinical interview showing that it often produced unreliable, or even misleading, data. A particular source of concern over the years has been the disagreement in diagnosis made by different clinicians interviewing the same patient.

More recently, there has been a concerted effort to address two sources of undependable information about diagnoses: development of operationally defined diagnostic decision rules (DSM-III, DSM-III-R, and DSM-IV) to reduce criterion variance, and development of structured interviews to reduce information variance (i.e., to aid in obtaining the necessary facts from the patient to be able to assign a diagnosis). Structured diagnostic interviews have been used extensively in clinical and epidemiological research. Satisfactory comprehensiveness of the information that can be collected with structured interviews is leading to their use in routine clinical practice. The availability of personal computers is leading to their increased use in patient self-administration of various structured interviews, including psychosocial history interviews and diagnostic interviews. It is possible to predict with confidence that diagnostic criteria will continue to evolve and change until laboratory correlates of diagnoses are identified. It is also possible to predict that the use of structured interviews (computer-administered), in clinical practice as well as research, will expand dramatically because of the wealth of information they provide the clinician.

REFERENCES

Adams, K. M., & Heaton, R. K. (1987). Computerized neuropsychological assessment: Issues and applications. In J. N. Butcher (Ed.), *Computerized psychological assessment* (pp. 355–365). New York: Basic Books.

American Psychiatric Association. (1980). *Diagnostic and statistical manual of mental disorders (3rd ed.).* Washington, DC: Author.

American Psychiatric Association. (1987). *Diagnostic and statistical manual of mental disorders* (3rd ed., rev. ed.). Washington, DC: Author.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

American Psychiatric Association. (1995). *The Diagnostic and statistical manual of mental disorders-Primary care version* (4th ed.). Washington, DC: Author.

Anders, G. (1997, February 4). How nurses take calls and control the care of patients from afar. *The Wall Street Journal,* (pp. 1, 6).

Andreasen, N. C. (1988). Brain imaging: Applications in Psychiatry. *Science,* 239, 1381–1388.

Ash, P. (1949). The reliability of psychiatric diagnosis. *Journal of Abnormal and Social Psychology, 44,* 272–277.

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry, 4,* 561–571.

Boon, S., & Draijer, N. (1991). Diagnosing dissociative disorders in the Netherlands: A pilot study with the structured clinical interview for DSM-III-R dissociatve disorders. *American Journal of Psychiatry, 148* (4), 458–462.

Cho, M. J., Moscicki, E. K., Narrow, W. E., Rae, D. S., Locke, B. Z., & Regier, D. A. (1993). Concordance between two measures of depression in the Hispanic health and nutrition examination survey. *Social Psychiatry & Psychiatric Epidemiology, 28,* 156–163.

Choca, J., & Morris, J. (1992). Administering the Category Test by computer: Equivalence of results. *The Clinical Neuropsychologist, 6,* 9–15.

Clayer, J. R., McFarlane, A. C., & Wright, G. (1992). Epidemiology by computer. *Social Psychiatry & Psychiatric Epidemiology, 27,* 258–262.

DeLeon, P. H. (1997). The 104th congress-Very interesting times. *The Independent Practitioner, 17,* 25–27.

Derogatis, L. R., Lipman, R. S., Rickels, K., Uhlenhuth, E. H., & Covi, L. (1973). The Hopkins Symptom Checklist (HSCL): A measure of primary symptom dimensions. In P. Pichot (Ed.), *Psychological measurement in pharmacopsychiatry* (Vol. 7). Basel, Switzerland: S. Karger.

Edelbrock, C., & Costello, A. J. (1984). Structured psychiatric interviews for children and adolescents. In G. Goldstein & M. Herson (Eds.), *Handbook of psychological assessment* (pp. 276–290). New York: Pergamon Press.

Eder, R. W., & Ferris, G. R. (1989). *The employment interview: Theory, research, and practice.* Newbury Park, CA: Sage Publications, Inc.

Endicott, J., & Spitzer, R. L. (1978). A diagnostic interview. *Archives of General Psychiatry, 35,* 837–844.

Erdman, H. P., Klein, M. H., & Greist, J. H. (1985). Direct patient computer interviewing. *Journal of Consulting and Clinical Psychology, 53,* 760–773.

Ferriter, M. (1993). Computer aided interviewing and the psychiatric social history. *Social Work & Social Sciences Review, 4,* 255–263.

Friedman, M., & Roseman, R. H. (1974). *Type A behavior and your heart.* New York: Knopf.

Giannetti, R. A. (1985). *Giannetti on-line psychosocial history:* GOLPH *(Version 2.0).* Unpublished manuscript.

Giannetti, R. A. (1987). The GOLPH psychosocial history: Response-contingent data acquisition and reporting. In J. N. Butcher (Ed.), *Computerized psychological assessment* (pp. 124–144). New York: Basic Books.

Gold, M. S. (1986). *The good news about depression.* New York: Bantam Books.

Goodman, M., Brown, J., & Deitz, P. (1992). *Managing managed care: A mental health practioner's survival guide.* Washington, DC: American Psychiatric Press.

Goodman, M., Brown, J., & Deitz, P., (1996). *Managing managed care II: A handbook for mental health professionals* (2nd ed.). Washington, DC: Amercian Psychiatric Press.

Greist, J. H. (1984). Conservative radicalism: An approach to computers in mental health. In M. D. Schwartz (Ed.), *Using computers in clinical practice: Psychotherapy and mental health applications* (pp. 191–194). New York: Haworth Press.

Hamilton, M. (1967). Development of a rating scale for primary depressive illness. *British Journal of Social and Clinical Psychology,* 6, 278–296.

Heaton, R. K., Lehman, R. A. W., & Getto, C. J. (1980). *Psychosocial pain inventory.* Odessa, FL: Psychological Assessment Resources.

Helzer, J. E., Robins, L. N., Croughan, J. L., & Welner, A. (1981). Renard diagnostic interview. *Archives of General Psychiatry, 38,* 393–398.

Hollifield, M., Katon, W., & Morojele, N. (1994). Anxiety and depression in an outpatient clinic in Lesotho, Africa. *International Journal of Psychiatry in Medicine, 24,* 179–188.

Honigfeld, G., & Klett, C. (1965). The Nurses' Observation Scale for Inpatient Evaluation (NOSIE): A new scale for measuring improvement in chronic schizophrenia. *Journal of Clinical Psychology, 21,* 65–71.

Intake Evaluation Report-Clinician Version 3.0 [Computer sortware]. (1984). Indialantic, FL: Psychologistics, Inc.

Kanfer, F. H., & Saslow, G. (1969). Behavioral diagnosis. In C. M. Franks (Ed.), *Behavior therapy: Appraisal and status* (pp. 417–444). New York: McGraw-Hill.

Kaplan, H. I., & Sadock, B. J. (1988). *Synopsis of psychiatry* (5-th ed.). Baltimore: Williams & Wilkins.

Koppel, T. (1987, June 15). *Newsweek,* pp. 50–56.

Krug, S. E. (1987). *Psychware sourcebook, 1987–1988* (2nd ed.). Champaign, IL: MetriTech, Inc.

Krug, S. E. (1993). *Psychware sourcebook, 1993* (4th ed.). Champaign, IL: MetriTech, Inc.

Lee, E., & Chan, F. (1986). The use of diagnostic interview schedule with Vietnamese refugees. *Asian American Psychological Association Journal,* 36–38.

Lukin, M. E., Dowd, E. T., Plake, B. S., & Kraft, R. G. (1985). Comparing computerized versus traditional psychological assessment. *Computers in Human Behavior, 1,* 49–58.

Maheu, M. M. (personal communication, February 9, 1997)

Maple, F. F. (1994). The development of goal-focused interactive videodiscs to enhance student learning in interpersonal practice methods classes. Special Topic: Electronic tools for social work practice and educations: II. *Computers in Human Services, 11,* 333–346.

Miller, W. R., & Marlatt, G. A. (1984). *Manual: Comprehensive Drinker Profile.* Odessa, FL: Psychological Assessment Resources.

Overall, J. E., & Gorham, D. R. (1962). The Brief Psychiatric Rating Scale. *Psychological Reports, 10,* 799–812.

Robins, L. N., Helzer, J. E., Croughan, J., & Ratcliff, K. (1981). National Institute of Mental Health Diagnositc Interview Schedule. *Archives of General Psychiatry, 38,* 381–389.

Rogers, R. (1995). *Diagnostic and structured interviewing: A handbook for psychologists.* Odessa, FL: Psychological Assessment Resources, Inc.

Rubio-Stipec, M., Canino, G. J., Shrout, P., Dulcan, M., Freeman, D., & Bravo, M. (1994). Psychometric properties of parents and children as informants in child psychiatry epidemiology with the spanish diagnostic interview schedule for children (DISC.2). *Jounal of Abnormal Child Psychology, 22,*(6), 703–720.

Siassi, I. (1984). Psychiatric interview and mental status examination. In G. Goldstein & M. Hersen (Eds.), *Handbook of psychological assessment* (pp. 259–275). New York: Pergamon Press.

Spielberger, C., Gorsuch, R., & Lushene, R. (1970). *The State-Trait Anxiety Inventory (STAI) test manual.* Palo Alto, CA: Consulting Psychologists Press.

Spiker, D. G., & Ehler, J. G. (1984). Structured psychiatric interviews for adults. In G. Goldstein & M. Hersen (Eds.), *Handbook of psychological assessment* (pp. 291–304). New York: Pergamon Press.

Spitzer, R. G., Endicott, J., & Robins, E. (1975). Clinical criteria for diagnosis and DSM-III. *American Journal of Psychiatry, 132,* 1187–1192.

Spitzer, R. L., Endicott, J., & Robins, E. (1978). Research diagnostic criteria rationale and reliability. *Archives of General Psychiatry, 35,* 773–782.

Spitzer, R. L., Forman, J. B. W., & Nee, J. (1979). DSM-III field trials: I. Initial interrater diagnostic reliability. *American Journal of Psychiatry, 136,* 815–817.

Walters, B. (1970). *How to talk with practically anybody about practically anything.* New York: Dell.

Wang, C., Liu, W. T., Zhang, M., Yu, E. S. H., Xia, Z., Fernandez, M., Lung, C., Xu, C., & Qu, G. (1992). Alcohol use, abuse, and dependency in Shanghai. In J. E. Helzer & G. J. Canino (Eds.), *Alcoholism in North America, Europe, and Asia* (pp. 264–286). New York: Oxford University Press.

Wiens, A. N. (1983). The assessment interview. In I. B. Weiner (Ed.), *Clinical methods in psychology* (2nd ed., pp. 3–57). New York: Wiley.

Wiens, A. N. (1990). Structured clinical interviews for adults. In G. Goldstein & M. Hersen (Eds.), *Handbook of psychological assessment.* New York: Pergamon Press.

Wiens, A. N., & Matarazzo, J. D. (1983). Diagnostic interviewing. In M. Hersen, A. E. Kazdin, & A. S. Bellack (Eds.), *The clinical psychology handbook* (pp. 309–328). New York: Pergamon Press.

Wing, J. K., Birley, J. L. T., & Cooper, J. E. (1967). Reliability of a procedure for measuring and classifying "present psychiatric state." *British Journal of Psychiatry, 113,* 449–515.

Wilson, L. G., & Young, D. (1988). Diagnosis of severely ill inpatients in China: A collaborative project using the structured clinical interview for DSM-III (SCID). *Journal of Nervous & Mental Disease, 176,* 585–592.

Wittenborn, J. R. (1955). *Manual: Wittenborn Psychiatric Rating Scales.* New York: Psychological Corporation.

Wittenborn, J. R. (1984). Psychological assessment and treatment. In G. Goldstein & M. Hersen (Eds.), *Handbook of psychological assessment* (pp. 405–420). New York: Pergamon Press.

World Health Organization. (1984). *WHO psychiatric disability assessment schedule (WHO/DAS).* Geneva, Switzerland. Author.

World Health Organization. (1993). *Composite international diagnostic interview.* Geneva, Switzerland. Author.

Young, G., O'Brien, J. D., Gutterman, E. M., & Cohen, P. (1987). Structured diagnostic interviews for children and adolescents. *Journal of the American Academy of Child and Adolescent Psychiatry, 26,* 611–620.

PART VII

PERSONALITY ASSESSMENT

This Page Intentionally Left Blank

CHAPTER 16

OBJECTIVE PERSONALITY ASSESSMENT

Elahe Nezami
James N. Butcher

INTRODUCTION

Historically, personality has been considered synonymous with character. Earliest historical writings document our fascination with understanding character and personality. Moralists, philosophers, writers, and politicians alike have throughout the centuries offered a myriad of ideas on the subject of personality. Early accounts place personality among the most important elements in predicting one's future fortunes.

In 50 B.C. Publilius Syrus wrote "Every man's character is the arbiter of his fortune." Of interest to Ralph Waldo Emerson was the stability of character. "A character is like an acrostic or Alexandrian stanza," wrote Emerson, "read it forward, backward, or across, it still spells the same thing." Jean De La Bruyere, 17th century moralist, spoke of the value of high character. He observed, "It is fortunate to be of high birth, but it is no less so to be of such character that people do not care to know whether you are or not."

Attempts to assess personality have a broad range of sophistication. In the late 18th and early 19th centuries, one approach to assessing personality was to feel the individual's head and examine the location of "bumps" on the skull (phrenology). From the same era there are accounts of attempts to determine character using a person's facial features (physiognomy). Graphology (examination of a person's handwriting), is another approach, first used as a means to assessing personality. This latter technique is thought by some to hold important clues to personality.

While primitive methods to assess personality have by and large given way to more empirically based techniques in the 20th century, some colorful methods still persist. Witness this approach offered by former U.S. President Ronald Reagan: "You can tell a lot about a fellow's character by his way of eating jelly beans." This method notwithstanding, in the last 75 years great progress has been made in the study of personality, giving personality assessment a strong empirical foundation.

HISTORY OF PERSONALITY ASSESSMENT

Despite the long and rich history of attempts to understand personality, formal personality assessment has a relatively short history. In 1917, Woodworth introduced the first formal self-report questionnaire. Initially designed to serve as a mass psychiatric screening for World War I draftees, the Woodworth Personal Data Sheet (PDS) was a paper-and-pencil test targeting neurotic symptoms. After the war, many inventories were constructed, each modeled after the PDS.

These early tests have been followed by continued progress in test development. The 11th edition of the *Mental Measurement Yearbook* (Kramer & Conoley, 1992) is a testimony to the proliferation of objective personality assessments in recent years. One hundred thirty-five personality tests are included in this edition, making up 28.3 percent of all tests reviewed. Remarkably, 79.3 percent of these instruments were added since the publication of the 10th edition of the yearbook in 1985. Clearly, despite the brief history of formal personality assessment, the field of psychological assessment has witnessed remarkable growth and recorded a lengthy list of breakthrough achievements that is truly impressive.

UTILITY OF PERSONALITY ASSESSMENT

Assessment plays a critical role in clinical psychology today, contributing valuable information useful in making professional diagnoses, selecting from various treatment options, and quantifying therapeutic change. Accurate initial information gathered through formal assessment is both time- and cost-effective. Aside from clinical settings, personality assessment has found favor in a wide range of other areas including business, education, and the legal system.

TEST CONSTRUCTION AND CONTEMPORARY STATUS OF OBJECTIVE PERSONALITY ASSESSMENT

Various strategies have been employed in the construction of personality tests. One strategy that has resulted in development of valid scales is the *empirical criterion keying method.* This strategy is summarized by Cohen, Swerdlik, and Smith (1992) in the following way:

1. Create a number of test items that presume to measure one or more traits.
2. Administer the proposed test items to at least two groups of people:
   a. a "criterion group" composed of people you know to possess the trait being measured, and
   b. a control group of people who are presumed not to possess the trait in question
3. Items that significantly discriminate with respect to the criterion and control groups are retained, whereas those items that do not discriminate between the two groups are discarded. (p. 416).

In this method, items are selected based on "their significance in distinguishing between groups of people differing on the criterion of interest." Two popular personality assessment instruments initially constructed using this method are the Minnesota Multiphasic Personality Inventory (MMPI) (Dahlstrom, Welsh, & Dahlstrom, 1972; Hathaway & McKinley, 1940) and its "normal" personality counterpart, the California Psychological Inventory (CPI) (Baucom, 1985; Eysenck, 1985; Gogh, 1975; Megargee, 1972). A detailed description of the MMPI will be provided later in this chapter as an example of empirical criterion keying.

A second method, referred to variously as the factor-analytic, internal-consistency, or inductive method, uses statistical strategies for test construction. Wide availability of computer technology is largely responsible for the popularity of this method. In this approach, factor analysis of the inter-item correlations is used to determine the minimum number of factors required to construct *homogeneous scales*, capturing the essence of the items included in the test. The factor-analytic method was used in constructing the Guilford-Zimmerman Temperament Survey (Guilford & Zimmerman, 1956), the Comrey Personality Scales (Comrey, 1970), and the Sixteen Personality Factor Questionnaire (16PF) (Cattell, Eber, & Tatsuoka, 1970). We have selected the 16PF for a more in-depth description of the factor-analytic method.

Raymond Bernard Cattell began by considering personality-trait names and terms available in the English dictionary and the psychiatric literature. Judges read the list and reduced it to 171 distinguishable traits (Cattell, 1957). Further ratings by college students and factor analysis reduced the number of traits to 16. The 16PF provides information along the lines of these 16 traits for "normal populations." Over the years a large body of research data has been devoted to establishing the test's usefulness (Butcher, 1985; Zuckerman, 1985).

The Clinical Analysis Questionnaire, an expansion of the 16PF, attempts to cover dimensions of pathological personality functioning in

Table 16.1. Published Research Using Personality Assessments Since 1990

	1990	1991	1992	1993	1994	1995	TOTAL
MMPI	149	204	279	282	277		1191
MCMI	25	30	37	21	46		159
16PF	22	26	45	45	65		203
BPI	5	2	2	2	3		14
PAI	3	9	8	5	4		29
NEO-P	33	34	21	28	13		129
DPI	0	0	0	0	0		0
PRF	13	9	2	10	4		38
RORSCHACH	68	124	134	129	165		560
TAT	19	34	35	39	66		193

Note: Documented from PsychINFO.

addition to normal functioning. This test measures personality along the lines of 16 normal traits and 12 clinical dimensions (Krug, 1980). Therefore, it is suitable for both normal and clinical populations. The standard form consists of 187 items from the original 16PF and an additional 144 items to assess psychopathology. There is also a short form, consisting of 128 items from the 16PF and the 144 clinical items. Computerized scoring and interpretation are offered for this assessment. The adequacy of the additional 144 items for detecting psychopathology awaits further investigation. One of the problems associated with this personality test is the lack of adequate validity indicators. Winder, O'Dell, and Karson (1975) developed a Faking Bad scale for the original form. However, the research on the utility of this scale is limited. Therefore, caution should be exercised in its use with cases where faking might be a concern (Berry, Wetter, & Baer, 1995).

A third approach to test construction uses the logical or *rational method* of item selection. The Woodworth Psychoneurotic Inventory (Woodworth, 1917) employed the rational method of test construction in selection of recruits during World War I. The rational approach to item selection was also utilized in the construction of the Mooney Problem Checklist (Mooney and Gordon, 1950). This test has several forms for junior high school, high school, and college students, as well as adult populations. The Mooney Problem Checklist identifies problems for discussion in individual or group counseling. While both measures show promise as important measures of personality, at this time there is only limited research substantiating their use.

We conclude our discussion of personality test construction with a brief overview of general areas of current use. It should be noted that, due to space limitations, we have not attempted to provide an overview of all of the tests mentioned. The reader is referred to other available sources, such as the 11th Mental Measurement Yearbook (Kramer and Conoley, 1992) and Anastasi (1988) for a description of these personality assessments. In our initial survey we have focused on articles using one or more of the following tests: MMPI, 16PF, 16PF Clinical Questionnaire, NEO-PI, MCMI, BPI, DPI, PRF, PAI[1]. In keeping with the approach of the authors of the prior edition of this volume, we have chosen these instruments because they are the most widely used objective measures, they are designed to cover a wide variety of personality dimensions, and they represent a diversity of major methods of test construction.

Survey of Recent Literature

Objective personality assessment has generated numerous scientific publications in our recent literature. The number of studies using a wide range of objective personality assessments, followed by the leading projective personality assessments since 1990, is shown in Table 16.1. The MMPI is the most widely researched instrument for personality assessment. Sundberg (1961) documented the MMPI's unchallenged lead in the field of objective personality assessment in the late 1950s. As indicated in Table 16.1, the MMPI has successfully kept its lead in the field of objective personality

Table 16.2. Summary of Correlates of the MMPI Validity, Clinical, and Selected Special Scales

MMPI SCALE	SCALE CORRELATES
Validity Scales	
? (Cannot Say)	The number of items not answered or answered in both directions. A defensive or invalid profile with possible attenuation of scale scores is suggested if the ? raw score is 30 or more.
L (Lie)	Measures a rather unsophisticated or self-consciously "virtuous" test-taking attitude. Elevated scores suggest that the individual is presenting himself or herself in an overly positive light, attempting to create an unrealistically favorable view of his or her adjustment.
F (Infrequency)	Items on this scale are answered very infrequently by most people. A high score suggests not only an exaggerated pattern of symptom checking that is inconsistent with accurate self-appraisal, but also confusion, disorganization, or actual faking of mental illness.
K (Defensiveness)	High scores reflect an uncooperative attitude and an unwillingness or reluctance to disclose personal information or problems. Low scores suggest openness and frankness. K is positively correlated with intelligence and educational level, which should be taken into account in interpretation.
Clinical Scales	
1 (Hs, Hypochondriasis)	High scorers present numerous vague physical problems that tend to be chronic. They are generally unhappy, self-centered, complaining, hostile, and attention-seeking in their behavior.
2 (D, Depression)	Reflects depressed mood, low self-esteem, and feelings of inadequcy, High scorers are described as moody, dependent, pessimistic, distressed, high-strung, lethargic, over-controlled, and guilt-prone. Elevations may reflect great discomfort and need for change or symptomatic relief.
3 (Hy, Hysteria)	High scorers tend to rely on neurotic defenses such as denial and repression to deal with stress. They tend to be dependent, naive, outgoing, infantile, and narcissistic. Their interpersonal relations are often disrupted, and they show little insight into problems. High scorers show little interest in psychological processes and interpret psycholocial problems as physical ones. High levels of stress are often accompanied by the development of physical symptoms.
4 (Pd, Psychopathic, Deviate)	Measures antisocal behavior, such as rebelliousness, disrupted family relations, impulsiveness, difficulties with school or work, legal involvement, and alcohol or drug abuse. Personality disorders are likely among high scorers: they are outgoing, sociable, and likeable as well as deceptive, manipulative, hedonistic, exhibitionistic, inclined toward poor judgment, unreliable, immature, hostile, and aggressive. High scores usually reflect long-standing character problems that are highly resistant to change.
5 (Mf, Masculinity-Femininity)	High-scoring men are described as sensitive, aesthetic, passive, or feminine. They may show conflicts over sexual identity and low heterosexual drive. Low-scoring men are viewed as masculine, aggressive, crude, adventurous, reckless, practical, and having narrow interests. Because the direction of scoring is reversed, high-scoring women are seen as masculine, rough, aggressive, self-confident, unemotional, and insensitive. Low-scoring women are viewed as passive, yielding, complaining, fault-finding, idealistic, and sensitive.
6 (Pa, Paranoia)	Elevations on this scale are often associated with being suspicious, aloof, shrewd, guarded, worrying, and overly sensitive. High scorers may project or externalize blame and harbor grudges against others. They are generally hostile and argumentative.

(*continued*)

Table 16.2. (Continued)

MMPI SCALE	SCALE CORRELATES
7 (Pt, Psychasthenia)	High scorers are tense, anxious, ruminative, obsessional, phobic, and rigid. They frequently are self-condemning and guilt prone, feel inferior and inadequate, over-intellectualize and rationalize problems, and resist psychological interpretations.
8 (Sc, Schizophrenia)	High scorers have an unconventional, schizoid lifestyle. They are withdrawn, shy, and moody, and feel inadequate, tense, and confused. They may have unusual or strange thoughts, poor judgment, and erratic moods. Very high scorers may evince poor reality contact, bizarre sensory experiences, delusions, and hallucinations. They are generally uninformed and have poor problem-solving skills.
9 (Ma, Hypomania)	High scorers are viewed as sociable, outgoing, implusive, overly energetic, and optimistic. They have liberal moral views, are flighty, may drink excessively, are grandiose, imitable, impatient, and rarely "follow through" with their plans. They are manipulative and exaggerate their self-worth. Very high scorers may show affective disorder, bizarre behavior, erratic mood, impulsive behavior, and delusions.
O (Si, Social Introversion)	High scorers are viewed as introverted, shy, withdrawn, socially reserved, submissive, overcontrolled, lethargic, conventional, tense, inflexible, and gult-prone. Low scorers are extroverted, outgoing, gregarious, expressive, aggressive, talkative, impulsive, uninhibited, spontaneous, manipulative, opportunistic, and insincere in social relations.
Special Scales	
A (Anxiety)	This scale defines the first and the largest factor dimension in the MMPI. It measures general maladjustment or emotional upset. High scores reflect anxiety, tension, lack of efficiency, and open expression of numerous psychological complaints.
R (Regression)	This factor scale relates to reliance on denial and repression. High scores reflect uninsightful, overcontrolled, and inhibited behavior. These individuals tend to avoid problems and appear overly conventional.
Es (Ego Strength)	This scale was originally developed to predict successful response to psychotherapy. Subsequent research has shown it to be an indicator of a person's overall level of functioning. High scores reflect effective functioning and the ability to withstand stress. Such individuals tend to have psychological resources that will help them to cope with problems.
MAC (MacAndrew Addiction)	This scale was developed to distinguish alcoholic from nonalcoholic psychiatric patients. High scores are also associated with other addictive problems such as drug abuse and pathological gambling; MAC serves as a measure of addiction-proneness.

Note: Adapted from University of Minnesota Press (1984). *Users Guide for the Minnesota Report: Personnel Selection System.* Minneapolis, MN: Author.

assessment. Accordingly, we will devote the rest of the chapter to its use.

HISTORY OF THE MMPI

Over 55 years ago, Hathaway and McKinley used an empirical strategy to develop what has evolved into the most widely used and respected objective personality-assessment instrument around the world—the MMPI. They developed item clusters with maximum discriminant validity by comparing psychiatric patients with "normal" persons. These item clusters or scales were effective in describing behavior and personality as well as classifying psychopathology (see Dahlstrom & Dahlstrom, 1980, for a collection of historically important articles on the MMPI).

Initially, Hathaway and McKinley developed a pool of 1,000 items, covering contents and themes from a wide range of psychiatric diag-

noses. After eliminating redundant and unnecessary items, a reduced version consisting of 504 items was administered to a group of individuals referred to as "Minnesota normals." This group was made up largely of in-patients at the University of Minnesota Hospital, a group of students, some medical patients, and a group of Work Progress Administration workers. The clinical scales were developed by determining which items differentiated the "normal" group from various clinical reference groups. The clinical reference or the criterion group consisted of eight groups of inpatients from the University of Minnesota Hospital. Items that successfully differentiated the "normal" group from each criterion group were chosen for the scales. Consequently, eight clinical scales corresponding to the eight diagnostic categories were constructed: Hs (Hypochondrinsis), D (Depression), Hy (Hysteria), Pd (Psychopathic Deviate), Pa (Paranoia), Pt (Psychasthenia), Sc (Schizophrenia), and Ma (Hypomania).

Hathaway and McKinley included several validity scales in the original MMPI. (See Table 16.2.) In the MMPI, the "cannot say" scale is simply the total number of items that were not answered or were answered in both directions; a very high "cannot say" score lowers the scores on all scales and calls the validity of the test into question. The L scale (for Lie) was a rationally constructed scale composed of items designed to tap an individual's unwillingness to admit to commonly acknowledged minor faults. The F scale measures deviance of responses compared to the normative sample; it is composed of items which less than 10 percent of the "Minnesota normals" answered in the keyed direction. Finally, the K scale was designed to identify subtle clinical defensiveness. It is the only validity scale in the original MMPI constructed empirically by the method of contrasted groups. A group of people with known psychopathology but normal MMPI profiles was compared with a group of non-patients with normal profiles. The resulting K scale was later used to develop correction factors for defensiveness for several of the clinical scales (McKinley, Hathaway, & Meehl, 1948).

Two additional scales, Mf (Masculinity/Femininity) and Si (Social Introversion) were added to the basic MMPI profile later. Unlike the other clinical scales, Mf and Si were constructed using non-patient criterion groups. Items from these scales and inclusion of 16 repeated items brought the total number of items to 566. A more detailed description of the MMPI's construction and validation is presented by Welsh and Dahlstrom (1956).

Revision and Restandardization of the MMPI

In their overview of developments in the use of the MMPI, Butcher and Owen (1978) noted a number of problems and criticisms of the MMPI. Some of these problems involved the MMPI itself, specifically the need for a revision and restandardization of the inventory. Other problems concerned the relative inactivity on the part of the test distributor in keeping up with existing MMPI technology and in failing to provide necessary interpretive materials that had been developed for the MMPI. The problems cited included the need to update and broaden the MMPI item pool; the need for a new standardization of normal responses on a broader, more representative contemporary normative sample; and the need for more flexibility and willingness on the part of the test distributor to provide relevant test materials.

In 1982 the University of Minnesota Press initiated a major research effort to revise, update, and restandardize the MMPI (Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989). Some of the goals set for the redevelopment of the MMPI were the following:

1. Maintain the integrity of the existing validity, clinical, and widely used special scales of the test.
2. Revise and reword the language of some of the existing items that are out of date, awkward, sexist, or otherwise inappropriate. Language use has changed over the years, and the content of some of the MMPI items has become antiquated.
3. Broaden the item pool to include other contents not represented—for example, treatment compliance, amenability to change, relationship problems, work attitudes, and so on.
4. Develop new, up-to-date norms for the MMPI. The original "Minnesota normals" represented a small, regional, somewhat parochial sample of adults living 40 years ago. Regional norma-

tive samples collected in North Carolina by Diehl (1977) and in southern Minnesota by Colligan, Osborne, Swenson, and Offord, (1984) indicated that the old norms are not representative of response patterns of contemporary "normal" people, but no nationwide standardization has ever been undertaken.

5. Include separate forms of the MMPI for adults and adolescents. New items were included for the adolescent form of the MMPI that are specific to problems of adolescents.

The MMPI project committee (James Butcher of the University of Minnesota, Grant Dahlstrom of the University of North Carolina, Jack Graham of Kent State University, and Auke Tellegen of the University of Minnesota) chose to conduct a program of research that was both conservative and expansive in scope: the goal was to maintain the integrity of the original instrument while expanding its range of coverage, utility, and acceptability to clients. To insure continuity with the original MMPI and its extensive research base, the committee decided to include the entire existing MMPI item pool (550 items) in the experimental booklet so that the original items and scales could be studied in modern samples of normals. It was possible for users involved in the restandardization research to continue to score and interpret the original MMPI scales while collecting responses on the new instrument. The 16 repeated items in the original MMPI were deleted and replaced with new items described below. This change would not affect scoring of the basic MMPI scales, as the repetitions were not scored in the original MMPI. About 14 percent of the original items were changed because of dated language or content, sexist or otherwise objectionable wording, or awkward grammar. In a few instances, where items were so out of date as to be meaningless, new items were substituted. However, for comparison purposes, the original items were retained (along with the rewritten version) in the experimental booklets. Analyses presented by Ben-Porath and Butcher (1988) revealed that, out of 82 rewritten items, only nine showed significant differences in endorsement percentage when compared to their original version, and none of these differences held across both sexes. Rewriting the items did not change any item-scale correlations significantly, and so had no real effect on the psychometric characteristics of the MMPI.

In addition to retaining the original items, some in updated form, the experimental MMPI booklet was expanded by the addition of 154 new items designed to address problem areas not well represented in the original version of the test. These additional items were selected rationally through a broad sampling of views of MMPI experts as to which content domains needed further coverage. The separate adolescent form of the experimental MMPI booklet also contained the original 550 MMPI items, 50 new items dealing with treatment amenability, and 104 new items designed to provide better coverage of concerns and problems specific to adolescents.

The official MMPI-2 restandardization project involved an extensive collection of the updated MMPI, biographical, demographic, and life-event data on a national normative sample of adults (N=2,600) and adolescents (N=805 boys and 815 girls). The new subject population for the MMPI-2 was obtained through sampling normal volunteers from several regions of the United States: Minnesota, Ohio, North Carolina, Pennsylvania, Washington, Virginia, and California. Efforts were made to obtain a sample representative of the U.S. population by matching sample characteristics to major demographic characteristics reported in the 1980 census. For a large subsample of the adult population, both members of married couples were tested and asked to fill out behavioral and personality ratings of each other as well as assessing their marital relationship. Extra forms used in the national standardization provided descriptive data and MMPI response correlates that were not available for the original Minnesota normative sample.

Users will find many of the new inclusions in the MMPI-2 are as useful or even more useful than the original MMPI scales. For example, content interpretation of the MMPI has been expanded and improved by the development of several new MMPI-2 content scales that include the dimensions represented by new items in the MMPI-2 item pool (See Table 16.3: Butcher, Graham, Williams, & Ben-Porath, 1990). These scales were developed by a multistage, multi method approach starting with rational item group-

Table 16.3. MMPI-2 Content Scales

SCALE	DESCRIPTION OF CONTENT AND CORRELATES
ANX (Anxiety)	General symptoms of anxiety and tension, sleep and concentration problem, somatic correlates of anxiety, excessive worrying, difficulty making decisions, and a willingness to admit to these problems.
FRS (Fears)	Many specific fears and phobias including animals, high places, insects, blood, fire, storms, water, the dark, being indoors, dirt, and so on.
OBS (Obsessiveness)	Excessive rumination, difficulty making decisions, compulsive behaviors, rigidity, feelings of being overwhelmed.
DEP (Depression)	Depressive thoughts, anhedonia, feelings of hopelessness and uncertainty, possible suicidal thoughts.
HEA (Health Concerns)	Many physical symptoms across several body systems: gastointestinal, neurological, sensory, cardiovascular, dermatological, and respiratory. Reports of pain and of general worries about health.
BIZ (Bizarre Mentation)	Psychotic thought processes, auditory, visual, or olfactory hallucinations, paranoid ideation, delusions.
ANG (Anger)	Anger-control problems, irritability, impatience, loss of control, past or potential abusiveness.
CYN (Cynicism)	Misanthropic beliefs, negative expectations about the motives of others, generalized distrust.
ASP (Antisocial Practices)	Cynical attitudes, problem behaviors, trouble with the law, stealing, belief in getting around rules and laws for personal gain.
TPA (Type A)	Hard-driving, work-oriented behavior; impatience; irritability; annoyance; feelings of time pressure; interpersonally overbearing.
LSE (Low Self Esteem)	Low self-worth; overwhelming feelings of being unlikable, unimportant, unattractive, useless, and so on.
SOD (Social Discomfort)	Uneasiness around other, shyness, preference for being alone.
FAM (Family Problems)	Family discord, possible abuse in childhood, lack of love and affection or marriage, feelings of hate for family members.
WRK (Work Interference)	Behaviors or attitudes likely to interfere with work performance, such as low self-esteem, obsessiveness, tension, poor decision making, lack of family support, negative attitudes towards career or coworkers.
TRT (Negative Treatment Indicators)	Negative attitudes toward doctors and mental health treatment. Preference for giving up rather than attempting change. Discomfort discussing any personal concerns.

Note: From Butcher, Graham, Williams, and Ben-Porath, 1990.

ings and then proceeding through statistical item-selection techniques to improve individual scale homogeneity and reduce scale intercorrelations. Several studies aimed at testing and validating these scales against contemporary clinical populations have been conducted (Graham & Butcher, 1988; Keller & Butcher, 1991).

MMPI-2/MMPI-A

The MMPI-2 consists of 567 items and the MMPI-A, the adolescence version, contains 467. Both share strong research and clinical foundations with their predecessor. The MMPI/MMPI-2 and MMPI-A are recognized as the most widely used and researched personality tests available

(Lubin, Larsen, Matarazzo, & Seever, 1985; Lubin, Larsen, & Matarazzo, 1984; Watkins, Campbell, & McGregor, 1988; Piotrowski & Keller, 1989; Butcher & Rouse, 1995). They are also cited as the most frequently administered psychological tests in both inpatient and outpatient therapy (Cohen, Swerdlik, & Smith, 1992), providing "a global and comprehensive measure of personality functioning in addition to specific diagnostic information" (Ben-Porath & Waller, 1992). The MMPI-2 provides a wealth of information from the interpretation of six validity scales, 10 clinical scales, 15 content scales, 18 supplementary scales, and the Harris-Lingoes subscales.

Administration and Scoring

The MMPI-2 and MMPI-A are relatively easy to administer; it takes only about 90 minutes to complete the MMPI-2 and 60 minutes for the MMPI-A. A 5th- or 6th-grade level of reading is sufficient for comprehension of the items (Butcher, 1995a). In addition to the paper and pencil forms, audio-cassette or computer-automated administration is also available. Like most objective assessments, administration of the MMPI-2 and MMPI-A require standard procedures matching those employed in collecting normative data. If the subject is not able to complete the full test, an abbreviated version of MMPI-2/MMPI-A can be administered. The abbreviated form generates information on the standard validity and clinical indicators included in the first 370 items in the MMPI-2 booklet or the first 350 items in the MMPI-A booklet. Short forms (that is, forms that estimate scale scores from a reduced-item set) of the instrument are not appropriate alternatives to the standard MMPI (Butcher, Kendall, & Hoffman, 1980; Dahlstrom, 1980) and are not recommended for clinical use (Greene, 1991; Hart, McNeill, Lutz, & Adkins, 1986; Graham, 1993).

Hand-scoring templates are available for manual scoring of the MMPI-2 and MMPI-A. The obtained raw scores are used then for plotting individual profiles. For the MMPI-2 five of these raw scores (Hs, Pd, Pt, Sc, Ma) are corrected for defensiveness before the profile is drawn. Computerized scoring and interpretations are also available and will be discussed latter in this chapter.

Interpretation

Personality Functioning

MMPI-2 data are useful in providing information in several areas: attitudes toward assessment, cooperation, cognitive/ideation, mood and affect, conflict areas, coping styles, diagnostic considerations, and treatment recommendations.

1. Validity Indicators. The first step in interpreting any personality test profile is to establish its validity to assure the subject's cooperativeness in taking the test. The MMPI/MMPI-2 continues to have the most comprehensive validity indicators among all existing instruments of personality assessment (Bagby, Buis, & Nicholson, 1995; Berry, Baer, & Harris, 1991a; Schretlen, Wilkins, Van-Gorp, & Bobholz, 1992). The original validity indicators of the MMPI included "cannot say," L, F, and K. In addition, two new response-inconsistency indicators were introduced with the MMPI-2. Variable Response Inconsistency (VRIN) measures the subject's consistency in responding to the content of the items. This index is most effective in detecting random responding (Berry, Wetter, Baer, Larsen, Clark, & Monroe, 1992). An elevated VRIN score in combination with a high F is indicative of random responding or confusion. True Response Inconsistency (TRIN) detects indiscriminate "yea-saying" or "nay-saying" response patterns. Additional validity indicators include Back-Page Infrequency, or F(B) scale (Berry, Wetter, Baer, Widiger, Sumpter, Reynolds, & Hallam, 1991b) and Inpatient Psychopathology, or F(p) (Arbisi & Ben-Porath, 1995). F(B) consists of 40 infrequently endorsed items appearing in the latter part of the booklet and is included in order to assess the validity of the last 197 items. A new scale (S), measures the tendency to present oneself in a superlative manner. Focusing on symptom underreporting, the S scale is a newly published addition to the validity indicators of the MMPI-2 (Butcher & Han, 1995). The utility of the MMPI validity scales has been thoroughly researched and continues to generate much research in the literature (see Baer, Wetter, & Berry, 1992; Berry et al., 1991b; Graham, Watts, & Timbrook, 1991; Wetter, Baer, Berry, Smith, & Larsen, 1992; Pope, Butcher, & Seelen, 1993). The MMPI-2's validity scales are generally

thought to provide a comprehensive assessment of the subject's attitudes and cooperation in responding to the MMPI items (Graham, et al., 1991).

2. Configural Interpretation (Code type). Because of the intercorrelations among scales as well as the overlap among characteristics of clinical syndromes, it was often found that several MMPI scales tended to be elevated together. Thus, interpretation of a particular scale might vary depending on the relative elevations of other scales in the profile. For this reason it made sense to consider two or more clinical scales together in interpreting a profile. Referred to as the configural approach, a wealth of information has been gained from the original MMPI using configurations, or code types. The continuity of the MMPI-2 and MMPI, and minimal changes in the clinical scales are expected to produce similar code types for the two versions of the test (Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989; Butcher, Graham, & Ben-Porath, 1995; Graham, Timbrook, Ben-Porath, & Butcher, 1991).

The consistency of code types from one administration to a second administration of the same version, or from the original to the revised version, is dependent both on the quality of how code types are defined and what one expects to gain from them. Profile definition is determined as follows.

The quality of a code type is determined by subtracting the highest scale not included in the code type from the lowest clinical scale in the code type or by subtracting the lowest scale in the code type from the next highest clinical scale not included in the code type. Due to measurement error, a minimum T-score difference of five is needed to identify a meaningful code type, useful in interpretation (Butcher, Graham, & Ben-Porath, 1995). A poorly defined code type, for example, one with only a one- or two-point T-score difference, may not be as reliable as one that is more clearly defined. Even when repeated administrations of the original MMPI were considered, such code types were determined to be less stable and, therefore, of limited clinical utility (Graham, Smith, & Schwartz, 1986). Graham and colleagues (1986) reported that the level of agreement at retest for individuals with poorly defined two-point code types was modest at best. However, acceptable agreement was reached when well-defined code types were considered.

There is evidence that "congruence between MMPI and MMPI-2 code types is greater for well-defined code types than for poorly defined code types (Graham, Timbrook, Ben-Porath, & Butcher, 1991). Tellegen and Ben-Porath (1993) suggested that well-defined code types reduce assessment error by identifying more homogeneous and distinct groups. They asserted that "generally high congruences found for well-defined code types provide conceptually, empirically, and practically meaningful links between the MMPI-2 and the MMPI" (p. 498). Harrell, Honaker, and Parnell (1992) examined the congruency of MMPI and MMPI-2 profiles among psychiatric outpatients and found that, consistent with previous research, interpretation of well-defined code types is recommended for both the MMPI and MMPI-2. These investigators suggested that use of well-defined profiles is necessary for reliable clinical interpretation.

Continuity of code types between the MMPI-2 and its predecessor has been a topic of recent controversy. Dahlstrom (1992), and Humphrey and Dahlstrom, (1995) questioned the concept of scale definition across the two versions. Tellegen and Ben-Porath (1993), and Ben-Porath and Tellegen (1995) provided insight into the controversy and explained the possible code type discrepancies previously observed. They noted that code type discrepancies described by Dahlstrom generally were derived from samples within the normal range, thereby lacking well defined code types. They were able to document, as expected, that possible shifts in profile code types are limited to profiles that are not clearly defined. Tellegen and Ben-Porath (1993) were also able to show that concerns about different code-type configurations and interpretations are unwarranted with clearly defined code types. In sum, the most recent findings are in accordance with previous studies that level of code-type definition is directly related to the level of agreement between observed MMPI and MMPI-2 code-type configurations (Graham & Ben-Porath, 1995). When we consider agreement between the two forms in the context of test-retest data and code-type purity, we conclude that the MMPI-2 appears to resemble the MMPI very closely in terms of clinical scale scores. This debate strengthens previous statements by Graham, Timbrook, Ben-Porath, and Butcher, (1991) that con-

gruence between the two versions (MMPI and MMPI-2), and consequently reliability of interpretations, increases as the code types are more clearly defined.

Empirical research in different settings has yielded new information suggesting that MMPI research can be applied to interpret the MMPI-2 code types. Two such recent studies were conducted to examine the external correlates of MMPI-2 code types with outpatient clients and alcoholics. Graham and Ben-Porath (1995) validated MMPI-2 two-point code types of 1,020 clients from outpatient mental-health centers. The criterion information was available from SCL90-R and mental-status ratings based on clinical interviews. They found that MMPI-2 interpretive information based on clearly defined code types was consistent with extra-test correlates found in the literature on the original MMPI. Babcock (1995) compared the MMPI-2 profiles, including the two-point code types, of 93 alcoholics with previously reported MMPI profiles of alcoholics. Babcock demonstrated that MMPI-2 code types were consistent with previous MMPI research utilizing the same patient population. Therefore, he concluded that MMPI research is "directly applicable to the MMPI-2."

The comparability of the MMPI-2's and MMPI's application for different clinical diagnoses and with different ethnic populations has also generated research interest. Litz and colleagues (1991) examined the comparability of MMPI-2 and MMPI interpretations in differential diagnosis of posttraumatic stress disorder. They found good concordance between the basic scales and two-point code types using the MMPI and MMPI-2. They concluded that validity and clinical-scale patterns for post-traumatic stress disorder (PTSD) patients appear to be similar for the MMPI and MMPI-2. Whitworth and McBlaine (1993) examined the stability of MMPI and MMPI-2 results among white and Hispanic-Americans. One hundred ten white and 173 Hispanic Americans participated in this study. Using the MMPI-2 the researchers reported differences on four scales of L, K, Hy, and Pd among white and Hispanic Americans, consistent with previously reported differences using the original MMPI. Thus, they concluded that interpretations based on the MMPI-2 are congruent with interpretations based on the MMPI. Therefore, the available literature on the MMPI with Hispanics applies to the MMPI-2 as well.

In summary, comparability of the MMPI-2 and MMPI test results is demonstrated when clearly defined code types are considered (Graham, Timbrook, et al., 1991; Ben-Porath, Slutske, & Butcher,. 1989; Chojnacki & Walsh, 1992; Harrell, Honaker, & Parnell, 1992). The present state of the research confirms Vincent's (1990) conclusion stating that one can be reasonably confident that compatibility of the MMPI-2 with the original MMPI is as good as the compatibility of the original MMPI to itself. It is important to note that profiles with less clearly defined code types are best interpreted using a scale-by-scale interpretation instead of forcing such profiles into a less reliable code type and making configural interpretations of dubious value.

3. Content Based Interpretation. Historically, clinical interpretation of the MMPI using specific content of items was generally discouraged until the publication of the Wiggins scales (1969), and then this method received only minor attention. However, in recent years the content of endorsed items has gained increased prominence, adding valuable information to the clinical picture provided by the standard clinical scales alone. The basic premise of content interpretation rests on the assumption that the subject, in answering test items, is reacting and responding openly and directly to the MMPI. Consequently, the content of MMPI items might represent an important source of information not available through empirical test-interpretation procedures. In the following section we provide a general overview of some of the content-based interpretations available for the MMPI, preserved in the MMPI-2. This section is followed by a description of the new MMPI-2 content scales. Finally, a brief discussion of special scales is presented.

MMPI and MMPI-2 Content Interpretation

Several historical strategies for interpreting content for the MMPI should be noted. Welsh and Dahlstrom (1956) and Block (1965) used the factor-analytic approach to evaluate the underlying factors of the inventory. Two homogeneous content dimensions were distilled: anxiety (A) and repression (R). These two factors have been explored extensively by researchers and clinicians.

In another approach to item content, Harris and Lingoes (1968) used a rational strategy to develop subscales for several MMPI clinical scales. These subscales are provided for D, Hy, Pd, Pa, Sc, and Ma. Content-consistent subscales for Mf and Si were later developed by Serkownek (see Schuerger, Foerstner, Serkownek, and Ritz, 1987) based on factor analyses of the items in these two scales.

A third approach to item content involved the use of critical items to assess relevant content themes provided by the patient. In this strategy certain items are believed to hold a special significance in interpretation. Grayson (1951) is credited with the first attempt at using "critical items" to interpret the MMPI. However, no empirical validation for the items he identified are available. More than twenty years later another series of critical items was suggested (Koss & Butcher, 1973; Koss, Butcher, & Hoffman, 1976; Lachar & Wrobel, 1979). These critical items are empirically based and widely used in interpretation. However, it should be noted that the low reliability of single items limits this interpretive strategy (Koss, 1980).

New MMPI-2 Content scales

The MMPI-2 content scales are a set of new scales. In some cases, MMPI-2 content scales are updated versions of previously available constructs (e.g., depression). In other instances these scales are new constructs not previously identified on the MMPI (e.g., Type A behavior pattern, treatment amenability, and work problems). The MMPI-2 content scales reflect 15 content dimensions in the MMPI-2 item pool. (See Table 16.3.) These scales enjoy good internal consistency and are relatively independent of each other.

Several recent studies have explored the external validity of the MMPI content scales. Ben-Porath, Butcher, and Graham (1991) demonstrated incremental validity of two content scales, depression and bizarre mentation, in differential diagnosis of depression and schizophrenia, respectively. They documented the effectiveness of these two content scales in providing relevant diagnostic information not available through the sole examination of the standard clinical scales. In agreement with previous studies, another recent study documented the validity of the MMPI-2 content scales (Ben-Porath, McCully, & Almagor, 1993).

Interpretation of Special Scales

In this section, we discuss several special scales that have secured widespread popularity in both clinical and research settings. Among the most common experimental scales are the Ego Strength Scale (Es), a measure of tolerance for stress or ego strength (Barron, 1953); the Welsh Anxiety Scale (A), a factor scale designed as a measure of overcontrol; and the MacAndrew Alcoholism Scale (Mac), a measure of an individual's proneness to addiction (MacAndrew, 1965).

Several new supplementary scales are included in the MMPI-2. The Marital Distress Scale (MDS), the Addiction Acknowledgment Scale (AAS), and the Addiction Potential Scale (APS) are among these scales. Preliminary validation of the MDS has been reported by Hjemboe, Almagor, and Butcher (1992). Validity of the AAS and the APS has also been documented (Weed, Butcher, McKenna, & Ben-Porath, 1992; Greene, Weed, Butcher, Arredondo, & Davis, 1992).

In sum, preliminary results suggest the ability of new content and supplementary scales to make significant contributions in clinical interpretation of the MMPI-2. Table 16.2 provides a brief description of the basic validity and clinical scales for the MMPI. More detailed description of the many MMPI scales and scale combinations can be found in a number of useful texts (cf. Dahlstrom, Welsh, & Dahlstrom, 1972, 1975; Greene, 1991; Graham, 1993; Butcher & Williams, 1992).

Utility of the MMPI in Different Contexts

The MMPI has proven to be a valuable assessment instrument in inpatient psychiatric facilities, outpatient psychotherapy clinics, and counseling centers, assisting clinicians in the areas of differential diagnosis, treatment planning, and evaluation of treatment outcome. In addition, the MMPI has proved to be a valuable assessment tool in college counseling, contributing to development of rapport and appropriate counseling goals (Butcher & Graham, 1994). The ever-increasing need for treatment accountability

has made the MMPI the most widely used instrument in the emerging assessment-treatment feedback model now being incorporated in health maintenance organizations for treatment planning (Butcher, 1990; Erdberg, 1979; Finn & Tonsager, 1992; Quirk, Strosahl, Kreilkamp, & Erdberg, 1995). In addition, the MMPI has found a special place in medical hospitals. In these settings it has been used to screen for psychopathology (Keller & Butcher, 1991) and substance abuse (Butcher & Graham, 1994). In addition, the MMPI has been used to identify personality characteristics that might predispose people to a variety of medical problems. For example, one particular area of research focuses on the relation of the hostility-scale (Cook & Medley, 1954) scores to rates of cardiovascular mortality and morbidity (e.g., Barefoot, Dahlstrom, & Williams, 1983; Han, Weed, Calhoun, & Butcher, 1995).

In the court system, the MMPI is recognized as the most prominent forensic psychological assessment tool. It is used in court cases to evaluate for possible insanity, determine competency to stand trial, classify offenders, and conduct child-custody evaluations.

Finally, as one of the most frequently used objective personality assessments, the MMPI has a long history of use in personnel selection (Butcher, 1991, 1995a). It is used as a screening tool for preemployment evaluation for a variety of sensitive or stressful occupations. Such evaluations are desirable for occupations in which the emotional stability of employees is especially crucial (e.g., airline pilots: Butcher, 1994; and nuclear power plant employees: Dyer, Sajwaj, & Ford, 1993). The utility of the MMPI in the military has been well-documented. Butcher, Jeffrey, Cayton, Colligan, Devore, and Minegawa (1990) summarized the MMPI's long history and utility for screening purposes, selection of personnel for special duties, and evaluation of emotional consequences of enduring harsh environmental conditions.

In sum, recent research has explored use of the MMPI in a myriad of environments to perform a wide range of functions beyond those for which it was originally developed. Research in these areas has expanded our knowledge and invites further interest to extend our understanding of issues relevant to these different areas.

COMPUTERIZED OBJECTIVE PERSONALITY ASSESSMENT

The use of computers to interpret psychological tests began at the Mayo Clinic more than 35 years ago (Rome, Swenson, Mataya, McCarthy, Pearson, Keating, & Hathaway, 1962). Since that time, this approach has found increasing acceptability in the mental health profession. Today computers are utilized in all stages of assessment, including administration, scoring, research, and clinical interpretation (Butcher, 1995). Reviews and various automated interpretive systems can be found in Butcher (1987, 1994, 1995). Computers decrease the amount of time required for interpretation of a personality assessment. However, the need for understanding assessment and knowledge about test construction and validation is not eliminated through the use of computers. Rather, with expansion of computer-generated programs for use in test scoring and interpretation, there is an ever-increasing need to document appropriate test construction and validation (Butcher, 1995a) before any computer-generated psychological assessment can be used.

Automated Test Administration

Three features of objective personality assessments make them appealing for automated administration. First, the test-stimulus material is highly standardized. Second, there are limited structured-response alternatives (e.g., true or false options). Following similar instructions, automated MMPI administration is compatible with the more traditional paper-and-pencil administration (Rozensky, Honor, Rasinski, & Tovian, 1986; Russell, Peace, & Mellsop, 1986; White, Clements, & Fowler, 1985). Third, there are clearly established (validity) test correlates that can be automatically applied to a particular profile.

Automated Test Scoring and Interpretation

The first IBM machine utilized in test scoring was developed in 1935. Today, over 60 years later, computers are used in scoring a wide range of objective personality assessments. The use of

computers has not been limited to objective personality-assessments. Computers are also used in the scoring and interpretation of projective personality assessment measures such as the Rorschach test (Piotrowski, 1980; Exner, 1987). A comprehensive list of psychological software and services commercially available up to 1987 is provided by Butcher (1987).

A wide range of options is available for the computerized interpretation of the MMPI. One of the first available formats was the mail-in service, which is still in use today. In using this service, MMPI answer sheets are completed using the traditional paper-and-pencil format and mailed to the service for scoring. If the clinician has a personal computer available, software programs for computer processing is another option. In this case, either on-line or traditional paper-and-pencil administration may be used for scoring the MMPI. Also, a scanner can be attached to a personal computer and used to score a large number of answer sheets in a short period of time.

It is important to keep in mind that automated reports are based on data collected from both actuarial tables (based on empirical relationships) and clinical experience. Introducing clinical experience as an element in the automated interpretation requires close attention. The accuracy of any such automated report is directly related to the expertise, knowledge, research, and clinical experience of the clinician who writes the program and fills the gaps in the actuarial tables. Therefore, it is strongly recommended that one first inquire about the adequacy and accuracy of interpretation services (Fowler, 1987). Butcher (1995b) suggests several questions that the potential user of a computerized test interpretation should keep in mind:

1. Does the procedure on which the computer interpretation system is based have an adequate network of established validity research?
2. Do the system developers have the requisite experience with the particular test(s) to develop reliable, valid interpretation?
3. Is there a sufficient amount of documentation available on the system? Is there a published *user's guide* available to explain the test and system variables?
4. Is the system flexible enough to incorporate new research information as it becomes available?
5. Do the system developers follow the APA guidelines for computer-based tests?
6. Do the reports contain an effective evaluation of potentially invalidating response conditions?
7. Does the system closely follow the empirically validated test correlates?
8. Does the company providing computer interpretation services have a qualified technical staff to deal with questions or problems that could emerge?
9. Are the reports sufficiently annotated to indicate appropriate interpretive cautions? (p. 80)

Based on updated American Psychological Association (APA) guidelines, the professional using the scoring service is responsible for assessing the relevance of computerized interpretations for particular clients. The APA code of ethics holds psychologists responsible for their interpretation and recommendations based on psychological assessment (APA, 1992). For this reason, it is incumbent upon psychologists to choose a reliable and valid computerized interpretation program. At the same time, it is also important to emphasize that computerized interpretations generate hypotheses to be considered by the professional, keeping in mind the unique characteristics of the test taker in conjunction with any special circumstances. Information from the computer reports that is validated by other sources should receive additional weight in refining the emerging clinical picture.

Pros and Cons of Computerized Psychological Assessment

Computerized assessment owes much of its recent growth and status to the unique advantages that computers offer to the task of psychological assessment in comparison to clinician derived assessments. First, computers are time and cost efficient. Computerized reports can be available shortly after the completion of the test administration, saving valuable professional time.

Another advantage of using computers in psychological assessment is their accuracy in scoring,

inasmuch as computers are less subject to human error when scoring (Allard, Butcher, Faust, & Shea 1995; Skinner, & Pakula, 1986).

Third, computers provide more objective and less biased interpretations by minimizing the possibility of selective interpretation of data.

A fourth advantage of computerized reports is that they are usually more comprehensive and thorough than clinicians' reports. In a computerized interpretation, the test taker's profile is examined in comparison to many other profiles. Therefore, test information can be more accurately used to classify the individual, while describing the behaviors, actions, and thoughts of people with similar profiles. In sum, a well-designed statistical treatment of test results and ancillary information will generally yield more valid conclusions than an individual professional using the same information (APA, 1986).

Finally, computerized test administration may be more interesting to some subjects, who may also feel less anxious responding to a computer monitor (Rozensky et al., 1986) than the more personal context of a paper-and-pencil test.

While the advantages of computerized assessment are many, this method is not totally problem-free. One major problem associated with automated administration, scoring, and interpretation is misuse by unqualified professionals. Skinner and Pakula (1986) suggest that computerized assessment may inadvertently encourage use by professionals without adequate knowledge and experience. It is important to keep in mind that the validity of the information obtained by computerized psychological assessment can be ensured only in the hands of a professional with adequate training and experience with the particular test in question. Turkington (1984) pointed out that computerized assessment, when used by those with little training or skill in test interpretation, can do more harm than good.

A second risk of the computer-assisted assessment is that mental-health professionals might become excessively dependent on computer reports, and accordingly become less active in personally interpreting test data. Computerized reports cannot take the place of important clinical observations, which provide essential information to be integrated with results from formal testing (Butcher, 1995b).

A third problem comes from the fallacy that computer-generated assessments yield information that is necessarily factual. Matarazzo (1983, 1986) cautioned professionals against the face validity of computer-generated interpretations. It cannot be assumed that computer assessments generate precise scientific statements that cannot be questioned. Computer-based conclusions are not chiseled in stone, and a critical review of such interpretation is necessary for their credible use.

Fourth, b computer statements in a computer report might not provide specific information about the test taker useful for diagnostic purposes. Practitioners should be cautious of "Barnum-type" statements that some computer reports may generate. Basing clinical decisions on this type of statement can lead to inaccurate recommendations (Butcher, 1995).

Finally, a computerized report might include statements that do not apply to every patient. It is important to keep in mind that computer reports are general descriptions of profiles. It is quite possible that individuals with similar profiles will not possess all of the characteristics identified by a particular profile. It is incumbent on the professional to ascertain the accuracy of test reports for each individual client (Butcher, 1995b).

Selection of Computer-based Services

Computer-interpretation services generate reports varying in accuracy of interpretation (Eyde, Kowal, & Fishburne, 1987). Therefore, in choosing a computerized interpretation program it is important to choose a reliable program that represents the test's database in compiling the narrative information. Eyde et al. (1987, 1993) compared accuracy of seven commercially available computerized programs and concluded that reports were diverse in their usefulness and accuracy. In comparing different reports they determined that the Minnesota report "received the highest number of accuracy ratings and the lowest number of inaccuracy ratings for the clinical cases." Butcher (1988) compared the computer-based interpretations (Minnesota Personnel Report, National Computer Systems [NCS]) with clinicians' reports of 262 airline pilot applicants. They reported high agreement between the computer and the clinicians' adjustment rating of the applicants. There was 98.5 percent agreement between the computer and clinicians' decisions on classifying an applicant as emotionally stable. An 88 percent concordance between com-

puter and clinician-based judgments was documented in classifying an applicant as possibly having emotional problems. Therefore, computer-generated interpretation results were well in line with the professional judgment of clinicians.

Adaptive Test Administration

The merits of adaptive test administration has been a focus of attention for decades. Adaptive assessment individually tailors a set of items to the unique qualities of each person during the testing process (Butcher, 1987). The advent of sophisticated computer abilities paved the road for adaptive computerized assessment. During an on-line administration of the test, the computer scores each response, presents the next item based on the previous responses, and terminates the test when certain objectives are met (Butcher, 1987).

One of the most appealing advantages of adaptive test administration is its flexibility in altering the order of questions presented to the subject based on previous responses. In a clinical interview, a psychologist might bypass an entire section comprised of questions regarding particular aspects of manic episodes (e.g., antecedents, onset, duration) if a prior interview question determined the absence of manic episodes for that particular patient. This kind of flexibility in computer-based interviews and assessments is a recent and welcomed development. Another advantage of adaptive computerized assessment is its efficiency in providing the same type of results as paper-and-pencil administration in considerably less time. Because an adaptive computerized assessment test is individualized for a particular test taker, a smaller number of items need be administered.

> Research investigating the [utility] of adaptive testing strategies in personality assessment" could be summarized in four areas, using different methods: "(a) the prediction of full scores on the conventional paper and pencil form of the test, (b) the adaptive topological approach, (c) the countdown strategy, and (d) methods based on item response theory (IRT).

The latter two approaches have been more extensively studied. The countdown method is one of the adaptive computerized strategies that has been used with the MMPI/MMPI-2 (Butcher, Keller, & Bacon, 1985). This test format allows for expediency in classifying test takers as "normal" or "abnormal." Specifically, it provides information on whether each individual's T-scores on different MMPI scales exceed 70. With the countdown method, item administration for each scale is terminated when the number of unendorsed items equals the number of items on the scale minus the cutoff (the number of endorsed items which correspond to a T-score of 70) plus one, or when that items answered in the keyed direction add up to the cutoff. For example, if a scale is 30 items long, and 20 endorsed items are required to obtain a T-score of 70 (the cutoff), as soon as an individual fails to endorse 11 items, administration of that scale can be terminated because there is no possibility of the subject exceeding a T-score of 70. This way, items from each scale are administered until the individual reaches the number of endorsed items required to reach a T-score of 70 or until it is established that the individual can not reach this T-score even if the remaining items in that scale are endorsed in a scorable direction.

Item-Response theory (IRT) is the basis for one of the more psychometrically sophisticated adaptive testing techniques. Briefly, according to Weiss (1985), IRT-based adaptive testing

> [S]elects items that provide maximum levels of item information at an individual's currently estimated trait level. In addition, IRT-based methods of scoring tests permit estimated trait level. In addition, IRT-based methods of scoring tests permit estimation of individual's trait levels based on their responses to one or more items. As a consequence, an item can be administered and an estimate can be made of the individual's level on the trait. After the administration of an item and estimation of trait, the new trait level is used to select the next item to be administered to that examinee. (p. 783)

A comprehensive discussion of this topic is presented by Weiss (1985) and Weiss and Vale (1987). The use of IRT adaptive administration has not produced favorable results with the MMPI (Carter, 1982). Carter and Wilkinson (1984) proposed that some of the MMPI items do not possess discriminant validity in this model. They suggested that IRT adaptive administration, which takes item discrimination into account, could be useful. However, some experts believe that IRT-based adaptive administration for tests that are empirically keyed is inappropriate (Ben-Porath & Butcher, 1986). One of the major assumptions of IRT is trait unidimensionality, which entails that factorially derived

assessments are more appropriate for this adaptive procedure.

Several lines of data confirm the comparability of the computerized adaptive (countdown method) and standard administration of the MMPI-2. Simulated computer data confirmed comparable diagnostic decisions and up to 38 percent item saving with this procedure (Ben-Porath et al. 1989; Slutske, Ben-Porath, & Butcher, 1988). In addition, the comparability of the results was confirmed when the scores of college students from the computer adaptive administration were compared with the completed administration of all items (Slutske, Ben-Porath, Roper, Nguyen, & Butcher, 1990). In another study, Roper, Ben-Porath, & Butcher (1991) administered both the computer adaptive version and the standard MMPI-2 to 155 college students one week apart. They demonstrated that the adaptive approach was comparable to the standard administration of the MMPI in generating identical clinical interpretations. While the preliminary data on the use of adaptive computerized testing with normal populations is promising, the efficacy of this method with psychiatric patients awaits further investigation.

SUMMARY

The widespread use of MMPI provides a good example of how, as human beings, our natural interest in understanding and exploring personality stretches across language, geographical, and cultural boundaries. Presently, there are over 150 MMPI translations and 25 translations of the MMPI-2 in 45 countries. The successful adaptation and wide-spread use of this instrument across the world is a strong testimony to its utility, validity, and generalizability.

If past behavior is a good predictor of future behavior, then it is likely that our desire for understanding personality will continue to produce advances in this field. Undoubtedly, there will be new psychological instruments and modifications of existing ones in the future that are sparked by computer-technological advances. As the unique features of electronic computers continue to find favor among professionals in their effort to understand personality, computer-based assessment will likely be expanded even further.

NOTES

1. NEO-PI and its revised version are based on the five-factor theory of personality (Costa & McCrae, 1985; Costa & McCrae, 1992). MCMI-III: Million Clinical Multiaxial Inventory III (Millon, 1994) is a personality assessment useful in individuals with known psychopathology. DPI: Jackson Differential Personality Inventory (Jackson, 1972) and PRF: Personality Research Form (Jackson, 1984) are mostly used as research instruments.

REFERENCES

Allard, G., Butcher, J. N., Faust, D., & Shea, M. (1995). Errors in hand scoring objective personality tests: The case of the Personality Diagnostic Questionnaires Revised (PDQ-R). *Professional Psychology Research and Practice, 26,* 304–308.

American Psychological Association. (1986). *American Psychological Association guidelines for computer-based tests and interpretations.* Washington, DC: Author.

American Psychological Association. (1992). Ethical principles of psychologists and code of conduct. *American Psychologist, 47,* 1597–1611.

Anastasi, A. (1988). *Psychological testing.* New York: McMillan Publishing Company.

Arbisi, P. A., & Ben-Porath, Y. S. (1995). An MMPI-2 infrequent response scale for use with psychopathological populations: The F(p) Scale. *Psychological Assessment, 7,* 424–431.

Babcock, D. J. (1995). *Clinical equivalence of the MMPI-2 with the MMPI in alcoholic male veteran samples.* Paper presented at the 25th Annual Symposium on Recent Developments of the MMPI (MMPI-2), Minneapolis, MN.

Baer, R. A., Wetter, M. W., & Berry, D. T. (1992). Detection of under reporting of psychopathology on the MMPI: A meta-analysis. *Clinical Psychology Review, 12,* 509–525.

Bagby, R., Buis, T., & Nicholson, R. A. (1995). Relative effectiveness of the standard validity scales in detecting fake-bad and fake-good responding: Replication and extension. *Psychological Assessment, 7,* 84–92.

Barefoot, J. C., Dahlstrom, W., & Williams, R. B. (1983). Hostility, CHD incidence, and total mortality: A 25 year follow-up study of

255 physicians. *Psychosomatic Medicine, 45,* 59–63.

Barron, F. (1953). An ego-strength scale which predicts response to psychotherapy. *Journal of Consulting Psychology, 17,* 327-333.

Baucom, D. H. (1985). Review of California Psychological Inventory. In J. V. Mitchell (Ed.), *The ninth mental measurements yearbook* (pp. 250–252). Lincoln, NE: University of Nebraska Press.

Ben-Porath, Y. S., & Butcher, J. N. (1986). Computers in personality assessment: A brief past, and ebullient present, and an expanding future. *Computers in Human Behavior, 2,* 167–182.

Ben-Porath, Y. S., & Butcher, J. N. (1988, March). *Exploratory analyses of rewritten MMPI items.* Paper presented at the 23rd Annual Symposium on Recent Developments in the Use of the MMPI, St. Petersburg, Florida.

Ben-Porath, Y. S., Butcher, J. N., & Graham, J. R. (1991). Contribution of the MMPI-2 content scales to the differential diagnosis of schizophrenia and major depression. *Psychological Assessment, 3,* 634–640.

Ben-Porath, Y. S., McCully, E., & Almagor, M. (1993). Incremental validity of the MMPI-2 content scales in the assessment of personality and psychopathology by self-report. *Journal of Personality Assessment, 61,* 557–575.

Ben-Porath, Y. S., Slutske, W. S., & Butcher, J. N. (1989). A real-data simulation of computerized adaptive administration of the MMPI. *Personality Assessment: A Journal of Consulting and Clinical Psychology, 1,* 18–22.

Ben-Porath, Y. S., & Tellegen, A. (1995). How (not) to evaluate the comparability of MMPI and MMPI-2 profile configurations: A reply to Humphrey and Dahlstrom. *Journal of Personality Assessment, 65,* 52–58.

Ben-Porath, Y. S., & Waller, N. G. (1992). Five big issues in clinical personality assessment: A rejoinder to Costa and McCrae. *Psychological Assessment, 4,* 23–25.

Berry, D. T., Baer, R. A., & Harris, M. J. (1991a). Detection of malingering on the MMPI: A meta-analysis. *Clinical Psychology Review, 11,* 585–591.

Berry, D. T., Wetter, M. W., & Baer, R. A. (1995). Assessment of malingering. In J. N. Butcher, (Ed.), *Clinical personality assessment: Practical approaches.* New York: Oxford University Press.

Berry, D. T., Wetter, M. W., Baer, R. A., Larsen, L., Clark, C., & Monroe, K. (1992). MMPI-2 random responding indices: Validation using a self-report methodology. *Psychological Assessment, 4,* 340–345.

Berry, D. T., Wetter, M. W., Baer, R. A., Widiger, T. A., Sumpter, J. C., Reynolds, S. K., & Hallam, R. A. (1991b). Detection of random responding on the MMPI-2: Utility of F, Back F and VRIN scales. *Psychological Assessment, 3,* 418–423.

Block, J. (1965). *The challenge of response sets: Unconfounding meaning, acquiescence, and social desirability in the MMPI.* New York: Appleton-Century-Crofts.

Butcher, J. N. (1995b). How to use computer-based reports. In J. N. Butcher, (Ed.), *Clinical Personality assessment: Practical approach.* New York: Oxford University Press.

Butcher, J. N. (1995a). Important considerations in the use of automated MMPI-2 reports. In J. N. Butcher, & J. R. Graham (Eds.), *Topics in MMPI-2 & MMPI-A Interpretation.* Minneapolis, MN: University of Minnesota Press.

Butcher, J. N. (1994). Psychological assessment by computer: Potential gains and problems to avoid. *Psychiatric Annals, 24,* 20–24.

Butcher, J. N. (1991). Screening for psychopathology: Industrial applications of the Minnesota Multiphasic Personality Inventory-2 (MMPI-2). In J. Jones, B. D. Steffey, & D. Bray (Eds.), *Applying psychology in business: The manager's handbook.* Boston: Lexington Books.

Butcher, J. N. (1990). *The MMPI-2 psychological treatment.* New York: Oxford University Press.

Butcher, J. N. (1988, March). *Use of the MMPI in personnel screening.* Paper presented at the 23rd Annual Symposium on recent developments in the use of the MMPI, St. Petersburg, Florida.

Butcher, J. N. (1987). *Psychological assessment.* New York: Basic Books, Inc.

Butcher, J. N. (1985). Review of Sixteen personality Factor Questionnaire. In J. V. Mitchell (Ed.), *The ninth mental measurements yearbook* (pp. 1391–1392). Lincoln, NE: University of Nebraska Press.

Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (1989). *Minnesota Multiphasic Personality Inventory-2 (MMPI-2): Manual for administration and scoring.* Minneapolis, MN: University of Minnesota Press.

Butcher, J. N., & Graham, J. R. (1994). The MMPI-2: A new standard for personality assessment and research in counseling settings. [Special Issue]. *Measurement and Evaluation in Counseling and Development, 27,* 131–150.

Butcher, J. N., Graham, J. R., & Ben-Porath, Y. S. (1995). Methodological problems and issues in MMPI/MMPI-2/MMPI-A research. *Psychological Assessment, 7,* 320–329.

Butcher, J. N., Graham, J. R., Williams, C. L., & Ben-Porath, Y. S. (1990). *Development and use of the MMPI-2 content scales.* Minneapolis, MN: University of Minnesota Press.

Butcher, J. N., & Han, K. (1995). Development of an MMPI-2 scale to assess the presentation of self in superlative manner. The S Scale. In J. N. Butcher, & C. D. Spielberger (Eds.), *Advances in personality assessment.* Hillsdale, NJ: LEA Press.

Butcher, J. N., Keller, L. S., & Bacon, S. F. (1985). Current developments and future directions in computerized personality assessment. *Journal of Consulting and Clinical Psychology, 53,* 803–815.

Butcher, J. N., Kendall, P. C., & Hoffman, N. (1980). MMPI short forms: Caution. *Journal of Consulting Psychology, 48,* 275–278.

Butcher, J. N., Jeffrey, T., Cayton, T. G., Colligan, S., Devore, J. R., & Minegawa, R. (1990). A study of active duty military personnel with the MMPI-2. *Military Psychology, 2,* 47–61.

Butcher, J. N., & Owen, P. L. (1978). Objective personality inventories: Recent research and some contemporary issues. In B. Wolman (Ed.), *Handbook of clinical diagnosis of mental disorders* (pp. 475-545). New York: Plenum Press.

Butcher, J. N., & Rouse, S. V. (1995). Personality: Individual differences and clinical assessment. *The annual review of Psychology, 47,* 87–111.

Butcher, J. N., & Williams, C. L. (1992). *Essentials of MMPI-2 and MMPI-A Interpretations.* Minneapolis, MN: University of Minnesota Press.

Carter, J. E. (1982). *A computerized adaptive version of the MMPI based on the Rasch model.* Unpublished doctoral dissertation, University of Chicago, Chicago Circle.

Carter, J. E., & Wilkinson, L. (1984). A latent trait analysis of the MMPI. *Multivariate Behavioral Research, 19,* 385-407.

Cattell, R. B. (1957). *Personality and motivation, structure and measurement.* Yonkers, NY: New World.

Cattell, R. B., Eber, H. W., & Tatsuoka, M. M. (1970). *Handbook for the Sixteen Personality Factor Questionnaire (16PF).* Champaign, IL: Institute for Personality and Ability Testing.

Chojnacki, J. T., & Walsh, W. (1992). The consistency of scores and configural patterns between the MMPI and MMPI-2. *Journal of Personality Assessment, 59,* 276–289.

Cohen, R. J., Swerdlik, M. E., & Smith, D. K. (1992). *Psychological testing and assessment: An introduction to tests and measurement.* Mountain View, CA: Mayfield Publishing Company.

Colligan, R. C., Osborne, D., Swenson, W. M., & Offord, K. P. (1984). The MMPI: A contemporary normative study. New York: Praeger.

Comrey, A. L. (1970). *EDITS manual for the Comrey Personality Scales.* San Diego, CA: Educational and Industrial Testing Service.

Cook, W. N., & Medley, D. M. (1954). Proposed hostility and pharisaic-virtue scales for the MMPI. *Journal of Applied Psychology, 38,* 414–418.

Costa, P. T., McCrae, R. R. (1985). *The NEO Personality Inventory manual.* Odessa, FL: Psychological Assessment Resources.

Costa, P. T., & McCrae, R. R. (1992). The NEO-PI-R/NEO-FFI professional manual. Odessa, FL: Psychological Assessment Resources, Inc.

Dahlstrom, W. G. (1980). Altered versions of the MMPI. In W. G. Dahlstrom & L. Dahlstrom (Eds.), *Basic readings on the MMPI: A new selection on personality measurement.* Minneapolis, MN: University of Minnesota Press.

Dahlstrom, W. G. (1992). Comparability of two-point high-point code patterns from original MMPI norms to MMPI-2 norms for the restandardization sample. *Journal of Personality Assessment, 59,* 153–164.

Dahlstrom, W. G., & Dahlstrom, L. (Eds.). (1980). *Basic readings on the MMPI: A new selection on personality measurement.* Minneapolis, MN: University of Minnesota Press.

Dahlstrom, W. G., Welsh, G. S., & Dahlstrom, L. E. (1972). *An MMPI handbook Vol. 1: Clinical interpretation* (rev. ed.). Minneapolis, MN: University of Minnesota Press.

Dahlstrom, W. G., Welsh, G. S., & Dahlstrom, L. E. (1975). *An MMPI handbook Vol. 2 Research applications* (rev. ed.). Minneapolis, MN: University of Minnesota Press.

Diehl, L. A. (1977). The relationship between demographic factors, MMPI scores and the social readjustment rating scale. *Dissertation Abstracts International, 38* (5-B), 2360.

Dyer, J. B., Sajwaj, T. E., & Ford, T. W. (1993, March). *MMPI-2 normative and comparative data for nuclear power plant personnel who were approved or denied securing clearances for psychological reasons.* Paper presented at the 28th Annual Symposium on recent developments in the use of the MMPI/MMPI-2, St. Petersburg, FL.

Erdberg, P. (1979). A systematic approach to providing feedback from the MMPI. In C. S. Newmark (Ed.), *MMPI Clinical and Research Trends,* (pp. 328–342). New York: Praeger.

Exner, J. E. (1987). Computer assistance in Rorschach interpretation. In J. N. Butcher (Ed.), *Computerized psychological assessment: A practitioner's guide* (pp. 218–235). New York: Basic Books.

Eyde, L., Kowal, D., & Fishburne, F. J. (1987, August). *Clinical implications of validity research on computer based test interpretations of the MMPI.* Paper presented at the annual meeting of the American Psychological Association, New York, NY.

Eyde, L., Kowal, D., & Fishburne, F. J. (1993). *The computer and the decision making process.* Hillsdale, NJ: LEA Press.

Eysenck, H. J. (1985). Review of California Psychological Inventory. In J. V. Mitchell (Ed.), *The ninth mental measurements yearbook.* Lincoln, NE: University of Nebraska Press.

Finn, S. E., & Tonsager, M. E. (1992). Therapeutic effects of providing MMPI-2 test feedback to college students awaiting therapy. *Psychological Assessment, 4,* 278–287.

Fowler, R. D. (1987). Developing a computer-based test interpretation system. In J. N. Butcher (Ed.), *Computerized psychological assessment: A practitioner's guide* (pp. 50–63). New York: Basic Books.

Gogh, H. G. (1975). *California Psychological Inventory* (rev. manual). Palo Alto, CA: Consulting Psychologists Press.

Graham, J. R. (1993). *MMPI-2: Assessing personality and psychopathology.* New York: Oxford University Press.

Graham, J. R., & Ben-Porath, Y. S. (1995, March). *Correlates of MMPI-2 code types in an outpatient sample.* Paper presented at the 30th Annual Symposium on recent developments in the use of the MMPI, MMPI-2, and MMPI-A, St. Petersburg, FL.

Graham, J. R., & Butcher, J. N. (1988, March). *Differentiating schizophrenic and major affective disordered inpatients with the revised form of the MMPI.* Paper presented at the 23rd Annual Symposium on Recent Developments in the Use of the MMPI, St. Petersburg. FL.

Graham, J. R., Smith, R. L., & Schwartz, G. F. (1986). Stability of MMPI configurations for psychiatric inpatients. *Journal of Consulting and Clinical Psychology, 54,* 375–380.

Graham, J. R., Timbrook, R. E., Ben-Porath, Y. S., & Butcher, J. N. (1991). Code-type congruence between MMPI and MMPI-2: Separating fact from artifact. *Journal of Personality Assessment, 57,* 205–215.

Graham, J. R., Watts, D., & Timbrook, R. E. (1991). Detecting fake-good and fake-bad profiles with the MMPI-2. *Journal of Personality Assessment, 57,* 264–277.

Grayson, H. M. (1951). *Psychological admissions testing program and manual.* Los Angeles: Veterans Administration Center, Neuropsychiatric Hospital.

Greene, R. L. (1991). *MMPI-2/MMPI: An interpretive manual.* Boston: Allyn & Bacon.

Greene, R. L, Weed, N. C., Butcher, J. N., Arredondo, R., & Davis, H. G. (1992). A cross-validation of MMPI-2 substance abuse scales. *Journal of Personality Assessment, 58,* 405–510.

Guilford, J. P., & Zimmerman, W. S. (Eds.). (1956). Fourteen dimensions of temperament. *Psychological Monographs, 70* (No. 417).

Han, K., Weed, N., Calhoun, R., & Butcher, J. N. (1995). Psychometric characteristics of the MMPI-2 Cook-Medley Hostility Scale. *Journal of Personality Assessment, 63,* 567–585.

Harrell, T. W., Honaker, L. M., & Parnell, T. (1992). Equivalence of the MMPI-2 with the MMPI in psychiatric patients. *Psychological Assessment, 4,* 460–465.

Harris, R. E., & Lingoes, J. C. (1968). *Subscales for the Minnesota Multiphasic Personality Inventory.* Unpublished manuscript.

Hart, T. R., McNeill, J. W., Lutz, D. J., & Adkins, T. G. (1986). Clinical comparability of the standard MMPI and MMPI-2. *Professional Psychology: Research and Practice, 17,* 269–272.

Hathaway, S. R., & McKinley, J. C. (1940). A multiphasic personality schedule (Minnesota): I. Construction of the schedule. *Journal of Psychology, 10,* 249-254.

Hjemboe, S., Almagor, M., & Butcher, J. N. (1992). Empirical assessment of marital distress: The Marital Distress Scale (MDS) for the MMPI-2. In J. N. Butcher & C. D. Spielberger (Eds.), *Advances in psychological assessment: Volume 9.* Hillsdale, NJ: Lawrence Erlbaum.

Humphrey, D. H., & Dahlstrom, W. G. (1995). The impact of changing from the MMPI to the MMPI-2 on profile configuration. *Journal Personality Assessment, 64,* 428–439.

Jackson, D. N. (1972). *Differential Personality Inventory.* London, Ontario: Author.

Jackson, D. N. (1984). *Personality Research Form manual.* Port Huron, MI: Research Psychologist Press.

Keller, L. S., & Butcher, J. N. (1991). *Assessment of chronic pain with the MMPI-2.* Minneapolis, MN: University of Minnesota Press.

Koss, M. P. (1980). Assessing psychological emergencies with the MMPI. In J. N. Butcher, G. Dahlstrom, M. Gunther, & W. Schofield (Eds.), *Clinical notes on the MMPI.* Roche Psychiatric Service Institute Monograph Series. Nutley, NJ: Hoffman-LaRoche.

Koss, M. P., & Butcher, J. N. (1973). A comparison of psychiatric patients' self-report with other sources of clinical information. *Journal of Research in Personality, 7,* 225-236.

Koss, M. P., Butcher, J. N., & Hoffman, N. G. (1976). The MMPI critical items: How well do they work? *Journal of Consulting and Clinical Psychology, 44,* 921-928.

Kramer, J. J., & Conoley, J. C. (1992). *The eleventh mental measurements yearbook.* Lincoln, NE; University of Nebraska Press.

Krug, S. E. (1980). *Clinical Analysis Questionnaire manual.* Champaign, IL: Institute for Personality and Ability Testing.

Lachar, D., & Wrobel, T. A. (1979). Validation of clinicians' hunches: Construction of a new MMPI critical item set. *Journal of Consulting and Clinical Psychology, 47,* 277-284.

Litz, B. T., Penk, W. E., Walsh, S., Hyer, L., Blake, D. D., Marx, B., Keane, T. M., & Bitman, D. (1991). Similarities and differences between MMPI and MMPI-2 applications to the assessment of posttraumatic stress disorder. *Journal of Personality Assessment, 57(2),* 238–253.

Lubin, B., Larsen, R. M., & Matarazzo, J. (1984). Patterns of psychological test usage in the United States 1935–1982. *American Psychologist, 39,* 451–454.

Lubin, B., Larsen, R. M., Matarazzo, J., & Seever, M. F. (1985). Psychological test usage patterns in five professional settings. *American Psychologist, 40,* 857–861.

MacAndrew, C. (1965). The differentiation of male alcoholic outpatients from nonalcoholic psychiatric patients by means of the MMPI. *Quarterly Journal of Studies on Alcohol, 26,* 238-246.

Mattarazzo, J. D. (1983). Computerized psychological testing. *Science, 221,* 323.

Mattarazzo, J. D. (1986). Computerized clinical psychological test interpretations: Unvalidated plus all mean and no sigma. *American Psychologist, 41,* 14–24.

McKinley, J. C., Hathaway, S. R., & Meehl, P. E. (1948). The MMPI: VI. The K scale. *Journal of Consulting Psychology, 12,* 20-31.

Megargee, E. I. (1972). *The California Psychological Inventory handbook.* San Francisco: Jossey-Bass.

Millon, T. (1994). *Millon Clinical Multiaxial Inventory III.* Minneapolis, MN: National Computer System.

Mooney, R. L., & Gordon, L. V. (1950). *Mooney Problem Checklist.* New York: Psychological Corporation.

Piotrowski, Z. A. (1980). CPR: The psychological x-ray in mental disorders. In J. B. Sidowski, J. H. Johnson, & T. A. Williams (Eds.), *Technology in mental health care delivery systems* (pp. 85–108). Norwood, NJ: Ablex.

Piotrowski, Z. A., & Keller, J. W. (1989). Psychological testing in outpatient mental health facilities: A national study. *Professional Psychology: Research and Practice, 20,* 423–425.

Pope, K. S., Butcher, J. N., & Seelen, J. (1993). *MMPI/MMPI-2/MMPI-A in court: Assessment, testimony, and cross-examination for*

expert witnesses and attorneys. Washington, DC: American Psychological Association.

Quirk, M. P., Strosahl, K., Kreilkamp, T., & Erdberg, P. (1995). Personality feedback consultation in a managed mental health care practice. *Professional Psychology: Research and Practice, 26,* 27–32.

Rome, H. P., Swenson, W. M., Mataya, P., McCarthy, C. E., Pearson, J. S., Keating, F. R., & Hathaway, S. R. (1962). Symposium on automation techniques in personality assessment. *Proceedings of the Staff Meetings of the Mayo Clinic, 37,* 61–82.

Roper, B. L., Ben-Porath, Y. S., & Butcher, J. N. (1991). Comparability of computerized adaptive and conventional testing with the MMPI-2. *Journal of Personality Assessment, 57,* 278–290.

Rozensky, R. H., Honor, L. F., Rasinski, K., & Tovian, S. (1986). Paper-and-pencil versus computer-administered MMPIs: A comparison of patients' attitudes. *Computers in Human Behavior, 2,* 111–116.

Russell, G. K., Peace, K. A., & Mellsop, G. W. (1986). The reliability of a micro-computer administration of the MMPI. *Journal of Clinical Psychology, 42,* 120–122.

Schretlen, D., Wilkins, S. S., Van-Gorp, W. G., & Bobholz, J. H. (1992). Cross validation of a psychological test battery to detect faked insanity. *Psychological Assessment, 4,* 77–83.

Schuerger, J. M., Foerstner, S. B., Serkownek, K., & Ritz, G. (1987). History and validates of the Serkownek subscales for MMPI scales 5 and 0. *Psychological Reports, 61,* 227-235.

Skinner, H. A., & Pakula, A. (1986). Challenge of computers in psychological assessment. *Professional Psychology: Research and Practice, 17,* 44–50.

Slutske, W. S., Ben-Porath, Y. S., & Butcher, J. N. (1988, March). *A real-data simulation study of adaptive MMPI administration.* Paper presented at the 23rd Annual Symposium on recent developments in the use of MMPI, St. Petersburg, Florida.

Slutske, W. S., Ben-Porath, Y. S., Roper, B., Nguyen, P., & Butcher, J. N. (1990). *An empirical study of the computer adaptive MMPI-2.* Paper presented at the 25th Annual Symposium on Recent Developments of the MMPI (MMPI-2), Minneapolis, MN.

Sundberg, N. D. (1977). *Assessment of Persons.* Englewood Cliffs, NJ: Prentice Hall.

Tellegen, A., & Ben-Porath, Y. S. (1993). Code type comparability of the MMPI and MMPI-2: Analysis of recent findings and criticism. *Journal Personality Assessment, 61,* 489–500.

Turkington, C. (1984). The growing use and abuse of computer testing. *APA Monitor, 7,* 26.

University of Minnesota Press. (1984). Users guide for the Minnesota report: Personnel selection system. Minneapolis, MN: Author.

Vincent, K. R. (1990). The fragile nature of MMPI code types. *Journal of Clinical Psychology, 46,* 800–802.

Watkins, C. E., Campbell, V. L., & McGregor, P. (1988). Counseling psychologists' uses of the opinions about psychological tests: A contemporary perspective. *Counseling Psychologist, 16,* 476–486.

Weed, N. C., Butcher, J. N., McKenna, T., & Ben-Porath, Y. S. (1992). New measures for assessing alcohol and drug abuse with the MMPI-2: The APS and AAS. *Journal of Personality Assessment, 58,* 389–404.

Weiss, D. J. (1985). Adaptive testing by computer. *Journal of Consulting and Clinical Psychology, 53,* 774–789.

Weiss, D. J., & Vale, C. D. (1987). Computerized adaptive testing for measuring abilities and other psychological variables. In J. N. Butcher (Ed.), *Computerized psychological assessment: A practitioner's guide* (pp. 325–343). New York: Basic Books.

Welsh, G. S., & Dahlstrom, W. G. (Eds.), (1956). *Basic readings on the MMPI in psychology and medicine.* Minneapolis, MN: University of Minnesota Press.

Wetter, M. W., Baer, R. A., Berry, D. T., Smith, G. T., & Larsen, L. (1992). Sensitivity of MMPI-2 validity scales to random responding and malingering. *Psychological Assessment, 4,* 369–374.

White, D. M., Clements, C. B., & Fowler, R. D. (1985). A comparison of computer administration with standard administration of the MMPI. *Computers in Human Behavior, 1,* 153–162.

Whitworth, R. H., & McBlaine, D. D. (1993). Comparison of the MMPI and MMPI-2 administered to Anglo and Hispanic-American university students. *Journal of Personality Assessment, 61,* 19–27.

Wiggins, J. S. (1969). Content dimensions in the MMPI. In J. N. Butcher (Ed.), *MMPI: Research developments and clinical applications.* New York: McGraw-Hill.

Winder. P., O'Dell, J., & Karson, S. (1975). New motivational distortion scales for the 16PF. *Journal of Personality Assessment, 39,* 532–537.

Woodworth, R. S. (1917). *Personal Data Sheet.* Chicago: Stoelting.

Zuckerman, M. (1985). Review of Sixteen Personality Factor Questionnaire. In J. V. Mitchell (Ed.), *The ninth mental measurements yearbook* (pp. 1392-1394). Lincoln, NE: University of Nebraska Press.

This Page Intentionally Left Blank

CHAPTER 17

RORSCHACH ASSESSMENT

Philip Erdberg

INTRODUCTION

The assessment technique that Hermann Rorschach introduced in 1921 has certainly had its share of critics. But even they must concede the resilience of an instrument that, against considerable odds, has now survived well into its second half-century.

Rorschach died a year after the test's initial publication, leaving the fledgling instrument in the care of three of his associates. In little more than a decade it had traveled across the Atlantic. Once in America, it found itself with five groups of increasingly diverging adoptive parents whose differences ultimately became so extensive as to threaten its identity. Methodological criticism came from outside the Rorschach community as well, and there were suggestions that the test be discarded entirely.

But by the mid-1970s, a new consolidation had integrated the best of what had been learned during the half century of divergence, and the Rorschach now appears to have entered what may well be its healthiest years to date. The history of the test's development, a review of its elements, and some descriptions of new directions are the subjects of this chapter.

HISTORY AND DEVELOPMENT

The idea that associations to ambiguous visual stimuli could help in understanding a person is an ancient one. Early writings suggest that the classical Greeks were interested in the interaction of ambiguity and the person's characterization of reality (Piotrowski, 1957). By the 15th century, both Da Vinci and Botticelli had postulated a relationship between creativity and the processing of ambiguous materials (Zubin, Eron, & Schumer, 1965). Use of inkblots as stimuli for imagination achieved substantial popularity in Europe during the 19th century. A parlor game called Blotto asked participants to create responses to inkblots, and a book by Justinius Kerner (1857) contained a collection of poetic associations to inkblot-like designs.

As the 19th century ended, several workers in the professional community were beginning to utilize inkblots in the study of a variety of psychological operations. Krugman (1940) reports that Binet and Henri used inkblots to study visual imagination as early as 1895. Tulchin (1940) notes that Dearborn's work at Harvard (which resulted in 1897 and 1898 publications) employed inkblots as part of an experimental approach to the study of consciousness. Another American investigator, Whipple (1910), also used a series of inkblots as a way of studying what he called "active imagination." Rybakow (1910), working in Moscow, developed a series of eight blots to tap imaginative function, and Hens (1917), a staff member at Bleuler's clinic in Zurich, used inkblots with children, nonpatient adults, and psychiatric patients.

The young Swiss psychiatrist Hermann Rorschach thus was not the first to involve inkblots in the study of psychological processes when he began his project in 1911. But his work was qualitatively different from anything that had preceded

it, in that he used inkblots to generate data from which extensive personality descriptions could be developed. Rorschach's preliminary but remarkably farsighted *Psychodiagnostik* was published in 1921. Tragically, he died within a year, at the age of 38, of complications of appendicitis.

It was three of Rorschach's friends, Walter Morgenthaler, Emil Oberholzer, and George Roemer, who insured that the insights and challenges of *Psychodiagnostik* were not lost. Morgenthaler had championed the book's publication against some resistance from the Bircher publishing house. Oberholzer followed up by insuring that an important posthumous paper (Rorschach & Oberholzer, 1923) was published, and all three continued to teach the test and encourage adherents. One of Oberholzer's students, David Levy, took the test to Chicago, where he established the first American Rorschach seminar in 1925.

Although each could have, neither Oberholzer nor Levy moved into a clear position as Rorschach's successor, and once in America, the test was adopted by five psychologists of very different backgrounds—Samuel Beck, Bruno Klopfer, Zygmunt Piotrowski, Marguerite Hertz, and David Rapaport. Of the five, only Beck, through the opportunity of a year's fellowship with Oberholzer in Zurich, was able to study with someone who had worked directly with Rorschach.

With little in the way of common heritage or experience, the five Americans soon diverged in directions consistent with their theoretical orientations. They ultimately produced five independent Rorschach systems, each attracting adherents and each generating a body of literature and clinical lore. The history of the Rorschach from the late 1920s to the early 1970s is, to a large extent, the history of the development and elaboration of these five systems.

Beck completed the first American Rorschach dissertation in 1932. He followed it with a number of journal articles and published his *Introduction to The Rorschach Method* in 1937. He completed the elaboration of his system with additional books in 1944, 1945, and 1952, with revised editions published through 1967.

Klopfer had his first direct contact with the Rorschach in 1933. After a series of articles, which included a description of a scoring system (Klopfer & Sender, 1936), he published *The Rorschach Technique* with Douglas Kelley in 1942. Elaborations of his system occurred in books in 1954, 1956, and 1970.

Piotrowski was a member of a seminar offered by Klopfer in 1934, but within two years he had moved toward the creation of an independent system. His work culminated with the publication of *Perceptanalysis* in 1957.

Hertz, who worked briefly with Levy and Beck, utilized the Rorschach in her dissertation in 1932 and continued research with the test for the decade after at the Brush Foundation in Cleveland. Sadly, the nearly 3,000 cases she had amassed and an almost completed manuscript describing her system were inadvertently destroyed when the Foundation closed. Although she never produced another book, her steady stream of journal articles and ongoing seminars led to the consolidation of a Hertz system by 1945.

Rapaport became interested in the Rorschach in the late 1930s and published a paper describing it in detail as part of a review of projective techniques in 1942. The first volume of *Diagnostic Psychological Testing* was published with Merton Gill and Roy Schafer in 1945, with the second volume following a year later. Schafer extended the system with additional books in 1948, 1954, and 1967, and Robert Holt edited a revised edition of the original two volumes in 1968.

With publication of Piotrowski's book in 1957, all five systems were essentially complete. Each was taught independently and, during this period of divergence, each accumulated its own body of research and clinical literature. When Exner did a comprehensive review of the systems in 1969, he concluded that there were actually five overlapping but discrete tests. Each of the systematizers had taken Rorschach's 10 inkblots and used some of the ideas in *Psychodiagnostik* to create an instrument consistent with his or her training and theoretical stance. Each asked the subject to respond to the cards and each attempted some sort of inquiry as a way of clarifying how the person had generated responses. Each had developed a format for coding or "scoring" various aspects of the percept. And each system then generated an interpretation on the basis of the data that had been gathered. But, at every level from administration to interpretation, there were major differences among the five systems.

These differences made sense in the context of the different theoretical positions and methodologies that the five systematizers brought to each aspect of the Rorschach. At the coding level, Beck's rigorous positivism and behavioral training emerged in his insistence on normative and valida-

tional backing for the various elements. Klopfer's phenomenological background allowed examiners greater leeway in using their own experience for reference in coding the same material. Rapaport, Hertz, and Piotrowski used methodological approaches between those of Beck and Klopfer. At the interpretive level, Rapaport's extensive utilization of psychoanalytic concepts separated his work from the less theory-based interpretive strategies of the other four systems.

Describing all the ways the five systems differ one from another is an immense task. But a distinction suggested by Weiner (1977) identifies the critical question whose answer allows the characterization of an approach: How does the system conceptualize the nature of Rorschach data? Weiner suggested that the Rorschach can be viewed either as a perceptual-cognitive task or as a stimulus to fantasy.

The perceptual-cognitive stance assumes that the basic Rorschach task is to structure and organize an ambiguous stimulus field. The way a person accomplishes this task is directly *representative* of real-world behavior he or she will demonstrate in other situations requiring the same kinds of operations. As an example, people who solve the inkblots by breaking them into details that they combine into meaningful relationships ("two ladies stirring a pot of soup as a butterfly glides by") might be expected to engage their day-to-day tasks in a similarly energetic, integrative manner.

The focus in the perceptual-cognitive conceptualization of the Rorschach is not on the words but rather the structure of the person's responses, such as choice of location or integration of blot areas. This reliably quantifiable description of Rorschach response structure is then utilized to generate descriptions of how the person is likely to behave elsewhere. These descriptions are based on the body of validity studies that link Rorschach structural variables with real-world behavior.

The stimulus-to-fantasy approach, on the other hand, views the Rorschach as an opportunity for the person to project material about internal states onto the ambiguity of the blots. The person's productions are seen as *symbolic* of his or her dynamics. As an example, the percept of "two desperately disappointed people" might be utilized to infer a state of interpersonal conflict on the part of the person producing the response.

The focus in the stimulus-to-fantasy approach is on the actual words, and this content helps create hypotheses not about likely behavior but rather about internal states. Here the interpreter utilizes his or her theoretical framework and clinical experience to link symbols and dynamics.

There is some question about where Rorschach himself should be placed in terms of the perceptual-cognitive versus stimulus-to-fantasy distinction, but it is likely that he would have taken a middle-ground position that included both approaches. He had specifically criticized Hens' 1917 work for its focus solely on content and imagination. In doing this, he differentiated himself from Hens and, by implication, from most of the earlier inkblot workers, whose focus had been on verbalization and creative processes. Rorschach stated that his primary interest was what he called "the pattern of perceptive process," as opposed to the content of inkblot responses. *Psychodiagnostik* itself is almost totally in the perceptual-cognitive camp, with particular attention to issues of form, movement, and color. The 1923 posthumous paper added the structural element of shading.

And yet Rorschach was well trained in the work of Freud and Jung. He almost certainly would have been comfortable with Freud's 1896 description of projection as a mechanism by which individuals endow external material with aspects of their own dynamics—and with Frank's classic 1939 paper that suggested that stimuli such as inkblots could serve as "projective methods" for evoking this process. Indeed, Roemer (1967) states that Rorschach saw value in content analysis, citing a 1921 letter suggesting that he envisioned the technique as including both structural and symbolic material.

A review of their actual work in dealing with Rorschach material suggests that all five of the American systematizers also saw the data as having both perceptual-cognitive and symbolic components, but with differences in relative emphasis. Beck stayed closest to the structural aspects. Rapaport was most willing to place major emphasis on the verbalizations, but even he wrote "one can learn more about the subject sometimes by looking at a response from the point of view of its perceptual organization, and at other times by looking at it from the point of view of the associative processes that brought it forth" (Rapaport, Gill, & Schafer, 1946, p. 274).

Despite their overlap, each of the systems solidified and developed specialized terminology and literature. Clinicians schooled in one approach found it increasingly difficult to communicate with

those trained in other systems as each approach went its own way.

The purpose of Exner's development of the Comprehensive System (1974, 1978, 1982, 1986a, 1991, 1993, 1995) has been to provide the Rorschach community once again with a common methodology, language, and literature base. The accumulated literature of all five systems was reviewed, and some new research was undertaken.

Using reliability and validity as criteria for inclusion, the project yielded a constellation of empirically defensible elements that forms the structural aspect of the system. Content analysis is a secondary but significant part of the Comprehensive System. The approach to the handling of data as symbolic material can be characterized as dynamic but not specifically linked to any single theory of personality operation. What follows is a description of the elements of the Comprehensive System.

THE RORSCHACH ELEMENTS

A frequent role for the Rorschach clinician is as consultant to the intervention process, responding to questions raised by another professional. Typically the referral is made in the hope that personality assessment can supplement the referring professional's observations and provide additional understanding and guidance for intervention decisions. Because this is the way the test is often employed, this chapter presents a review of Rorschach elements and their supporting literature in the format of a series of clinical questions.

What is the Person's Preferred Style of Coping with Need States?

Faced with stressful situations, some individuals depend mostly on internal resources, while others are more likely to seek interaction with the outside world as a way of coping. The Erlebnistypus (*EB*) first proposed by Rorschach (1921) predicts which of these response tendencies is more likely for a particular person. A substantial number of studies (Molish, 1967; Singer & Brown, 1977; Exner, 1986a, provide reviews) have lent support to Rorschach's hypothesis that individuals who use a preponderance of human movement (*M*) in creating their Rorschach percepts (introversives) tend to rely on inner resources, while those (extratensives) who involve relatively more chromatic color (*FC*, *CF*, and *C*) are likely to seek interaction as a way of solving problems. Rorschach also identified a third response style, the ambitent, to describe individuals who do not have clearly skewed introversive or extratensive profiles.

A concurrent validity study described by Exner (1978) illustrates some behavioral correlates of the three styles and provides some reference points for contrasting them. Academically matched college students identified by the Rorschach as either introversives, extratensives, or ambitents participated in a logical analysis task. The experiment employed an apparatus in which each decision provided feedback about the combination of moves necessary to reach a solution. The students were scored on total moves to solution, total errors, total repeated moves, time between moves, and time to solution. The introversive group was characterized by fewer moves, longer times between moves, and fewer repeated moves and errors. The extratensive group had more moves and shorter times between moves than the introversives. The ambitents had the greatest number of moves, the greatest number of repeated moves and errors, and, most importantly, took the longest amount of time to get to solutions. Although stylistically very different, the introversives and extratensives did not differ significantly on the bottom-line variable: time to solution.

We can speculate from these data that the introversive group used a more "thoughtful" approach, processing feedback internally, while the extratensive group utilized more trial-and-error interaction with the environment and less internal processing. If we use time to solution as a measure of the efficiency of the problem-solving styles, the introversives and extratensives emerged as equally efficient. It is the ambitents whose approach was less effective.

A substantial amount of data (Exner, 1993) on both non-patient and pathological adults is consistent with the problem-solving study described above in suggesting that ambitents may be least able to cope with a variety of stressful situations. Within a non-patient U.S. sample, ambitents made up only 20 percent of the group, with the remaining 80 percent made up about equally of introversives (36 percent) and extratensives (44 percent). In contrast, ambitents accounted for 56 percent of an inpatient-depressive sample, 41 percent of an outpatient character-disorder group, and 30 percent of an inpatient-schizophrenic sample.

The *EB* is very much an enduring trait that predicts on an ongoing basis how individuals will cope with stressful situations. In one study, Exner (1986a) re-tested 39 clearly identified introversives and extratensives at one year. On re-test, 38 of the 39 still displayed the Rorschach style that had characterized them a year earlier.

The issue of problem-solving style as described by the *EB* has far-ranging implications for the clinician. We might speculate, for example, that spouses whose styles are different would have difficulty confronting a problem as a couple, since one member's style of using interaction to "talk the problem through" might interfere with the other person's need for solitude to "mull things over."

How Likely is the Person's Preferred Coping Style to Work? What Kinds of Problems Will it Encounter?

Although the introversive and extratensive approaches are stylistically very different, they both represent task-oriented coping approaches. They are both volitional strategies that the person calls on to handle problems. For that reason, Beck (1960) suggested that the summation of the human movement and color determinants (*EA*) could be utilized as a measure of the person's available psychological resources, those coping strategies the individual could decide to apply to solve problems. A clue as to whether these organized strategies can be expected to work is provided by considering another summation, the *es*, and its relation to *EA*.

The *es*, a variable suggested by Exner (1986a) comprises the unorganized psychological material that impinges on the person in unexpected and often disorganizing ways. Using *EA* as a measure of accessible coping strategies and *es* as a measure of nondeliberate, intrusive psychological material, Exner developed the *D* score, a scaled difference score that is generated by comparing *EA* to *es*. When the *D* score is zero or in the positive range, it suggests that "*under most circumstances*, sufficient resources are available to be able to initiate and direct behavior in a deliberate and meaningful way, and that stimulus demands being experienced generally do not exceed the capacities of the subject for being able to control behavior" (p. 315). If *D* is in the minus range, Exner suggests the likelihood of a situation in which "the frequency and/or intensity of stimulus demands exceeds the range of responses that can be formulated or implemented effectively" (p. 316).

If the *es* represents the totality of unorganized psychological material tapped by the Rorschach, a review of its components can help specify the kinds of difficulties that may interfere with the person's deliberate coping style, be it introversive, extratensive, or ambitent. Two Rorschach elements appear to be associated with disruptive ideation and four others with intrusive affect. We will discuss each of these components of the *es* in detail.

Although there are fewer validational studies for the animal movement (*FM*) determinant than for many other Rorschach variables, the available literature is consistent. It suggests that *FM* is associated with the experience of unorganized need-state ideation that intrudes into consciousness with an intensity that demands action. What happens then may well depend on the robustness of the person's coping strategies for dealing with the need state to which he or she has been alerted. When the unorganized ideation identified with *FM* is greater than the organized ideational style associated with *M*, the probability of impulsive behavior may go up.

Two studies (Exner & Murillo, 1975; Exner, Murillo, & Cannavo, 1973) found that when *FM* is greater than *M*, the likelihood of post-hospitalization relapse is greater for a variety of psychiatric patients. We can speculate that one of the reasons for the relapsers' inability to operate outside the hospital involved ongoing need states for which they did not have sufficient coping and delaying strategies and to which they responded impulsively and inappropriately.

Another sort of disruptive ideational experience is that which appears to be associated with the use of inanimate movement (*m*) in formulating Rorschach percepts. While the *FM* experience seems to involve ideation about internal need states, *m* appears associated with ideation provoked by the experience of stressful circumstances over which the person has little control.

Subject groups as varied as Navy personnel under severe storm conditions, depressed psychiatric inpatients the day before a first electoconvulsive therapy (ECT) treatment, parachute trainees the evening before their first jump, and hospital patients the day before elective surgery all showed more *m* in the Rorschachs they produced at these times than in baseline records (Shalit, 1965; Exner, 1986a). A series of temporal consistency studies (Exner, 1986a) suggests that the test-retest correla-

tions of *m* are notably lower than those of most other Rorschach variables, supporting the conceptualization of *m* as a situational or "state" variable.

The next group of Rorschach components associated with the individual's non-volitional operations appears to involve the experience of painful emotion as opposed to disruptive ideation. Each of these components is associated with a somewhat different type of distressing emotional experience, and it will be helpful to review each separately.

The use of the Rorschach's light-dark or shading features to formulate a percept involving texture (*FT*, *TF*, or *T*) appears to reflect the experience of a need for interpersonal contact that has more of an "emotional" than an "intellectual" quality. As an example, recently separated or divorced individuals who had not yet established new emotional relationships produced 2.7 times as much texture in their records as a group of demographically matched controls who rated their marriages at least average for stability and happiness (Exner, 1986a). Texture is the most frequent of the shading determinants, with most non-patients producing one texture determinant. It would appear that the distribution of texture for several patient groups has a more bimodal quality, with members of these groups producing either no texture or more than one texture determinant. We can speculate that these extremes are associated with disruptions in effective interpersonal function, with the high-texture individuals manifesting greater than average interpersonal neediness and the no-texture individuals uninterested in relationships that have an affective component.

Another Rorschach variable that can contribute to the amount of painful emotion impinging on the person involves the use of the shading features to formulate a percept of depth or dimensionality (*FV*, *VF*, or *V*). Use of this vista determinant appears associated with the sort of introspection that produces an unrealistically negative self-evaluation. Exner's data (1993) suggest that vista is relatively rare (20 percent) in adult non-patients, while it occurs in 55 percent of the members of an inpatient depressive sample. Exner and Wylie (1977), Exner (1986a), and Arffa (1982) have found that this generally rare variable is frequently present in the records of suicidal adults, adolescents, and children.

A third source of disruptive emotion is the experience that can be linked to the general use of the light-dark features of the blots (*FY*, *YF*, and *Y*). These diffuse shading determinants appear to suggest that the person is experiencing feelings of helplessness or resignation in the face of a stressful situation that demands action. Exner (1978) followed psychotherapy patients in a longitudinal study and found that those who were able to terminate by 18 months were characterized by significant decreases in the amount of diffuse shading in their records. Patients who were still in therapy at 18 months had about the same amount of diffuse shading as when they had begun treatment. We can hypothesize that diffuse shading is associated with an experience of helplessness in the face of stressful demands.

The final Rorschach variable contributing to disruptive emotion involves the utilization of the white-gray-black features of the blots (*FC'*, *C'F*, and *C'*). This achromatic color determinant appears related to the experience of containing affect instead of allowing its discharge. Exner (1986a) noted that several groups who could be expected to inhibit affective discharge—psychosomatics, obsessives, schizoids, and depressives who did not make suicide attempts—showed significantly more achromatic color in their Rorschachs than individuals whose behavior suggested less containment of affect (character disorders and depressives who made suicide attempts). It would appear that individuals who use achromatic color in producing their Rorschach percepts tend to internalize affect, with the resulting pressure and disequilibrium that this painful limiting of emotional expression can produce.

These, then, are the sources of intrusive ideation and emotion that appear to have Rorschach correlates. When these disruptive elements predominate in a person's psychological operations, they can interfere significantly with the ability to utilize task-oriented coping strategies effectively.

What is the Quality of the Person's Reality Testing?

The individual's ability to converge on percepts that are frequently seen or can be easily shared with others is a Rorschach indicator thought to be representative of accurate function in other day-to-day activities. Although there have been different methodologies used in establishment of this indicator (Kinder, Brubaker, Ingram, & Reading, 1982; Exner, 1986a), it is fair to say that Rorschach and all the systematizers since have viewed "form quality" as an important variable. The consistent

sense throughout all the systems is that form quality describes the individual's ability to operate conventionally, a sort of conflict-free ego function. This skill manifests very early in normal development. It is fascinating to note that the perceptual accuracy of non-patient seven-year-olds is within a few percentage points of that found for 16-year-olds and for adults (Exner & Weiner, 1995). Exner (1986a) provides an extensive review suggesting that significant deficits in Rorschach form quality are likely to be associated with "major impairment" (p. 369).

Rorschach's original recommendation was that percepts be differentiated on the basis of "good" versus "poor" form. Elaborations of this basic dichotomy have allowed for greater specificity in describing the individual's reality testing. Using a modification of an approach suggested by Mayman (1970), Exner (1993) divides convergent form responses into those involving superior articulation and those whose articulation is less elaborate. He divides non-convergent form responses into percepts that are not commonly seen but that do not significantly misrepresent external reality and those whose reality is arbitrary and internally driven.

This sort of distinction can be of substantial value in the assessment of schizophrenia, where the specification of the extent to which reality is internally informed may be diagnostic. Harder and Ritzler (1979), for example, found that a good form versus poor form dichotomy was unable to differentiate between psychotics and nonpsychotics in their inpatient sample, while approaches that made finer gradations within the good- and poor-form categories could differentiate the groups quite accurately.

Exner's (1993) X-% was designed to provide a single measure that indicates how frequently the individual has significantly distorted the blot contours in the production of percepts. The mean X-% for Exner's non-patient sample is 7 percent, while the mean for a sample of inpatient schizophrenics is 37 percent. An X-% in this range suggests that the individual's representation of reality may be so internally driven that ability to operate convergently would be compromised.

A contrast that may have substantial value in describing the person's reality testing is the distinction between perceptual convergence in relatively affect-free situations (F+%) versus situations involving more affective complexity (X+%). Typically, these two indicators are highly correlated, but when they differ markedly, the divergence may have clinical significance. We can speculate, for example, that individuals whose affect-free reality testing is significantly better than their perceptual convergence in emotionally toned situations might do well in structured treatment settings but would tend to have difficulty if they were placed in more ambiguous and complex environments.

How Mature and Complex are the Person's Psychological Operations?

There are a variety of ways to approach the inkblots, some of them involving substantially more complexity than others. These distinctions appear to be of value in describing the sophistication of the person's psychological operations in day-to-day settings.

Meili-Dworetzki (1956) found that as children increase in age, their location, selection, and integration of blot details become more complex. Smith (1981) classified 2nd- and 6th-grade students in terms of the Piagetian stages of cognitive development and found that children at the higher stages more frequently chose the whole blot as the location for their percepts and integrated various details into meaningful relationships.

Exner and Weiner (1995) found that these organized responses ("two waiters placing dishes on a table") increased from 23 percent in their six-year-old normative sample to 35 percent for their 16-year-olds, while vague percepts ("some kind of cloud") decreased from 13 percent to four percent. In the Comprehensive System, the coding of developmental quality encompasses a range from very diffuse percepts to those demonstrating the complex integration of form-dominated objects. Developmental-quality findings may thus provide data about the sophistication with which the person approaches the world.

Blends are percepts in which more than one determinant is used in producing the response ("two people talking as a red butterfly appears in the background"). Exner (1986a) suggests that, although there may be a very modest relation to intelligence, blends probably specifically reflect psychological complexity and awareness of the intricacies of oneself and one's environment. He goes on to suggest that either a very large or very small number of blends may be problematical. The large numbers may be associated with an immobi-

lizingly overcomplex style, while the low-blend individual may be characterized by limited ability or willingness to entertain complex alternatives when responding to demands.

With What Frequency and Efficiency Does the Person Attempt to Organize the Environment?

As noted above, an individual can either take a sort of "conservation of energy" approach to the Rorschach or attempt to organize the blot more energetically. The economical approach limits percepts to a single detail, while the more energetic style involves utilizing either the whole blot or integrating two or more details into a meaningful relationship. With only a few exceptions (the whole percepts on cards I and V, for example), responses involving wholes or the integration of details appear to represent a more organizationally challenging process. The frequency with which the person attempts this sort of energy-consuming integration (*Zf*) may provide a useful prediction of style in approaching the elements of the day-to-day world.

Although *Zf* does have a modest correlation with intelligence (Exner, 1974), it would appear that other stylistic variables play at least as great a part in determining how likely it is that the individual will attempt the synthesizing sorts of operations that this index reflects. For example, Exner (1986a) speculates that high *Zf* may be associated with a person's need for intellectual attainment or with a very precise way of dealing with detail. Low *Zf*, on the other hand, may reflect unwillingness to engage the complexity of the stimulus field.

Whatever the frequency of the person's organizational attempts, it is important to know whether each attempt is likely to be efficient or not. An index developed by Exner (1974) can be of substantial value in describing the precision of the person's integrative efforts. This index, the *Zd*, provides a measure of whether, for any given number of organizational attempts, the overall complexity of an individual's integrative operation is greater, less, or about the same as that of a primarily non-patient group studied by Wilson and Blake (1950).

Individuals with a high positive *Zd* tend to bring in more complexity per organizational attempt than the Wilson and Blake sample. They can be described as overincorporators. At the other extreme, a high negative *Zd* implies that the individual has involved less complexity in organizational attempts than the members of the normative sample did, an under-incorporative style.

A series of studies summarized by Exner (1993) suggests that the Rorschach finding of under-incorporative or over-incorporative style is associated with some quite consistent behavioral tendencies, whether the subjects were youngsters playing Simon Says, high school students doing a perceptual-spatial task, college students guessing from incomplete verbal data, or adults solving a serial-learning problem. The under-incorporators were characterized by fast speed but many errors, responding before they had fully scanned and processed the data. The over-incorporators tended to be more cautious in their response style, taking much longer to act. They waited for more data, sometimes to the point of redundancy, before making their decisions.

Both these extreme styles can be maladaptive. The under-incorporators run the risk of making decisions that were not informed by all the relevant data. The over-incorporators' need for "complete" data can be immobilizing, particularly in situations that involve time pressure or deadlines.

What is the Extent and Quality of the Person's Self-Focus?

Several Rorschach variables provide information about self-image. We will review each in detail.

There is some suggestion that use of the symmetrical properties of the blots to generate percepts involving pairs (2) or reflections (*Fr* or *rF*) is associated with self-focus. Exner (1973, 1986a) found that pair and reflection responses were positively associated with self-focused answers on a sentence-completion task and with mirror-looking behavior in a group of engineering job applicants waiting for an interview. These findings led to establishment of the Egocentricity Index, a weighted percentage of the number of reflections and pairs in a person's record.

If someone's Egocentricity Index is significantly higher than the norm for his or her age group, it suggests the likelihood of greater self-involvement than is developmentally appropriate. If it is notably lower than the age-group mean, there is a likelihood of the sort of negative self-concept that is seen in depressed individuals.

It would appear that one of the components of the Egocentricity Index, reflection responses, represents a somewhat more primitive and intense form of self-focus. Exner (1986a) reports that 20 percent of an outpatient character-disorder sample have at least one reflection response, as opposed to a seven-percent figure for adult nonpatients.

Presence of two Rorschach determinants, vista (*FV, VF,* or *V*) and form dimensionality (*FD*), suggest that the person is devoting some time to self-inspection. The form-dimensionality determinant suggests a somewhat more even-handed version of this self-focus. As noted above, vista is associated with an intensely devaluing self-appraisal in which the person is unable to put positive and negative aspects in perspective.

Morbid content on the Rorschach also has implications for negative self-concept. Exner (1993) reports presence of at least one morbid content percept in 69 percent of an inpatient depressive sample as opposed to 51 percent of his non-patient group. He reports that therapy patients who have three or more morbid content percepts were rated by their therapists as having more negative attitudes toward themselves and their presenting problems and less optimism about the future than patients without this Rorschach finding. Exner suggests that elevations in morbid content may indicate "that the self-image is conceptualized by the subject to include more negative and possibly damaged features than is commonplace and, second, that the orientation toward the self, and probably toward the environment, is marked by considerable pessimism" (p. 397).

With What Balance of Activity and Passivity Does the Person Interact with the World?

A differentiation of whether the person's movement responses are active ("someone building a house") or passive ("a bird gliding through the sky") appears to have substantial promise as a way of predicting a variety of important non-Rorschach behaviors. Exner (1974) found that acute schizophrenics, patients hospitalized for character disorders, and patients with a variety of diagnoses but a common history of assaultiveness were characterized by significantly more active movement responses. Chronic schizophrenics and depressives had significantly more passive movement percepts. Even more important, though, was his finding that approximately 70 percent of psychiatric patients had a skewed active-passive mix, while about the same percentage of non-patients had a more balanced mix of the two kinds of movement responses.

If the person's Rorschach active-passive balance is skewed one way or the other, a series of studies summarized by Exner (1993) suggests the likelihood of cognitive inflexibility in a variety of situations. When the progress of adolescents treated for behavioral problems was evaluated by significant others, most of those rated as improved had shifted from a skewed to a more balanced active-passive ratio. Almost all of those rated as unimproved had not made this shift.

Women whose active-passive mix was skewed were also characterized by a relatively rigid style when the actions of the central figure of their daydreams were evaluated. Women with more balanced active-passive ratios shifted much more frequently between active and passive modes for their daydreams' central figure. Psychoanalytically oriented therapists rated patients with skewed active-passive ratios lower for insight, progress, and overall session-effectiveness and higher for redundancy than they did a group of patients with a more even distribution of active and passive percepts (Exner, 1986a). High-school students with balanced active-passive ratios were able to come up with significantly more unusual or "creative" uses for familiar objects both singly and in combination than an academically matched group of students with very skewed active-passive ratios.

The common theme throughout these studies is that the Rorschach finding of a skewed distribution of active and passive percepts appears to be associated with the sort of cognitive rigidity that may limit the variety of the person's coping behaviors.

Two studies summarized by Exner (1986a) are of interest in suggesting correlates of the particular Rorschach finding of a skew in the direction of more passive percepts. When a measure of behavioral passivity was administered to the significant others of 279 outpatients, those with a passive skew in their Rorschachs were rated much higher for a variety of passive behaviors. Even more specifically, it would appear that when passive percepts for human movement (*Mp*) exceed active percepts (*Ma*), the person's ideation may be characterized by a sort of "magical thinking" that depends on the intervention of others at stressful times.

Two groups of non-patient adults, identified by the presence or absence of *Mp* greater than *Ma*, were asked to write endings for TAT stories in which the protagonist was portrayed as being in some sort of difficult situation. The *Mp* individuals brought new characters into their stories with significantly greater frequency. These "interveners," not the original protagonists, were significantly more often instrumental in initiating some sort of resolution. Exner describes this style as the "Snow White" feature: "being more likely to take flight into passive forms of fantasy as a defensive maneuver, and also being less likely to initiate decisions or behavior if the alternative that others will do so is available" (1986a, p. 374).

How Does the Person Respond to "Emotional" Experience?

Klopfer and Kelley (1942) and Beck (1961) suggested that the proportion of responses given to the three fully chromatic blots (cards VIII, IX, and X) may provide data about responsiveness to emotionally charged experiences in daily life. Non-patients typically give about 40 percent of their responses to these three cards. The Affective Ratio compares the number of responses to the three fully chromatic blots to the number of responses to the other seven cards. It is formulated so that high scores mean that the person has been proportionately overresponsive, and low scores mean that he or she has "backed away" from the fully chromatic inkblots. A series of studies summarized by Exner (1993) suggests that as the Affective Ratio goes up, so does receptiveness to emotionally complex situations and willingness to involve this material in making decisions.

It is noteworthy that patients are often at the extremes of the Affective Ratio, with bimodal distributions suggesting that they either underrespond or overrespond to the fully chromatic blots (Exner, 1986a). When tested again after treatment, those patients who were rated by significant others as improved had moved into the normal range with greater frequency than those rated as unimproved. It would appear that either underresponsiveness or overresponsiveness to the emotional parts of experience has the potential for generating maladaptive function.

If we view the Affective Ratio as providing a probability statement about how likely it is that emotional stimuli will be processed and responded to, an important next question concerns how well moderated the response will be when it does occur. The person's integration of form and color on the Rorschach appears to be associated with this sort of moderation. Form-dominated color (*FC*) percepts ("a swallow-tail butterfly—here's the head and wings and it's red") are associated with well-modulated affective responding, while *CF* or *C* color-dominated form or pure color percepts ("blood—it's all red") are more likely to be associated with intense emotional displays.

Gill (1966) found that ability to delay responses in a problem-solving task was associated with significantly more *FC* percepts. Individuals who could not delay their responses were characterized by significantly more *CF* and *C*. Adult non-patients typically have at least as much *FC* as *CF* and *C* in their records. Exner (1993) reports that patient groups are more likely to be skewed outside of this normative balance.

The configuration of the two kinds of data—one reflecting responsiveness to affective stimuli and the other predicting how well structured the person's responses will be—can be of value in describing his or her overall approach to emotional experience. For example, we could speculate that a person who is overresponsive to emotionally charged situations (high Affective Ratio) and who does not structure this material well (*CF* plus *C* greater than *FC*) would be likely to manifest relatively frequent episodes of poorly modulated emotional discharge.

What is the Quality of Interpersonal Function?

Several Rorschach variables are helpful in describing how an individual is likely to operate in the interpersonal world. Most people—patients and non-patients alike—tend to give between four and six human-content percepts in the course of the 10 inkblots. Absence of human content is normatively unexpected and apt to be associated with significant interpersonal difficulties. Most human content involves whole humans. Percepts with human parts or mythical humans are less frequent. They may suggest an inaccurate understanding of the interpersonal world.

Another Rorschach variable that has interpersonal implications is the aggressive movement (*AG*) response. Summarizing a series of studies with this variable, Exner (1986a) concludes that

elevations in aggressive movement responses "signify an increased likelihood for aggressive behaviors, either verbal or nonverbal, and... also indicate attitudes toward others that are more negative and/or hostile than is customary" (p. 405).

A composite variable developed by Exner (1986a), the Isolation Index, appears useful in its ability to identify individuals whose social network is tenuous and who are interpersonally isolated. Presence of an elevated Isolation Index in the context of a record that has other suggestions of interpersonal withdrawal—low Affective Ratio, absence of texture percepts, low number of whole humans—may describe an individual whose interpersonal competence and interests are significantly limited.

The hypervigilance index (Exner, 1987) was developed by analyzing Rorschach data to see if it could discriminate a subset of patients described by their therapists as avoiding close interpersonal relationships. Individuals who are positive on this index devote a substantial amount of time to an apprehensive scanning of the interpersonal field, and their sense of pessimism and distrust concerning the motives of others is noteworthy.

A variable Exner (1988) named cooperative projection (*COP*) can be coded for human or animal movement percepts in which there is a clearly cooperative relationship. *COP* percepts appear at least once in 79 percent of a sample of adult nonpatient Rorschachs but in only 38 percent of adult character-disorder records (Exner, 1993). Although research is still preliminary, *COP* appears to be a stable variable that is associated with positive rankings by peers and possibly with favorable treatment outcome.

FUTURE DIRECTIONS

The kinds of questions for which the Rorschach can provide data are very much those the practicing clinician faces on a daily basis. A study by Ritzler and Alter indicated that the test is now taught in 93 percent of APA-approved graduate clinical psychology programs in the United States and Canada, and an increasing number of clinicians are seeking continuing-education training (Exner, 1988). To a very great extent, it has been the availability of an ongoing research base that has sparked the Rorschach's renaissance, and a substantial amount of work is currently in progress. Five major research areas are of particular importance.

Discriminant function and logistic regression techniques that differentiate the Rorschachs of externally identified individuals—be they hypervigilant, perfectionistic, or violent—bring a powerful methodology to the instrument. This configural technique takes advantage of computer technology and allows very extensive utilization of the Rorschach's data yield.

The collection of normative data on a variety of clinically relevant groups is an equally important area on the research frontier. Availability of an increasing number of skilled Rorschach clinicians will undoubtedly lead to the accumulation of important reference data over the next years. The work of Gacono and Meloy (1994), for example, has provided comprehensive descriptions of the Rorschach characteristics of several groups of aggressive and psychopathic individuals.

Advances in computer and psychophysiological technology now make possible increasingly sophisticated basic science research that will shed greater light on the way that Rorschach responses are generated. A series of studies (Exner, 1978) suggested that the blots are scanned very quickly and that the person processes substantially more data than he or she reports. What happens in the interim between the person's scanning of the blot and the articulation of a response appears to involve a very complex process that is the focus of substantial current research. It is likely that advances in brain-imaging technology will allow even more sophisticated understanding of the processes underlying Rorschach-response production.

A fourth significant research area takes advantage of the fact that the Rorschach is typically utilized as part of a test battery. The study of its interaction with other personality instruments, and with cognitive and neuropsychological techniques, will be of substantial importance to practicing clinicians. Ganellen (1996), as an example, has provided an extensive summary describing both research and clinical integration of the Rorschach and the MMPI-2.

Finally, the Rorschach's potential as a rich source of content data continues to be an important area of study. A variety of workers are creating approaches that link Rorschach content material with constructs generated from formal personality theory. Examples include the work of Lerner (1991) and Cooper, Perry, and Arnow (1988).

SUMMARY

Perhaps the survival of Rorschach's deceptively simple technique can be traced to its extraordinarily comprehensive ability to tap the complexity of human psychological operation. It has been this very richness of data that sometimes made for controversy as various workers argued over how best to conceptualize and process the test's varied yield. Those years of divergence seem to be over. The integrative stance that currently characterizes Rorschach research and practice has turned the controversies into alternative and complementary approaches to an instrument whose breadth we are still charting after well over half a century.

REFERENCES

Arffa, S. M. (1982). Predicting adolescent suicidal behavior and the order of Rorschach measurement. *Journal of Personality Assessment, 46*, 563–568.

Beck, S. J. (1937). *Introduction to the Rorschach Method: A manual of personality study*. American Orthopsychiatric Association Monograph, No. 1.

Beck, S. J. (1944). *Rorschach's test. I: Basic processes*. New York: Grune & Stratton.

Beck, S. J. (1945). *Rorschach's test. II: A variety of personality pictures*. New York: Grune & Stratton.

Beck, S. J. (1952). *Rorschach's test. III: Advances in interpretation*. New York: Grune & Stratton.

Beck, S. J. (1960). *The Rorschach experiment: Ventures in blind diagnosis*. New York: Grune & Stratton.

Beck, S. J., Beck, A. G., Levitt, E., & Molish, H. B. (1961). *Rorschach's test. I: Basic processes* (3rd ed.). New York: Grune & Stratton.

Beck, S. J., & Molish, H. B. (Eds.). (1967). *Rorschach's test. II: A variety of personality pictures*. New York: Grune & Stratton.

Cooper, S., Perry, J., & Arnow, D. (1988). An empirical approach to the study of defense mechanisms. I. Reliability and preliminary validity of the Rorschach defense scale. *Journal of Personality Assessment, 52*, 187–203.

Exner, J. E. (1973). The Self Focus Sentence Completion: A study of egocentricity. *Journal of Personality Assessment, 37*, 437–455.

Exner, J. E. (1974). *The Rorschach: A comprehensive system. Vol. 1*. New York: Wiley.

Exner, J. E. (1978). *The Rorschach: A comprehensive system. Vol. 2: Current research and advanced interpretation*. New York: Wiley.

Exner, J. E. (1986a). *The Rorschach: A comprehensive system, Vol. 1* (2nd ed.): *Basic foundations*. New York: Wiley.

Exner, J. E. (1986b). Some Rorschach data comparing schizophrenics with borderline and schizotypal personality disorders. *Journal of Personality Assessment, 50*, 455–471.

Exner, J. E. (1987). *Alumni newsletter*. Asheville, NC: Rorschach Workshops.

Exner, J. E. (1988). *Alumni newsletter*. Asheville, NC: Rorschach Workshops.

Exner, J. E. (1991). *The Rorschach: A comprehensive system, Vol. 2* (2nd ed.): *Interpretation*. New York: Wiley.

Exner, J. E. (1993). *The Rorschach: A comprehensive system, Vol. 1* (3rd ed.): *Basic foundations*. New York: Wiley.

Exner, J. E., & Murillo, L. G. (1975). Early prediction of post-hospitalization relapse. *Journal of Psychiatric Research, 12*, 231–237.

Exner, J. E., Murillo, L. G., & Cannavo, F. (1973). Disagreement between patient and relative behavioral reports as related to relapse in nonschizophrenic patients. Washington DC: Eastern Psychological Association.

Exner, J. E., & Weiner, I. B. (1982). *The Rorschach: A comprehensive system, Vol. 3: Assessment of children and adolescents*. New York: Wiley.

Exner, J. E., & Weiner, I. B. (1995). *The Rorschach: A comprehensive system, Vol. 3* (2nd ed.): *Assessment of children and adolescents*. New York: Wiley.

Exner, J. E., & Wylie, J. R. (1977). Some Rorschach data concerning suicide. *Journal of Personality Assessment, 41*, 339–348.

Frank, L. K. (1939). Projective methods for the study of personality. *Journal of Psychology, 8*, 343–389.

Gacono, C. B., & Meloy, J. R. (1994). *The Rorschach Assessment of Aggressive and Psychopathic Personalities*. Hillsdale, NJ: LEA.

Ganellen, R. J. (1996). *Integrating the Rorschach and the MMPI-2 in Personality Assessment*. Mahwah, NJ: LEA.

Gill, H. S. (1966). Delay of response and reaction to color on the Rorschach. *Journal of Projective Techniques and Personality Assessment, 30*, 545–552.

Harder, D. W., & Ritzler, B. A. (1979). A comparison of Rorschach developmental level and form-level systems as indicators of psychosis. *Journal of Personality Assessment, 43*, 347–354.

Hens, S. (1917). *Szynon phantasieprufung mit formlosen klecksen be, schulkindern, normalen erwach-*

seuen und genteskranken. Unpublished dissertation. Zurich.

Kerner, J. (1857). Klexographien. In R. Pissen (Ed.), *Kerners werke.* Berlin: Boag and Co.

Kinder, B., Brubaker, R., Ingram, R., & Reading, E. (1982). Rorschach form quality: A comparison of the Exner and Beck systems. *Journal of Personality Assessment, 46,* 131–138.

Klopfer, B., Ainsworth, M. D., Klopfer, W. G., & Holt, R. R. (1954). *Developments in the Rorschach technique, Vol. I. Technique and theory.* Yonkers: World Book.

Klopfer, B., Ainsworth, M. D., Klopfer, W. G ., & Holt, R. R. (Eds.). (1956). *Developments in the Rorschach technique, Vol. II. Fields of application.* Yonkers: World Book.

Klopfer, B., & Kelley, D. (1942). *The Rorschach technique.* Yonkers: World Book.

Klopfer, B., Meyer, M. M., & Brawer, F. (Eds.). (1970). *Developments in the Rorschach technique. Vol. III, Aspects of personality structure.* New York: Harcourt Brace, Jovanovich.

Klopfer, B., & Sender, S. (1936). A system of refined scoring symbols. *Rorschach Research Exchange, 1,* 19–22.

Krugman, M. (1940). Out of the inkwell. *Rorschach Research Exchange, 4,* 91–101.

Lerner, P. (1991). Psychoanalytic theory and the Rorschach. Hillsdale, NJ: Analytic Press.

Mayman, M. (1970). Reality contact, defense effectiveness, and psychopathology in Rorschach form-level scores. In B. Klopfer (Ed.), *Developments in the Rorschach technique, III.* New York: Harcourt Brace, Jovanovich.

Meili-Dworetzki, G. (1956). The development of perception in the Rorschach. In B. Klopfer (Ed.), *Developments in the Rorschach technique, II.* Yonkers, NY: World Book.

Molish, H. B. (1967). Critique and problems of the Rorschach. A survey. In S. J. Beck & H. B. Molish (Eds.), *Rorschach's Test, Vol. II.* New York: Grune & Stratton.

Piotrowski, Z. (1957). *Perceptanalysis.* New York: Macmillan.

Rapaport, D., Gill, M., & Schafer, R. (1945, 1946). *Diagnostic psychological testing.* Chicago: Yearbook Publishers.

Rapaport, D., Gill, M., & Schafer, R. (1968). *Diagnostic psychological testing* (rev. ed.). R. R. Holt (Ed.). New York: International Universities Press.

Roemer, G. (1967). The Rorschach and Roemer symbol test series. *Journal of Nervous and Mental Disorders, 144,* 185–197.

Rorschach, H. (1921). *Psychodiagnostik.* Bern: Bircher. (English translation, Bern: Hans Huber, 1942).

Rorschach, H., & Oberholzer, E. (1923). The application of the form interpretation test. *Zeitschrift fur die Gesamte Neurologie und Psychiatrie,* 82. Also in H. Rorschach (1942). *Psychodiagnostik.* Bern: Hans Huber).

Rybakow, T. (1910). *Atlas for experimental research on personality.* Moscow: University of Moscow.

Schafer, R. (1948). *The clinical application of psychological tests.* New York: International Universities Press.

Schafer, R. (1954). *Psychoanalytic interpretation in Rorschach testing.* New York: Grune & Stratton.

Schafer, R. (1967). *Projective testing and psychoanalysis.* New York: International Universities Press.

Singer, J. L., & Brown, S. L. (1977). The experience type: Some behavioral correlates and theoretical implications. In M. A. Rickers-Ovsiankina (Ed.), *Rorschach psychology.* Huntington, NY: Krieger.

Shalit, B. (1965). Effects of environmental stimulation of the *M, FM,* and *m* responses in the Rorschach. *Journal of Projective Techniques and Personality Assessment, 29,* 228-231.

Smith, N. M. (1981). The relationship between the Rorschach whole response and level of cognitive functioning. *Journal of Personality Assessment, 45,* 13–19.

Tulchin, S. H. (1940). The pre-Rorschach use of inkblot tests. *Rorschach Research Exchange, 4,* 1–7.

Weiner, I. B. (1977). Approaches to Rorschach validation. In M. A. Rickers-Ovsiankina (Ed.), *Rorschach psychology.* Huntington, NY: Krieger.

Whipple, G. M. (1910). *Manual of mental and physical tests.* Baltimore: Warwick and York.

Wilson, G., & Blake, R. (1950). A methodological problem in Beck's organizational concept. *Journal of Consulting Psychology, 14,* 20–24.

Zubin, J., Eron, L. D., & Schumer, F. (1965). *An experimental approach to projective techniques.* New York: Wiley.

This Page Intentionally Left Blank

PART VIII

BEHAVIORAL ASSESSMENT

This Page Intentionally Left Blank

CHAPTER 18

BEHAVIORAL ASSESSMENT OF CHILDREN

Ross W. Greene
Thomas H. Ollendick

INTRODUCTION

As Mash and Terdal (1988) have noted, assessment and classification of children begins virtually at the point of conception and continues throughout childhood, and may take various forms depending on one's orientation, emphasis, and goals. A "behavioral assessment" is, presumably, one emanating from a "behavioral" orientation. Yet, notions as to what constitutes a behavioral orientation have become increasingly varied and blurred, and the precise components—the *who*, *what*, and *how*—of behavioral assessments of children have, in turn, become less definitive. From our perspective, behavioral assessment of children can be characterized by several critical features.

First, behavioral assessments are guided by a broad *social learning theory framework*. This orientation, which is described more fully in the pages that follow, is distinguished by its empirical, databased approach to the study of persons. Consistent with this emphasis, behavioral assessments rely predominantly upon *objective* data and minimally upon subjective data or high levels of inference. In this respect, a behavioral assessment differs from a "dynamic formulation"; while both might involve hypotheses regarding mechanisms underlying a child's patterns of thought and behavior, the latter is typically characterized by psychodynamic concepts for which confirmatory objective data are scarce. The emphasis on objectivity also necessitates consideration of *developmental* and *cultural norms*, and has ramifications for the selection of assessment procedures and for the types of conclusions one may draw from the information obtained through the assessment process.

A social learning orientation has implications for the *what* of assessment. While "objective measurement" has previously referred only to the direct observation of highly discrete overt behaviors and the identification of their controlling variables (in the behavior analytic tradition), from our point of view assessments adhering to this orientation tend to be unnecessarily limited in scope and utility. From a social learning perspective, behavioral assessments of children focus upon both *overt* and *covert* behaviors, with the latter referring to affective states and various cognitive processes (e.g., expectancies, perceptions, beliefs, subjective values) that may exert considerable influence on overt behavior. As such, the *how* of objective measurement refers to a broad array of assessment instruments and practices tapping the overt and covert domains and possessing *satisfactory psychometric properties*. (Due to space limitations, we do not provide a comprehensive discussion of psychometric issues in this chapter; interested readers are referred to Anastasi, 1982, and Breen, Eckert, & DuPaul, 1996, for discussion of these issues.)

Second, we believe behavioral assessments are distinguished by their *breadth* and *comprehensiveness*. In this respect, a behavioral assessment differs from a diagnostic assessment, the primary goal of which is to arrive at diagnostic categorizations. While diagnostic assessments provide important data about a broad range of behavioral categories, they frequently rely on a limited number of reporters (often only one) for assessment information and ordinarily provide little highly specific information about *contexts* which may contribute to variations in a child's behavior. This emphasis on breadth and comprehensiveness stems from an understanding that situational factors exert a powerful influence on the frequency, intensity, and duration of a given behavior. Thus, the *who* and *what* of behavioral assessments must extend well beyond the overt and covert behaviors of an identified child to encompass the multiple persons (e.g., parents, teachers, siblings, peers) who interact (and have interacted previously) with the child and the multiple settings (e.g., various contexts within the home, school, and other environments) in which these interactions occur (or have occurred). In this respect, behavioral assessments are also consistent with the themes of developmental psychopathology (e.g., Cicchetti, 1984, 1993; Rutter & Garmezy, 1983) and with theoretical models of development emphasizing reciprocal influences on behavior (e.g., Sameroff & Chandler, 1975).

The emphasis on breadth and comprehensiveness—along with limitations inherent in different assessment procedures—also has implications for the *how* of behavioral assessments. Because no single procedure or reporter is viewed as sufficient to provide a comprehensive understanding of a child, behavioral assessments are *multi-modal*, meaning that multiple reporters and measurement procedures are employed. In general, we might expect assessment conclusions to be more definitive if there is stability in a child's behavior across time and contexts and general agreement across reporters and assessment procedures. In instances of inconsistency—for example, certain behaviors appear to be more frequent and intense in interactions within a particular environment—the evaluation process must focus on factors that may account for this inconsistency. Unless these inconsistencies are clarified, targets and procedures for intervention will remain unclear. Pinpointing situations in which target behaviors are exacerbated or improved helps identify the persons who, and contexts that, should be targeted for intervention and thereby allows more educated and fine-tuned selection of intervention options.

Third, a behavioral assessment is best conceptualized as a *fluid, exploratory, hypothesis-testing process,* the goal of which is to *understand* a given child, group, family, and/or social ecology, often for the purpose of formulating and evaluating specific intervention strategies (Ollendick & Greene, 1990; Ollendick & Hersen, 1984). In this respect a behavioral assessment differs from, for example, a one-time mental status examination, in which the goal is to achieve an efficient summary of a child's mental state, often for the purpose of determining whether the child requires immediate psychiatric hospitalization. Mental status examinations clearly serve a critical function and provide extremely useful information; however, the time-limited nature of such evaluations often forces overly definitive, concrete, narrow conclusions driven by necessity of categorization (e.g., "this child is not psychotic and reports no suicidal or homicidal ideation or intent; while his behavior is of concern and may signal an emerging bipolar illness, he does not require hospitalization at this time"). The time constraints involved in a mental status examination may also encourage the invocation of fairly amorphous explanations and recommendations (e.g., "clearly the stress caused by the acrimonious relationship between this child's mother and father is impacting upon his already tenuous emotional state; he and his parents have been referred to a mental health clinic in their community for consultation").

Inherent in a behavioral assessment is the understanding that a definitive understanding of a child's overt and covert behavior is extremely difficult to achieve. As such, initial conclusions (e.g., "the stress caused by the acrimonious relationship between this child's mother and father is impacting upon his already tenuous emotional state...[and] emerging bipolar illness") are understood as *hypotheses* that await verification or revision based on additional information, with pursuit of additional information viewed as an ongoing process. This process continues even during intervention, which, aside from being an attempt to ameliorate a child's difficulties, can be understood as an opportunity to obtain additional information based on the child's response to treatment. Thus, child behavioral assessments de-emphasize quick, definitive conclusions and focus more on obtaining information that is directly relevant to treatment. "Relevance for treatment" refers to the clinical utility of information in pinpointing treat-

ment goals, the selection of targets for intervention, the design and implementation of interventions, and the evaluation of intervention outcomes (Greene, 1995; Mash & Terdal, 1988). A related concept—treatment utility—refers to the degree to which assessment strategies are shown to contribute to beneficial treatment outcomes (Hayes, Nelson, & Jarrett, 1987). For example, concluding that a child "has" a given disorder typically provides little useful information about the contexts in which different behaviors associated with the disorder occur and are most problematic, and may be of only minimal assistance in formulating an initial approach to intervention.

In sum, the behavioral assessment of children can be understood as *a fluid, exploratory, hypothesis-testing process—guided by social learning principles—in which a range of specific procedures is used in order to achieve a broad, comprehensive understanding of a given child, group, and social ecology, and to formulate and evaluate specific intervention strategies.* It may be obvious that this process has evolved in response to various theoretical and technological advances. We believe it is helpful to understand this evolution in a historical context. Thus, before expanding upon the various assessment components delineated above, a brief historical overview may be useful.

HISTORY OF BEHAVIORAL ASSESSMENT

In a very direct way, the history and evolution of behavioral assessment closely parallels the history and evolution of behaviorism. It is fair to state that behaviorism evolved at least in part as a reaction against psychoanalytic theory, which emphasized concepts perceived by behaviorists as subjective, unobservable and, therefore, unscientific. One of the goals of the early behaviorists was to elucidate a philosophy of understanding and studying persons driven by the objective procedures that typified the "harder" sciences. From a behavioral perspective, psychoanalytic concepts, such as drives, conflicts, psychosexual stages of development, unconscious defense mechanisms, and the like, could not be measured objectively and had to be *inferred* from a person's self-report or behavior. Thus, such concepts had no place in a scientific theory of human behavior and could not be invoked to explain mechanisms controlling a given behavior. Early behaviorists believed that only directly observable behaviors were worthy of study (Skinner, 1953). Inference was eschewed. A person's self-report was viewed not only as unreliable but also as uninterpretable, since interpretation was viewed as the epitome of subjectivity. Direct observation of behavior was—and in many respects, still is—the hallmark of behavioral assessment. Thus, rather than interpreting behaviors—for example, a child who chronically hits other children as being "hostile" or "enraged"—behaviorists would instead strive to operationalize or define hitting behavior in objective terms (e.g., "punching," "smacking," "kicking," "scratching," "eye-poking," etc.) and embark on an assessment process that involved direct measurement of the frequency, intensity, and duration of these behaviors under varying conditions or situations so that the "function" of the behavior could be understood, predicted, and subsequently altered.

By emphasizing science, behaviorists also spurred much research examining the normative developmental course of children's behavior. Whereas various interpretations had previously been applied to certain childhood behaviors (e.g., hitting is a sign of internal rage, bedwetting is a primitive attempt to extinguish this internal rage, clingy behavior is indicative of enmeshment, and so forth), researchers began to demonstrate that childhood behaviors once considered "deviant" were actually fairly normative at certain ages (e.g., it is not developmentally unusual for two year olds to be impulsive and inflexible, three year olds to void in their beds at night, five year-olds to reverse letters, and 12 year olds to be anxious about their physical appearance). Thus, the increased emphasis upon determining whether a particular behavior or pattern of behavior was *developmentally deviant* was by no means coincidental (see Campbell, 1989, for further discussion of this issue).

Behaviorists also refuted notions of "personality" as consisting of stable and enduring traits and of behavior as consistent across situations and over time (we place "personality" in quotations because early behaviorists would have objected to use of this term, given its ambiguous parameters). As noted above, the behavioral view has instead emphasized the *situational* nature of behavior. Let us return again to the child who hits other children. A behaviorist would assume that the child's acts of aggression do not occur at all times but rather under certain conditions and settings and would therefore attempt to identify those variables that

elicit and maintain aggressive acts in those situations. In other words, the assessment process would extend well beyond the clinic and would document the child's behavior at various points in time at home, school, the playground, and so forth. Recall that in behavioral assessment the goal is not to make sweeping generalizations about a child's behavior but rather to reach a clear understanding of the conditions under which certain specific behaviors occur. Restated, the focus of assessment is on what the child *does* in a given *situation* rather than on what he or she *has* or *is* (Mischel, 1968).

Not coincidentally, these emphases on directly observable stimuli and situational factors led to the conclusion that behavior occurs by virtue of *conditioning* (i.e., *learning*, *experience*). By manipulating environmental conditions, behaviorists were able to demonstrate convincingly the manner in which behavior could be shaped, elicited, maintained, and eliminated. Briefly, in *classical conditioning* a neutral stimulus (one that elicits no particular response) is repeatedly presented along with a stimulus that reflexively elicits a particular response, with the neutral stimulus alone eventually eliciting the same response. In *operant conditioning* behavior is governed by its consequences; behaviors that are rewarded are more likely to be repeated whereas behaviors that are punished are less likely to be repeated.

Literally thousands of published scientific studies have offered compelling evidence for the role of classical and operant conditioning as influences on the behavior of humans and other animals. Nonetheless, driven by the influential work of Rotter (e.g., 1954, 1966, 1972), Bandura (e.g., 1971, 1973, 1986), Mischel (e.g., 1968, 1973, 1979), and others, behaviorists began to recognize that other factors influenced human behavior in addition to operant and classical conditioning. These *cognitive behaviorists* or *social learning theorists* (we use these terms interchangeably in this chapter) extended the work of early behaviorists in a variety of ways, perhaps most significantly by introducing the notion that cognitions and affective states exert significant influence upon learning and behavior, and by proposing that much learning occurs indirectly by observing others (vicarious learning) rather than directly through classical and operant conditioning.

True to their scientific roots, social learning theorists proposed and attempted to study specific categories of cognition and the manner in which these cognitions influence learning and human behavior. Mischel (1973, 1984) dubbed these categories *cognitive social learning person variables* (CSLPVs) and, because of their significance for behavioral assessment, they are worthy of brief overview here. *Competencies* refers to a person's actual skills and capacities, be it intelligence, reading skills, social skills, problem-solving skills, and so on; *encoding strategies* refer to the manner in which a person interprets and perceives themselves, others, and events; *expectancies* are a person's expectations of the likely outcomes of a particular behavior or stimulus and the degree to which a person believes he or she can actually perform a particular behavior; *subjective values* refers to a person's preferences and aversions, likes and dislikes, and so on (i.e., what stimuli are rewarding to the person and what stimuli are punishing); and *self-regulatory systems and plans* refers to a person's capacity for and manner of self-imposing goals and standards and self-administering consequences.

How might these social learning person variables be of value in the assessment of a child who hits other children? Were we to assess the child's competencies, we might, for example, consider the child's capacities or skills for formulating behavioral alternatives (besides hitting) for dealing with frustration, and modulating emotions associated with frustration. The child's encoding strategies might also be relevant to assessment: Do the child's overly hostile attributions about actions of other children lead to frequent aggressive retribution? Do the child's perceptions of the impact of his or her hitting behavior on others differ from that which would be considered normative and adaptive? Also relevant are the child's expectancies: Does the child believe that negotiating difficulties with other children will lead to advantageous outcomes? Does the child believe he or she could successfully perform such negotiations in an efficacious manner? Were we to assess subjective values we might find that the child finds pummeling other children to be rewarding or may view this approach as optimal for resolving difficulties with others. Finally, we might find the child to have a severely limited capacity for self-regulation; while he or she is able to generate alternatives for dealing with frustration, believes that such alternatives would lead to advantageous outcomes, values these advantageous outcomes, and is remorseful over having strayed from these values, he or she may lack the capacity to inhibit the response of hitting under frustrating conditions

(e.g., Barkley, 1994). Thus, information about CSLPVs might lead to a more comprehensive conceptualization of "the problem," might identify situations in which "the problem" is more or less likely to occur, and might result in more productive interventions than, for example, merely punishing hitting and rewarding more adaptive behaviors. As discussed below, consideration of social learning person variables may also lead to an improved understanding of the persons with whom a given child interacts. Consistent with the notion that a child's behavior is situational, assessment of competencies, encoding strategies, expectancies, subjective values, and self-regulatory systems and plans of significant persons in the child's life is critical.

As we have already noted, social learning theory also proposes that much learning occurs through observational or vicarious processes rather than solely through classical and operant conditioning. In other words, hitting may have entered a child's repertoire by observing others hitting (e.g., at home, on television) rather than solely by the child having been reinforced for hitting. We may also learn by *hearing about* others' behavior or outcomes, or by reading about them. Thus, there are virtually limitless opportunities and situations through which behaviors may enter a child's repertoire. Therefore, behavioral assessment of a child must extend into important environmental domains in which the child has had the opportunity to observe, hear about, and learn such behaviors. This may include large-scale social systems such as schools and neighborhoods (Patterson, 1976; Wahler, 1976) which have been shown to have immediate and profound effects on individual behavior (see Winett, Riley, King, & Altman, 1989, for a more comprehensive discussion of this issue). Although inclusion of these additional factors serves to complicate the assessment process, they are indispensable to the goal of achieving a broad, comprehensive, and clinically useful assessment of a child.

In sum, child behavioral assessment has evolved from sole reliance on measurement of directly observable target behaviors into a broader approach that takes into account cognitive and affective processes, developmental issues, and social contexts that contribute to variations in a child's behavior. Needless to say, multi-method, multi-context behavioral assessment of children entails use of a wide range of specific procedures. Let us now turn to an overview of how these procedures connect with the various *components* of a behavioral assessment reviewed in these introductory sections. The overview is not organized sequentially but rather by assessment domains: *overt behavior, covert behavior,* and *contexts.* Procedures used in the assessment of a child's overt behavior tend to focus primarily on the issue of "What does the child do, and when?" while procedures utilized in assessment of covert behavior focus on issues of "What does the child think and feel, and when?" and "How do these thoughts and feelings connect with the child's overt behavior?" Finally, procedures used in the assessment of contexts center on issues of "How do overt and covert behaviors of adult caretakers and characteristics of environments contribute to variations in a child's overt and covert behavior?" Given space limitations, we have chosen not to provide an exhaustive overview of assessment instruments; instead, we provide examples of various measures and procedures which may assist in achieving each component.

ASSESSMENT OF CHILDREN'S OVERT BEHAVIOR

Various procedures may be employed in attempting to gather information about what overt behaviors a child exhibits and when (under what conditions) these behaviors occur. Cone (1978) has distinguished between direct and indirect methods of assessing a child's overt behavior; in *direct* assessment, the behaviors of interest are assessed *at the time and place of their occurrence,* whereas *interviews* with the child or other reporters, *self-reports* by the child, and *behavior ratings* by the child or other reporters are considered more *indirect* means of obtaining assessment information. We discuss each of these methods and their relative advantages and limitations below.

Direct Behavioral Observation

Direct observation involves the formal or informal observation of a child's overt behavior in various natural contexts, including home (e.g., during homework, dinner, bedtime, etc.), school (e.g., during recess, lunch, group discussions, independent work, etc.), and other domains (little league, friends' homes, etc.). In addition to providing a first-hand view of behaviors of interest, such

observations also afford an opportunity to observe situational factors which contribute to variations in a child's behavior. Johnson and Bolstad (1973) have characterized the development of naturalistic observation procedures as the major contribution of the behavioral approach to assessment and treatment of children. While this point is arguable, it is clear that direct observation involves the least inference of the assessment methods we will describe and, from our perspective, remains an indispensable component of a behavioral assessment of a child. Unfortunately, philosophical indispensability does not always translate into practice, and this may be particularly true of naturalistic observation procedures, which appear to be decreasing in popularity (Hops, Davis, & Longoria, 1995). Direct observations tend to be more time-consuming and inconvenient as compared to other methods of assessment, and there is no guarantee that target behaviors will actually occur during designated observation periods. Thus, in some cases it may be necessary to have significant others in the child's environment (e.g., parents, teachers) formally observe and record the child's behavior, to have children observe and record their own behavior, or to employ analogue procedures in which the goal is to observe the child in a simulated laboratory setting that closely resemble the actual setting(s) of interest.

In many instances, it may be desirable to formalize the direct observation process, especially if the observer wishes to quantify behaviors for the purpose of normative comparisons. In such instances, a measure such as the Child Behavior Checklist Direct Observation Form (CBC-DOF; Achenbach, 1986) may be useful. This measure consists of 96 items covering a broad range of children's behavior. The form is completed after observing a child in a given setting for 10 minutes after a recommended three to six separate occasions. Scores can be obtained on six factor-analytically-derived scales, including withdrawn-inattentive, hyperactive, nervous-obsessive, depressed, attention-demanding, and aggressive. An advantage of the CBC-DOF is its sensitivity to developmental issues (normative information based on a child's age and gender is available) and sound psychometric properties (McConaughy, Achenbach, & Gent, 1988).

In other cases, observers may wish to quantify a set of highly specific behaviors. An example of a system designed for such purposes is the Classroom Observation Code developed by Abikoff, Gittelman-Klein, and Klein (1977) and modified by Abikoff and Gittelman (1985), which is used to record behaviors associated with poor self-regulation in school settings. The system involves the systematic recording of various categories of behavior reflective of poor self-regulation, including (but not limited to) *interference* (verbal or physical behaviors that are disturbing to others), *off-task* (attending to stimuli other than the assigned work), *noncompliance* (failure to follow teacher instructions), *minor motor movement* (e.g., restlessness and fidgeting), *gross motor behavior* (e.g., leaving seat and/or engaging in vigorous motor activity), *physical aggression* (physical aggression directed at another person or destruction of others' property), and *solicitation of teacher* (e.g., raising hand, calling out to teacher).

In yet other cases, assessors may wish to *simultaneously* record the behavior of children and the adults with whom they are interacting for the purpose of capturing the reciprocal nature of adult/child interactions. Because such recording systems tend to be labor-intensive, their use is often restricted to research. However, one cannot overstate the conceptual appeal of conducting observations that attend simultaneously to child and adult behavior as so as to obtain information about the manner in which the behavior of each may lead to variations in the behavior of the other, even if such observations occur informally and in unstructured contexts. Examples of formal codings systems for the simultaneous recording of adult and child behavior are the Code for Instructional Structure and Student Academic Response (CISSAR; Stanley & Greenwood, 1981; Greenwood, Delquadri, Stanley, Terry, & Hall, 1985), which is used to record student-teacher interactions; and the Response Class Matrix (Barkley, 1981; Mash & Barkley, 1987), the Dyadic Parent-Child Interaction Coding System II (DPICS; Eyberg, Bessmer, Newcomb, Edward, & Robinson, 1994), and the Living in Family Environments system (LIFE; Hops, Biglan, Tolman, Sherman, Arthur, et al., 1990), which are used to record parent-child interactions.

As noted above, direct observation can also be accomplished via laboratory or analogue settings that are similar to, but removed from, the natural environment. Simulated observations are especially helpful when a target behavior is of low frequency, when the target behavior is not observed in the naturalistic setting due to reactivity effects (i.e., the child does not exhibit the target behavior

because he or she is aware of being observed), or when the target behavior is difficult to observe in the natural environment due to practical constraints. An example of such a simulated observation system is the Restricted Academic Situation (RAS) described by Barkley (1990), which has been used in the analogue assessment of poorly self-regulated children. In brief, the RAS involves placing a child in a playroom containing toys and a small worktable; the child is instructed to complete a packet of mathematics problems and is then observed for 15 to 20 minutes from behind a two-way mirror. Behavior coding occurs during this period, using clearly defined categories such as off-task, fidgeting, vocalizing, playing with objects, and out of seat. A major disadvantage of this system is its lack of normative data. The degree to behaviors observed during this and other analogue procedures generalizes to natural settings remains an important concern.

Behavior Ratings and Checklists

Behavior checklists are perhaps the most popular indirect means of gathering information about a child's overt behavior, and this popularity is presumably a function of the efficiency of such checklists. In their most common form, behavior checklists require adults to rate various items to indicate the degree to which a child exhibits certain behaviors. In general, checklists are useful in providing an overall description of a child's behavior, in specifying dimensions or response clusters that characterize the child's behavior, and in serving as outcome measures for the effectiveness of treatment. To reiterate, this method of assessment is considered *indirect* because it relies on retrospective descriptions of the child's behavior (by reporters whose impartiality is uncertain). Nonetheless, behavior checklists can provide a comprehensive and cost-effective picture of the child and his or her overall level of functioning and may be useful in eliciting information that may have been missed by other assessment procedures (Novick, Rosenfeld, Bloch, & Dawson, 1966). The better checklists—and obviously, those we view as most appropriate for an empirically driven assessment—undergo rigorous evaluations of reliability and validity and provide developmental norms. As with direct observation, assessors must be sensitive to potential biases of persons completing and interpreting checklists. Normative information helps protect against some forms of subjectivity, but not completely so. Behavior checklists cannot be viewed as a satisfactory replacement for direct observation; rather the two methods of assessment are best understood as complementary.

Among the most widely used of the "omnibus" checklists—those assessing a broad range of behaviors—are the Child Behavior Checklist (CBCL; Achenbach, 1991a), which is completed by parents, and the Child Behavior Checklist Teacher Report Form (CBC-TRF; Achenbach, 1991b), which is completed by teachers. The parent-completed CBCL is available in two formats depending on the age of the child being rated (i.e., two to three years and four to 18 years; we review only the latter here). The CBCL/4-18 consists of 112 items rated on a 3-point scale. Scored items can be clustered into three-factor analyzed profiles: social competence, adaptive functioning, and syndrome scales. The latter includes scales such as withdrawn, anxious/depressed, social problems, attention problems, delinquent behavior, and aggressive behavior. Social competency items examine the child's participation in various activities (e.g., sports, hobbies, chores) and social organizations (e.g., clubs, groups), and school (e.g., grades, placement, promotions). The teacher-completed CBC-TRF also consists of 112 items which are fairly similar, but not completely identical to, those found in the CBCL. The scored items from the CBC-TRF cluster into the same three-factor analyzed profiles. Both checklists have been extensively researched, provide detailed normative information, and have exceptional psychometric properties. Some of the items on the CBC-TRF and CBCL refer to directly observable behaviors (e.g., physically attacks people, bites fingernails, gets teased a lot) whereas others refer to cognitions and affective states (e.g., fears he or she might think or do something bad, likes to be alone, feels too guilty, feels worthless or inferior). As such, while we have placed our discussion of these checklists in this section on assessment of *overt* behavior, we wish to emphasize that they also provide information about select *covert* processes.

However, one should not assume that the excellent psychometric properties of the CBCL and CBC-TRF protect these instruments from the idiosyncratic perceptions and biases of raters. Many items are not clearly defined and do, in fact, require subjective judgments (e.g., acts too young for his or her age, demands a lot of attention, showing off or clowning,

too fearful or anxious) and therefore may contribute to variability in ratings by different adults (see Achenbach, McConaughy, & Howell, 1987). Thus, variability in responses of different raters may reflect not only the influence of contextual factors (i.e., the child's behavior varies depending on characteristics of different situations) but also the unique interpretations, expectations, and tolerances of different adult raters (Greene, 1995, 1996). Unlike direct naturalistic observation, behavior checklists are one or more steps removed from the behaviors themselves, and it may be difficult to clarify responses and explore potential response biases of raters.

In some instances, it may be useful to obtain more detailed information about a specific domain of behavior—such as social skills or self-regulation—than that provided by omnibus behavior checklists. Many "narrow band" checklists have been developed and may be useful in this regard. For example, the Social Skills Rating System (SSRS; Gresham & Elliott, 1990) is a 55-item questionnaire which provides information about a child's social behavior in three domains (social skills, problem behaviors, and academic competence). The SSRS includes parent-, teacher-, and self-rated formats. Numerous checklists are available for obtaining ratings of a child's self-regulation, including the ADD-H Comprehensive Teacher Rating Scale (ACTeRS; Ullman, Sleator, & Sprague, 1984, 1991), an instrument consisting of 24 items which cluster into four dimensions: attention, social problems, hyperactivity, and oppositional behavior. Both instruments are psychometrically sound and supported by normative data. (In other instances, children may be asked to complete checklists about their own behavior; we have reserved discussion of children's self-report instruments in our review of assessment of covert behavior in the next section).

Interviews

Of the many procedures employed by behavioral clinicians, the interview is the most widely used (Swann & MacDonald, 1978) and is also considered an indispensable part of assessment (Gross, 1984; Linehan, 1977; McConaughy, 1996); so indispensable, in fact, that we discuss behavioral interviews not only in this section as related to assessment of a child's overt behavior but also in subsequent sections relative to assessment of a child's covert behavior and of the behavior and cognitions of others in a child's environments. Similar to naturalistic observations, such interviews may be *informal* or *formal*.

As an assessment of overt behavior, informal interviews may yield a wide range of detailed information about a child's specific behavior and much preliminary information about possible controlling variables. Such information may assist in the early formulation of treatment plans and in the development of relationships with the child and his or her family (Ollendick & Cerny, 1981). As with direct observation and behavior checklists, clinicians must recognize that the impartiality and objectivity of informants is uncertain. At the risk of redundancy, reporters—adults and children—are active interpreters of behavior; thus, the accuracy and reliability of information they provide is always subject to verification.

The popularity of interviews may derive in part from a number of practical considerations, as well as to advantages they offer over other procedures (Gross, 1984). As we have noted, direct observations of target behaviors are essential but frequently inconvenient and impractical. Moreover, behavior checklists often do not permit detailed exploration or elaboration of responses. An informal interview permits the clinician to efficiently obtain a broad band of information about overall functioning as well as detailed information about specific areas, and this information can be used as the initial basis for early hypotheses and interventions. Informal interviews also involve greater flexibility than formal interviewing methods described below, allowing the clinician to build a relationship with the child and his or her family and to obtain information that might otherwise not be revealed. As noted by Linehan (1977), some family members may be more likely to divulge information verbally in the context of a professional relationship than to write it down on a form to be entered into a permanent file.

Formal interviews may be either structured or semi-structured. In general, structured interviews are oriented toward specific diagnostic categories and require parents (and/or children and adolescents) to endorse diagnostic items for the wide range of childhood disorders in the DSM-IV (American Psychiatric Association, 1994) or ICD-10 (World Health Organization, 1991). Clearly, such interviews facilitate collection of systematic data relative to a broad range of symptoms and diagnoses. However, structured interviews may be limited by (a) an overem-

phasis on diagnostic categories (to the exclusion of other important information), (b) weak self-report reliability for children under age 13, (c) low reliability between responses of children and parents, and (d) dichotomous present-or-absent scoring (McConaughy, 1996). Further, structured interviews often do not yield specific information about contextual factors; thus, even when a structured interview is used, subsequent informal interviewing is often required for the purpose of clarifying responses. Several reliable, valid "omnibus" structured diagnostic interviews are available, such as the Diagnostic Interview Schedule for Children—Version 2.3 (DISC-2.3; Shaffer, 1992), recently revised to reflect DSM-IV criteria. Other structured interviews are oriented toward a specific domain, such as anxiety, as with the Anxiety Disorders Interview Schedule for Children (ADIS; Silverman & Nelles, 1988). Various semi-structured interviews have been developed as an alternative to the rigid format and focus on categorical diagnoses that characterize structured interviews, including the Semistructured Clinical Interview for Children and Adolescents (SCICA: McConaughy & Achenbach, 1994), a child and adolescent self-report interview. Semi-structured interviews may also focus on a more specific domain such as social functioning, as in the case of the Social Adjustment Inventory for Children and Adolescents (SAICA: Orvaschel & Walsh, 1984).

ASSESSMENT OF A CHILD'S COVERT PROCESSES

We turn our attention now to the issues of "What does a child think and feel, and when?" and "How do these thoughts and feelings connect with a child's overt behavior?" If one acknowledges the influence of cognitive and affective processes on overt behavior, then the importance of asking and answering these questions is self-evident. In addition to obtaining information from adult caretakers about a child's covert processes—via interviews or behavior checklists—more direct methods for obtaining such information may also be useful. As described above cognitive social learning person variables (CSLPVs)—competencies, encoding strategies, expectancies, subjective values, and self-regulatory systems and plans—provide a framework for guiding the assessment of covert processes.

Interviews

Clearly, directly interviewing a child may yield much information about covert processes, assuming the child is willing and able to report on such processes. As we have noted, early behaviorists eschewed such interviews, maintaining that observable behavior was the least inferential method of obtaining assessment information. To a large extent, such negative bias against self-report was an outgrowth of early findings indicating that reports of subjective states did not always coincide with observable behaviors (Finch & Rogers, 1984). While congruence in responding is, in fact, not always observed, contemporary researchers have cogently argued that a child's *perceptions* of his or her behavior and its consequences may be as important for behavior change as the behavior itself (Finch, Nelson, & Moss, 1983; Ollendick & Hersen, 1984). We are aware of no structured interviews used specifically for the purpose of assessing CSLPVs, but informal interviews may yield critical information about competencies ("Besides hitting, what else could you do when your brother enters your room without permission"), encoding strategies ("Do the kids at school seem to like you?"), expectancies ("What would happen if you told your friend that her comment made you feel bad?"), subjective values ("How important is it to you to do well on that science test?"), and self-regulatory systems and plans ("Do you think you could begin keeping track of how often you speak out of turn in class?"). It may also be useful to obtain information in these domains through more formal means, to which we now focus our attention.

Direct Measures

An excellent example of a broad-based instrument used in the assessment of competencies is the IQ test. While one could reasonably argue that an IQ test—such as the Wechsler Intelligence Scale for Children—Third Edition (WISC-III; Wechsler, 1991)—is, in fact, a sampling of *overt* behavior, there seems little question that the broad conclusions tapped by IQ measures go well beyond the single sampling of behavior obtained during testing. Indeed, IQ tests may provide significant information about a child's competencies in a wide range of areas, such as acquired knowledge, abstract logical reasoning skills, problem-solving skills, long- and short-term auditory and visual memory, social judgment, sequencing skills, mental flexibility, distractibility,

frustration tolerance, and so forth, each of which may impact quite directly on a child's behavior. The impressive normative data available for measures such as the WISC-III add to its value and credibility as a direct measure of a broad range of competencies.

Other direct measures may be used to gauge CSLPVs. For example, computerized continuous performance tasks (CPT) have a long history in the assessment of sustained attention, vigilance, and impulse control (e.g., Rosvold, Mirsky, Sarason, Bransome, & Beck, 1956), skills which may fall under the general umbrella of self-regulatory systems and plans. Most CPTs, such as the Gordon Diagnostic System (GDS; Gordon, 1983), provide normative information. Other direct measures, such as the Matching Familiar Figures Test (MFFT; Kagan, 1966), have also been used to assess impulse control. The ecological validity of these measures has been questioned as related to the diagnosis of Attention Deficit Hyperactivity Disorder (e.g., Barkley, 1990; DuPaul, Anastopoulos, Shelton, Guevremont, & Metevia, 1992); nonetheless, these instruments directly tap areas related to self-regulatory systems and plans and may therefore by useful to the assessment process.

Self-Report Instruments

In addition to simply asking a child about covert processes in an interview format, various self-report instruments have also been developed to access this information. As with parent- and teacher-completed checklists, self-reports should be used with appropriate caution and due regard for their specific limitations. Because they generally involve the child's retrospective rating of attitudes, feelings, and behaviors, self-report instruments must be considered indirect methods of assessment (Cone, 1978).

Some self-report checklists focus on a broad range of overt and covert processes, as in the case of the Youth Self-Report (YSR; Achenbach, 1991c). The YSR includes items very similar to the CBCL and CBC-TRF described earlier, and is appropriate for children and adolescents aged 11 to 18 years. While many YSR items are reflective of overt behaviors (e.g., I bite my fingernails; I destroy my own things; I have nightmares), it is perhaps most useful for assessing a variety of covert behaviors (e.g., I'm too dependent on adults; I don't feel guilty after doing something I shouldn't; I am afraid of going to school; I feel that I have to be perfect; I feel that others are out to get me; my moods or feelings change suddenly). As such, the YSR may be used to assess various of the cognitive social learning person variables (e.g., encoding strategies, subjective values, expectancies).

Other self-report checklists may tap a more specific area of interest, such as anxiety. Few would argue against the notion that anxiety is comprised of both overt *and* covert behaviors. Indeed, anxiety may tap into numerous CSLPVs, including expectancies and encoding strategies. Moreover, some anxieties may not be manifested overtly in some children, and assessment methods which are primarily oriented toward overt behavior may therefore be less informative. Self-report scales such as the Revised Children's Manifest Anxiety Scale (RCMAS; Reynolds & Richmond, 1978), the State-Trait Anxiety Inventory for Children (STAIC; Spielberger, 1973), the Child Anxiety Sensitivity Index (CASI; Silverman, Fleisig, Rabian, & Peterson, 1991), the Social Anxiety Scale for Children—Revised (SASC-R; La Greca & Stone, 1993), and the Fear Survey Schedule for Children-Revised (FSSR-R; Ollendick, 1983; Ollendick, Matson, & Helsel, 1985) may be extremely useful for assessing covert aspects of children's anxieties. We will describe this latter instrument for purposes of illustration.

The FSSC-R is the revised version of the Fear Survey Schedule for Children (Scherer & Nakamura, 1968). In the revised scale, designed to be used with younger and middle-age children, the child is instructed to rate his or her fear level to each of 80 items on a 3-point scale. Children are asked to indicate whether a specific fear item (e.g., having to go to school, being punished by father, dark places, riding in a car) frightens them "not at all," "some," or "a lot." Factor analysis of the scale has revealed five primary factors: fear of failure or criticism, fear of the unknown, fear of injury and small animals, fear of danger and death, and medical fears (Ollendick, King, & Frary, 1989). Moreover, it has been shown that girls report greater fear than boys, that specific fears change developmentally, and that the most prevalent fears of boys and girls have remained unchanged over the past 30 years. Further, recent studies have shown that the instrument is sensitive to cultural influences (Ollendick, Yang, King, Dong, & Akande, 1996). Such information is extremely useful for determining whether a child of a specific age, gender, and culture is excessively fearful. Further, the instrument can be used to differentiate subtypes

of specifically phobic youngsters whose fear of school is related to separation anxiety (e.g., death, having parents argue, being alone) from those whose fear is due to specific aspects of the school situation (e.g., taking a test, making a mistake, being sent to the principal).

Still other instruments are available to tap a broad range of additional covert processes, such as self-concept (e.g., the Piers-Harris Children's Self-Concept Scale [Piers, 1984]), self-perception (the Self-Perception Profile for Children [Harter, 1985] and the Self-Perception Profile for Adolescents [Harter, 1988]), depression (e.g., the Children's Depression Inventory—Short Form [Kovacs, 1992]), locus-of-control (e.g., the Nowicki-Strickland Locus of Control Scale for Children [Nowicki & Strickland, 1973]), social relationships and assertion (e.g., Ollendick, 1984; Scanlon & Ollendick, 1985), and reinforcer preferences (e.g., Clement & Richard, 1976).

Self-Monitoring

Self-monitoring can be used to assess both overt and covert behavior; we have placed it in this section on covert processes for emphasis. Self-monitoring differs from self-report in that it constitutes an observation of clinically relevant thoughts and feelings at the time of their occurrence. As such, it is a more direct method of assessment. Self-monitoring requires a child to monitor his or her own cognitions/feelings ("I'm dumb," "I'm scared," "I'm unhappy," "I'm feeling out of control," etc.) and then to record their occurrence systematically. Typically, the child is asked to keep a diary, place marks on a card, or push the plunger on a counter as cognitions/feelings occur or immediately thereafter. Although self-monitoring procedures have been used with both children and adults, at least three considerations must be attended to when such procedures are used with younger children (Shapiro, 1984). The cognitions/feelings should be clearly defined, prompts to use the procedures should be readily available, and rewards for their use should be provided. Some children will be less attuned to their inner feelings and thoughts than others and may therefore require preliminary coaching. Other children may have difficulty remembering exactly which cognitions/feelings to monitor and how those covert behaviors are defined. For these reasons, it is generally considered desirable to provide the child a brief description of the targeted covert behaviors or, better yet, a picture of it, and to have the child record only one or two cognitions/feelings at a time. The key to successful self-monitoring in children is use of recording procedures that are highly portable, simple, time-efficient, and relatively unobtrusive.

ASSESSMENT OF CONTEXTS

In this section we discuss methods for measuring contextual factors—characteristics of adult caretakers and environments—which may contribute to variations in a child's overt and covert behavior. Recall that, from a social learning perspective, situational factors may exert considerable influence on a child's behavior. Methods for assessing contexts have become more plentiful as an appreciation for this influence has evolved. As might be expected, home and school contexts, and the adults who interact with children in these settings, have become popular targets of assessment. In an earlier section we noted the potential value of naturalistic observation as an avenue for obtaining information about the impact of various situational factors on a child's behavior. We turn our attention now to other methods for assessing contexts.

Interviews

Interviews with children and the adults who interact with them have the potential to yield significant information about contexts, assuming the interviewer is determined to obtain such information. Parents and teachers may not be highly sensitive to variations in a child's behavior in different situations and may need significant prompting along these lines ("Is Johnny aggressive *all* the time or primarily during certain activities?" "Is Susie inattentive during all class activities or especially during certain types of lessons or when certain task demands are present?" "Does Tony say he wishes he were dead only when he is tantrumming or also at other times?"). Interviews also allow the clinician the opportunity to formally or informally assess the overt and covert behaviors of important adults in the child's life and form initial impressions about the manner in which these behaviors affect the behavior of the identified child. Children rarely refer themselves for treatment; invariably, they are referred by adults whose perceptions (encoding strategies) of a child's problems will be

important to gauge. To what degree do the adults have accurate perceptions of the developmental deviance of a child's behavior? What attributions do the adults make about the child's behaviors (e.g., do they believe the behaviors are due to poor parenting, insensitive teachers, significant life events, biological factors, etc.)? Do the adults believe the child has the capacity (competencies) and motivation (subjective values) to alter behaviors defined as problematic?

In addition to encoding strategies, it will be important to assess other CSLPVs in the important adults in a child's life. For example, one of the most commonly recommended interventions for poorly self-regulated children is behavior management strategies (e.g, contingency contracting, time-out); however, one should not assume that all parents and teachers have equal capacities (*competencies*) for implementation of such strategies, or that all parents and teachers will "stick with the program" (*self-regulatory systems and plans*) for a sufficient period (see Greene, 1995, 1996, for more thorough discussion of the issue of "teacher-treatment compatibility"). These variations in competencies and self-regulatory systems and plans may have significant ramifications for intervention selection and treatment duration. If parents have attempted such strategies in the past, it may also be fruitful to inquire about their *expectancies* regarding the likely outcome of implementing such strategies? It may also be important to assess the parents' *subjective values* as regards parenting goals and priorities. Do the parents value independence in their children? Cleanliness? Timeliness? Honesty? Good grades? Is it important to the parents that their children attend Ivy League schools? To what degree are these subjective values compatible with the subjective values and competencies of the parents' children?

Group interviews often permit the interviewer an opportunity to observe problematic interactions (e.g., parent-child interactions, mother-father interactions, parent-teacher interactions) which may not have been included in the adults' original perception of the problematic behavior, and to explore discrepancies in reports and perceptions of all individuals present ("Mrs. Utley, your daughter seems to feel that you are constantly criticizing her...what's your sense about that?" "Mr. Jones, you don't seem as troubled by your son's behaviors as does your wife...do you feel that's an accurate perception on my part?" "I get the sense that there are very different opinions between school and home about the degree to which Adam is actually able to control his excessive motor activity").

Finally, while we have focused primarily on the covert domain in the discussion above, we wish to emphasize that interviews also provide an extremely valuable mechanism for assessing the overt behavior of important persons in the various environments in which the identified child interacts. In other words, while it is important to assess what a child's parent or teacher thinks and feels about the child, it will also be critical to assess what parents or teachers actually *do* in their interactions with the child. For example, while a parent's subjective values may suggest that she or he believes screaming is an inappropriate parenting strategy, the parent may nonetheless report that she or he screams at the child frequently. Such discrepancy between thought and action may be an important consideration in choosing efficacious treatments.

Checklists

Various checklists have been developed to provide information about a host of contextual factors, particularly in the home and school environments. One of the most widely used global measures of the family environment is the Family Environment Scale (Moos & Moos, 1986), a parent-completed measure which provides information about family functioning in the global domains of cohesiveness, expressiveness, and conflict. A second measure has also been used to assess family cohesion and adaptability (Family Adaptability and Cohesion Evaluation Scales-II; Olson, Portner, & Bell, 1982). A measure tapping the situational pervasiveness and severity of child behavior problems in different home settings (e.g., while playing with other children, when asked to do chores, when asked to do school homework) is the Home Situations Questionnaire (HSQ; Barkley & Edelbrock, 1987) and its revision (HSQ-R; DuPaul & Barkley, 1992). Although scores may be derived from the HSQ and HSQ-R based on the number of situations in which a child exhibits problems and the severity of these problems, these instruments are perhaps best utilized as informal devices for efficiently obtaining information about a child's behavior in a wide range of home situations. Studies have shown both instruments to possess adequate psychometric properties. Other instruments may measure narrower aspects of family functioning, such as family learning environment (e.g., the Family Learning Environment Scale; Marjoribanks, 1979).

Other checklists may be useful in assessing characteristics of parents, including parental psychopathology (e.g., SCL-90-R; Derogatis, 1983); parental expectations regarding children's developmental abilities (e.g., the Parent Opinion Questionnaire; Azar, Robinson, Hekimian, & Twentyman, 1984); dysfunctional discipline practices (e.g., the Parenting Scale: Arnold, O'Leary, Wolff, & Acker, 1993); the degree to which a parent finds interactions with a particular child to be stressful (e.g., the Parenting Stress Index; Abidin, 1986, 1990); the quality of the parents' relationship with each other (e.g., the Dyadic Adjustment Scale; Spanier, 1976, 1989); and the degree to which parents believe they have a sound working relationship with the child's other parent (e.g., the Parenting Alliance Inventory; Abidin & Brunner, 1995).

While the above instruments are completed by parents, some instruments have been developed to assess children's perceptions of contexts, including the Social Support Scale for Children (SPPC; Harter, 1986), which assesses the degree to which a child feels supported by others (such as parents, teachers, classmates, and close friends); the Children's Report of Parenting Behavior Inventory (CRPBI; Schluderman & Schluderman, 1970), which is designed to assess a child's perceptions of his or her parents' behavior; and the Social Support Appraisals Scale (APP; Dubow & Ullman, 1989), which measures children's subjective appraisals of social support provided by family, peers, and teachers.

The School Situations Questionnaire (SSQ; Barkley & Edelbrock, 1987) and its revision (SSQ-R; DuPaul & Barkley, 1992) are similar to the HSQ and HSQ-R described above, and assess the pervasiveness of behavior problems across various school situations (e.g., during lectures to the class, during class discussions, during individual deskwork). Studies have shown both instruments to possess adequate psychometric properties.

Other checklists may be useful in assessing characteristics of teachers. For example, the SBS Inventory of Social Behavior Standards and Expectations (Walker & Rankin, 1983) assesses teachers' behavioral expectations regarding students' behavior. Section I consists of 56 items describing adaptive student behaviors; respondents mark each item as "critical," "desirable," or "unimportant." Section II of the SBS gauges teachers' tolerances for various problematic behaviors. Section II consists of 56 items describing maladaptive student behaviors; respondents mark each item as "unacceptable," "tolerated," or "acceptable." The SBS has been shown to have excellent psychometric properties and has been used in previous studies to identify teacher expectations that are associated with effective teaching of behaviorally challenging students (e.g., Gersten, Walker, & Darch, 1988; Kauffman, Lloyd, & McGee, 1989; Kauffman, Wong, Lloyd, Hung, & Pullen, 1991).

A second measure, the Index of Teaching Stress (ITS; Greene & Abidin, 1995, Greene, Abidin, & Kmetz, 1997) assesses the degree to which a teacher experiences stress and frustration in teaching a given student, and was developed as a companion to the Parenting Stress Index described earlier. The ITS consists of two sections; the first is comprised of student behaviors which may induce stress or frustration in a teacher (this section includes five factors: ADHD, emotional lability/low adaptability, anxiety/withdrawal, low ability/learning disabled, and aggressive/conduct disorder). The second domain is comprised of various domains of teacher stress and frustration (this section includes four factors: self-doubt/needs support, loss of satisfaction from teaching, disrupted teaching process, and frustration working with parents). The utility of the ITS has been demonstrated in a longitudinal study exploring school outcome for children with ADHD (Greene, Beszterczey, Katzenstein, & Park, 1999).

Cultural Considerations

Numerous observers have called attention to the internationalization of the world and the "browning of America" (e.g., Malgady, Rogler, & Constantino, 1987; Vazquez Nuttall, De Leon, & Del Valle, 1990, Vazquez Nuttall, Sanchez, Borras Osorio, Nuttall, & Varvogil, 1996). As regards the topic of this chapter, this means that the assessment process is increasingly being applied to non-Caucasian children for whom English is not the primary language, and that use of assessment procedures that are gender-, culture-, and language-fair has become a major concern and of utmost importance.

Various cultural issues must be considered in the assessment process. Cultural differences may be expressed in child-rearing practices, family values, parental expectations, communication styles, nonverbal communication patterns, and family structure and dynamics (Vazquez Nuttall et al., 1996).

Behaviors characteristic of ethnic minority children may be seen as emotionally or behaviorally maladaptive by persons who have little or no appreciation for cultural norms (e.g., Prewitt-Diaz, 1989). Thus, cultural biases may occur early in the referral process. Fortunately, steps can be taken to minimize cultural biases in the assessment process itself.

Vazquez and colleagues (1996) have delineated a variety of steps which may be useful for incorporating cultural considerations into the assessment process: (1) including extended family members in the information-gathering process; (2) use of interpreters in interviewing; (3) familiarizing oneself with the culture of specific ethnic groups; and (4) using instruments that have been translated into the native language of reporters *and* for which norms are available for specific ethnic groups. With regard to this latter recommendation, significantly greater progress has occurred for the translation component than for the normative component. In sum, while considerable progress has been made to incorporate developmental considerations into assessment technology, the future challenge—to make similar progress in the cultural domain—is clearly before us.

SUMMARY

In the preceding pages we have described the various components of a behavioral assessment, reviewed these components in a historical context, and provided an overview of various assessment procedures which may be useful in conducting a thorough and comprehensive behavioral assessment. We have also asserted that perhaps the most pressing challenge for behavioral asssessment at this time is the development of culturally sensitive instruments and assessors.

Yet, in outlining components of a behavioral assessment, we should emphasize several important points. First, regardless of the procedures employed, child behavioral assessments must be conducted by persons with the training and experience to execute them in a knowledgeable fashion and the skills to analyze, organize, integrate, and communicate the vast array of information gathered in the assessment process for purposes of (a) arriving at a comprehensive understanding of a child's interactions with his or her environment(s); (b) requiring that additional information be collected when such an understanding has not been achieved; (c) making accurate judgments regarding the developmental deviance of a child's behavior; (d) determining the most appropriate persons and behaviors to be targeted for change and the interventions most likely to produce these desired changes; and (e) maintaining contact over the long term with various adults who continue to interact with the child and who are charged with implementation of interventions and/or are targets of intervention; and (f) monitoring the continuous, fluid assessment process and facilitating reformulation of "the problem" as necessary. Assessors must also be well-acquainted with the nature of information provided by each assessment procedure—in other words, what conclusions can and cannot be arrived at on the basis of the information provided by a particular instrument (Greene, 1995).

Second, while it may be obvious that, in making normative comparisons and using standardized instruments, assessors are employing certain aspects of a nomothetic approach to assessment (the application of general laws as applied to large numbers of children), we continue to view behavioral assessment of children as a primarily *idiographic* undertaking (concerned more with the uniqueness of a given child). Unlike the *nomothetic* approach, the idiographic perspective emphasizes the discovery of relationships among variables uniquely patterned in each child. As Mischel (1968) observed some years ago, "Behavioral assessment involves an exploration of the unique or idiosyncratic aspects of the single case, perhaps to a greater extent than any other approach" (p. 190). While we eagerly anticipate theoretical and technological advances of the future, we believe that this assertion regarding behavioral assessment will continue to ring true in the next millenium.

REFERENCES

Abidin, R. R. (1983). *Parenting Stress Index (PSI) manual*. Charlottesville, VA: Pediatric Psychology Press.

Abidin, R. R. (1990). *Parenting Stress Index (PSI) manual* (3rd ed.). Charlottesville, VA: Pediatric Psychology Press.

Abidin, R. R., & Brunner, J. F. (1995). Development of a Parenting Alliance Inventory. *Journal of Clinical Child Psychology, 24*(1), 31–40.

Abikoff, H., Gittelman-Klein, R., & Klein, D. (1977). Validation of a classroom observation code for hyperactive children. *Journal of Consulting and Clinical Psychology, 45*, 772–783.

Abikoff, H., & Gittelman, R. (1985). Classroom Observation Code: A modificaiton of the Stony Brook code. *Psychopharmacology Bulletin, 21*(4), 901–909.

Achenbach, T. M. (1986). *Manual for the Child Behavior Checklist Direct Observation Form.* Burlington: University of Vermont, Department of Psychiatry.

Achenbach, T. M. (1991a). *Manual for the Child Behavior Checklist and Revised Child Behavior Profile.* Burlington: University of Vermont, Department of Psychiatry.

Achenbach, T. M. (1991b). *Manual for the Teacher Report Form and 1991 Profile.* Burlington: University of Vermont, Department of Psychiatry.

Achenbach, T. M. (1991c). *Manual for the Youth Self-Report and 1991 Profile.* Burlington: University of Vermont, Department of Psychiatry.

Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. *Psychological Bulletin, 101,* 213–232.

American Psychiatric Association (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: American Psychiatric Association.

Anastasi, A. (1982). *Psychological testing* (5th ed.). New York: Macmillan.

Arnold, D. S., O'Leary, S. G., Wolff, L. S., & Acker, M. M. (1993). The Parenting Scale: A measure of dysfunctional parenting in discipline situations. *Psychological Assessment, 5,* 131–136.

Azar, S. T., Robinson, D. R., Hekimian, E., & Twentyman, C. T. (1984). Unrealistic expectations and problem-solving ability in maltreating and comparison mothers. *Journal of Consulting and Clinical Psychology, 52,* 687–691.

Bandura, A. (1971). *Social learning theory.* Englewood Cliffs, NJ: General Learning Press.

Bandura, A. (1973). *Aggression: A social learning analysis.* Englewood Cliffs, NJ: Prentice-Hall.

Bandura, A. (1986). *Social foundations of thought and action: A social-cognitive theory.* Englewood Cliffs, NJ: Prentice-Hall.

Barkley, R. A. (1981). *Hyperactive children: A handbook for diagnosis and treatment.* New York: Guilford Press.

Barkley, R. A. (1990). *Attention deficit hyperactivity disorder: A handbook for diagnosis and treatment.* New York: Guilford Press.

Barkley, R. A. (1994). Impaired delayed responding: A unified theory of attention deficit hyperactivity disorder. In D. K. Routh (Ed.), *The disruptive behavior disorders in children: Essays in honor of Herbert C. Quay* (pp. 11–57). New York: Plenum Press.

Barkley, R. A., & Edelbrock, C. S. (1987). Assessing situational variation in children's behavior problems: the Home and School Situations Questionnaires. In R. Prinz (Ed.), *Advances in behavioral assessment of children and families* (Vol. 3, pp. 157–176). Greenwich, CT: JAI Press.

Breen, M. J., Eckert, T. L., & DuPaul, G. J. (1996). Interpreting child behavior questionnaires. In M. J. Breen & C. R. Fiedler (Eds.), *Behavioral approach to assessment of youth with emotional/ behavioral disorders* (pp. 225–242). Austin, TX: Pro-Ed.

Campbell, S. B. (1989). Developmental perspectives in child psychopathology. In T. H. Ollendick & M. Hersen (Eds.), *Handbook of child psychopathology* (2nd ed.). New York: Plenum Press.

Cicchetti, D. (1984). The emergence of developmental psychopathology. *Child Development, 55,* 1–57.

Cicchetti, D. (1993). Developmental psychopathology: Reactions, reflections, and projections. *Developmental Review, 13,* 471–502.

Clement, P. W., & Richard, R. C. (1976). Identifying reinforcers for children: A Children's Reinforcement Survey. In E. J. Mash & L. G. Terdal (Eds.), *Behavior therapy assessment: Diagnosis, design, and evaluation* (pp. 207–216). New York: Springer.

Cone, J. D. (1978). The Behavioral Assessment Grid (BAG): A conceptual framework and taxonomy. *Behavior Therapy, 9,* 882-888.

Derogatis, L. (1983). *Manual for the Symptom Checklist 90-Revised (SCL 90-R).* Baltimore, MD: Author.

Dubow, E. F., & Ullman, D. G. (1989). Assessing social support in elementary school children: The Survey of Children's Social Support. *Journal of Clinical Child Psychology, 18,* 52–64.

DuPaul, G. J., Anastopoulos, A. D., Shelton, T. L., Guevremont, D. C., & Metevia, L. (1992). Multimethod assessment of Attention-Deficit Hyperactivity Disorder: The diagnostic utility of clinic-based tests. *Journal of Clinical Child Psychology, 21,* 394–402.

DuPaul, G. J., & Barkley, R. A. (1992). Situational variability of attention problems: Psychometric properties of the revised Home and School Situations Questionnaires. *Journal of Clinical Child Psychology, 21,* 178–188.

Eyberg, S., Bessmer, J., Newcomb, K., Edwards, D., & Robinson, E., (1994). *Dyadic Parent-Child Coding System II.* Unpublished manuscript, University of Florida, Gainesville.

Finch, A. J., Nelson, W. M., III, & Moss, J. H. (1983). A cognitive-behavioral approach to anger management with emotionally disturbed children. In A. J. Finch, W. M. Nelson, & E. S. Ott (Eds.), *Cognitive behavioral approaches to treatment with children.* Jamaica, NY: Spectrum Publications.

Finch, A. J., & Rogers, T. R. (1984). Self-report instruments. In T. H. Ollendick & M. Hersen (Eds.), *Child behavioral assessment: Principles and procedures.* Elmsford, NY: Pergamon Press.

Gersten, R., Walker, H., & Darch, C. (1988). Relationship between teachers' effectiveness and their tolerance for handicapped students. *Exceptional Children, 54,* 433–438.

Gordon, M. (1983). *The Gordon Diagnostic System.* Boulder, CO: Clinical Diagnostic Systems.

Greene, R. W. (1995). Students with ADHD in school classrooms: Teacher factors related to compatibility, assessment, and intervention. *School Psychology Review, 24*(1), 81–93.

Greene, R. W. (1996). Students with ADHD and their teachers: Implications of a goodness-of-fit perspective. In T. H. Ollendick & R. J. Prinz (Eds.), *Advances in Clinical Child Psychology* (pp. 205–230). New York: Plenum.

Greene, R. W., & Abidin, R. R. (1995). *The Index of Teaching Stress: A new measure of student-teacher compatibility.* Paper presented at the 27th Annual Meeting of the National Association of School Psychologists, Chicago, IL

Greene, R. W., Abidin, R. R., & Kmetz, C. (1997). The Index of Teaching Stress: A measure of student-teacher compatibility. *Journal of School Psychology, 35*(3), 239–259.

Greene, R. W., Beszterczey, S. K., Katzenstein T., & Park, K. (1999). Are students with ADHD more stressful to teach? Patterns and predictors of teacher stress in an elementary-age sample, under review.

Greenwood, C. R., Delquadri, J. C., Stanley, S. O., Terry, B., & Hall, R. V. (1985). Assessment of eco-behavioral interaction in school settings. *Behavioral Assessment, 7,* 331–347.

Gresham, F. M., & Elliott, S. N. (1990). *Social Skills Rating System manual.* Circle Pines, MN: American Guidance Service.

Gross, A. M. (1984). Behavioral interviewing. In T. H. Ollendick & M. Hersen (Eds.), *Child behavioral assessment: Principles and procedures.* Elmsford, NY: Pergamon Press.

Harter, S. (1985). *The Self-Perception Profile for Children: Revision of the Perceived Competence Scale for Children [manual].* Denver, CO: University of Colorado.

Harter, S. (1986). *Manual: Social Support Scale for Children.* Denver, CO: University of Denver.

Harter, S. (1988). *Manual for the Self-Perception Profile for Adolescents.* Denver, CO: University of Denver Press.

Hayes, S. C., Nelson, R. O., & Jarrett, R. B. (1987) The treatment utility of assessment: A functional approach to evaluating assessment quality. *American Psychologist, 42,* 963-974.

Hops, H., Biglan, A., Tolman, A., Sherman, L., Arthur, J., Warner, P., Romano, J., Turner, J., Friedman, L., Bulcroft, R., Holcomb, C., Oostenink, N., & Osteen, V. (1990). *Living In Familial Environments (LIFE) coding system: Training/procedures and reference manual for coders* (rev. ed.). Eugune: Oregon Research Institute.

Hops, H., Davis, B., & Longoria, N. (1995). Methodological issues in direct observation: Illustrations with the Living in Familial Environments (LIFE) coding system. *Journal of Clinical Child Psychology, 24*(2), 193–203.

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

Kagan, J. (1966). Reflection-impulsivity: The generality and dynamics of conceptual tempo. *Journal of Abnormal Psychology, 71,* 17–24.

Kauffman, J. M., Lloyd, J. W., & McGee, K. A. (1989). Adaptive and maladaptive behavior:Teachers' attitudes and their technical assistance needs. *Journal of Special Education, 23,* 185–200.

Kauffman, J. M., Wong, K. L. H., Lloyd, J. W., Hung, L. Y., & Pullen, P. L. (1991). What puts pupils at risk? An analysis of classroom teachers' judgments of pupils' behavior. *Remedial and Special Education, 12,* 7–16.

Kovacs, M. (1992). *Children's Depression Inventory.* Los Angeles: Multi-Health Systems.

LaGreca, A. M., & Stone, W. L. (1993). Social Anxiety Scale for Children—Revised: Factor structure and concurrent validity. *Journal of Clinical Child Psychology, 22,* 17–27.

Linehan, M. (1977). Issues in behavioral interviewing. In J. D. Cone & R. P. Hawkins (Eds.), *Behavioral assessment: New directions in clinical psychology.* New York: Brunner/Mazel.

Malgady, R., Rogler, L., & Constantino, G. (1987). Ethnocultural and linguistic bias in mental health

evaluation of Hispanics. *American Psychologist, 42*, 228–234.

Marjoribanks, K. (1979). *Families and their learning environments*. London: Routledge & Kegan Paul.

Mash, E. J., & Barkley, R. A. (1986). assessment of family interaction with the Response Class Matrix. In R. Prinz (Ed.), *Avances in behavioral assessment of children and families* (Vol. 2, pp. 29–67). Greenwich, CT: JAI Press.

Mash, E. J., & Terdal, L. G. (1988). Behavioral Assessment of child and family disturbance. In E. J. Mash & L. G. Terdal (Eds.), *Behavioral Assessment of Childhood Disorders* (pp. 3–65). New York: Guilford Press.

McConaughy, S. H. (1996). The interview process. In M. J. Breen & C. R. Fiedler (Eds.), *Behavioral approach to assessment of youth with emotional/behavioral disorders: A handbook for school-based practitioners* (pp.181–224). Austin, TX: ProEd.

McConaughy, S. H., & Achenbach, T. M. (1994). *Manual for the Semistructured Clinical Interview for Children and Adolescents*. Burlington: University of Vermont, Department of Psychiatry.

McConaughy, S. H., Achenbach, T. M., & Gent, C. L. (1988). Multiaxial empirically based assessment: Parent, teacher, observational, cognitive, and personality correlates of Child Behavior Profiles for 6-11-year-old boys. *Journal of Abnormal Child Psychology, 16*, 485–509.

Mischel, W. (1968). *Personality and assessment*. New York: Wiley.

Mischel, W. (1973). Toward a cognitive social learning reconceptualization of personality. *Psychological Review, 80*, 252–283.

Mischel, W. (1979). On the interface of cognition and personality. *American Psychologist, 34*, 740–754.

Mischel, W. (1984). Convergences and challenges in the search for consistency. *American Psychologist, 39*, 351–364.

Moos, R. H., & Moos, B. S. (1986). *Family Environment Scale manual* (2nd ed.). Palo Alto, CA: Consulting Psychologists Press.

Nowicki, S., & Strickland, B. R. (1973). A locus of control scale for children. *Journal of Consulting and Clinical Psychology, 40*, 148–154.

Novick, J., Rosenfeld, E., Bloch, D. A., & Dawson, D. (1966). Ascertaining deviant behavior in children. *Journal of Consulting and Clinical Psychology, 30*, 230-238.

Ollendick, T. H. (1983). Reliability and validity of the Revised-Fear Survey Schedule for Children (FSSC-R). *Behaviour Research and Therapy, 21*, 685-692.

Ollendick, T. H. (1984). Development and validation of the Children's Assertiveness Inventory. *Child and Family Behavior Therapy, 5*, 1–15.

Ollendick, T. H., & Cerny, J. A. (1981). *Clinical behavior therapy with children*. New York: Plenum Press.

Ollendick, T. H., & Greene, R. W. (1990). Behavioral assessment of children. In G. Goldstein & M. Hersen (Eds.), *Handbook of psychological assessment* (2nd ed.). New York: Pergamon.

Ollendick, T. H., & Hersen, M. (Eds.) (1984). *Child behavioral assessment: Principles and procedures*. New York: Pergamon Press.

Ollendick, T. H., King, N. J., & Frary, R. B. (1989). Fears in children and adolescents: Reliability and generalizability across gender, age, and nationality. *Behaviour Research and Therapy, 27*,19–26.

Ollendick, T. H., Matson, J. L., & Helsel, W. J. (1985). Fears in children and adolescents: Normative data. *Behaviour Research and Therapy, 23*, 465-467.

Ollendick, T. H., Yang, B., King, N. J., Dong, Q., & Akande, A. (1996). Fears in American, Australian, Chinese, and Nigerian children and adolescents: A cross-cultural study. *Journal of Child Psychology and Psychiatry, 37*, 213–220.

Olson, D. H., Bell, R. O., & Portner, J. A. (1982). *Manual for FACES II: Family Adaptability and Cohesion Scales*. St. Paul, MN: University of Minnesota, Family Social Science.

Orvaschel, H., & Walsh, G. (1984). *The assessment of adaptive functioning in children: A review of existing measures suitable for epidemiological and clinical services research*. Washington, DC: U.S. Department of Health and Human Services, NIMH, Division of Biometry and Epidemiology.

Patterson, G. R. (1976). The aggressive child: Victim and architect of a coercive system. In E. J. Mash, L. A. Hammerlynck, & L. C. Hardy (Eds.), *Behavior modification and families*. New York: Brunner/Mazel.

Piers, E. V. (1984). *Revised manual for the Piers-Harris Children's Self-Concept Scale*. Los Angeles: Western Psychological Services.

Prewitt-Diaz, J. (1989). *The process and procedures for identifying exceptional language minority children*. Sate College: Pennsylvania State University.

Reynolds, C. R., & Richmond, B. O. (1978). What I Think and Feel: A revised measure of children's manifest anxiety. *Journal of Abnormal Child Psychology, 6*, 271–280.

Rosvold, H. E., Mirsky, A. F., Sarason, I., Bransome, E. D., & Beck, L. H. (1956). A continuous performance test of brain damage. *Journal of Consulting Psychology, 20*, 343–350.

Rotter, J. B. (1954). *Social learning and clinical psychology*. Englewood Cliffs, NJ: Prentice-Hall.

Rotter, J. B. (1966). Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs, 80* (Whole No. 609).

Rotter, J. B. (1972). Beliefs, social attitudes, and behavior: A social learning analysis. In J. B.

Rotter, J. E. Chance, & E. J. Phares (Eds.), *Applications of a social learning theory of personality*. New York: Holt, Rinehart, & Winston.

Rutter, M., & Garmezy, N. (1983). Developmental psychopathology. In P. Mussen (Ed.), *Handbook of child psychopathology* (Vol. 4, pp. 775–911). New York: Wiley.

Scanlon, E. M., & Ollendick, T. H. (1985). Children's assertive behavior: The reliability and validity of three self-report measures. *Child and Family Behavior Therapy, 7*, 9–21.

Scherer, M. W., & Nakamura, C. Y. (1968). A fear survey schedule for children (FSS-FC): A factor-analytic comparison with manifest anxiety (CMAS). *Behaviour Research and Therapy, 6*, 173-182.

Schluderman, E., & Schluderman, S. (1970). Replicability of factors in children's report of parent behavior (CRPBI). *Journal of Psychology, 76*, 239–249.

Shaffer, D. (1992). *NIMH Diagnostic Interview Schedule for Children, Version 2.3*. New York: Columbia University, Division of Child and Adolescent Psychiatry.

Shapiro, E. S. (1984). Self-monitoring. In T. H. Ollendick & M. Hersen (Eds.), *Child behavioral assessment: Principles and procedures*. Elmsford, NY: Pergamon Press.

Silverman, W. K., Fleisig, W., Rabian, B., & Peterson, R. A. (1991). Childhood Anxiety Sensitivity Index. *Journal of Clinical Child Psychology, 20*, 162–168.

Silverman, W. K., & Nelles, W. B. (1988). The Anxiety Disorders Interview Schedule for Children. *Journal of the American Academy of Child and Adolescent Psychiatry, 27*, 772–778..

Skinner, B. F. (1953). *Science and human behavior*. New York: Macmillan.

Spanier, G. B. (1976). Measuring dyadic adjustment: New scales for assessing the quality of marriage and similar dyads. *Journal of Marriage and the Family, 38*, 15–28.

Spanier, G. B. (1989). *Dyadic Adjustment Scale: A manual*. North Tonawanda, NY: Multi-Health Systems.

Spielberger, C. D. (1973). *State-Trait Anxiety Inventory for Children: Preliminary manual*. Palo Alto, CA: Consulting Psychologists Press.

Stanley, S. O., & Greenwood, C. R. (1981). *CISSAR: Code for instructional structure and student academic response observer's manual*. Kansas City, KS: Juniper Gardens Children's Project, University of Kansas.

Swann, G. E., & MacDonald, M. L. (1978). Behavior therapy in practice: A rational survey of behavior therapists. *Behavior Therapy, 9*, 799-807.

Ullman, R. K., Sleator, E. K., & Sprague, R. L. (1984). A new rating scale for diagnosis and monitoring of ADD Children. *Psychopharmacology Bulletin, 20*, 160–164.

Ullman, R. K., Sleator, E. K., & Sprague, R. L. (1991). *The ADD-H Comprehensive Teacher's Rating Scale* (2nd ed.). Champaign, IL: MetriTech.

Vazquez Nuttall, E., DeLeon, B., & Del Valle, M. (1990). Best practice in considering cultural factors. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology II* (pp. 219–233). Washington, DC: National Association of School Psychologists.

Vazquez Nuttall, E., Sanchez, W., Borras Osorio, L., Nuttall, R. L., & Varvogil, L. (1996). Assessing the culturally and linguistically different child with emotional and behavioral problems. In M. J. Breen & C. R. Fiedler (Eds.), *Behavioral approach to assessment of youth with emotional/behavioral disorders: A handbook for school-based practitioners* (pp. 451–502). Austin, TX: ProEd.

Wahler, R. G. (1976). Deviant child behavior in the family: Developmental speculations and behavior change strategies. In H. Leitenberg (Ed.), *Handbook of behavior modification and behavior therapy*. Englewood Cliffs, NJ: Prentice-Hall.

Walker, H. M., & Rankin, R. (1983). Assessing the behavioral expectations and demands of less restrictive settings. *School Psychology Review, 12*, 274–284.

Wechsler, D. (1991). *Wechsler Intelligence Scale for Children—Third edition manual*. New York: Psychological Corporation.

Winett, R. A., Riley, A. W., King, A. C., & Altman, D. G. (1989). Preventive strategies with children and families. In T. H. Ollendick & M. Hersen (Eds.), *Handbook of child psychopathology* (2nd ed.). New York: Plenum Press.

World Health Organization (1991). *International classification of mental and behavioral disorders: Clinical descriptions and diagnostic guidelines* (10th ed.). Geneva: Author.

CHAPTER 19

BEHAVIORAL ASSESSMENT OF ADULTS

Stephen N. Haynes

Behavioral assessment is a powerful, evolving psychological assessment paradigm.[1] The assumptions and methods of behavioral assessment are congruent with cognitive-behavioral, social-learning, and behavior-analytic construct systems (e.g., Bandura, 1969; Haynes, & O'Brien, 1999; Johnston & Pennypacker, 1993; Nelson & Hayes, 1986; O'Donohue & Krasner, 1995). The paradigm includes diverse methods of assessment, such as naturalistic and analogue observation, self-monitoring, electrophysiological measurement, and behavioral interviews and questionnaires. The paradigm emphasizes minimally inferential constructs and an individualized approach to assessment. Environmental and reciprocal determinism, the quantification of lower-level psychological constructs, and time-series measurement strategies are also important elements.

Assessment is an important component in the behavioral treatment of adult behavior problems. Assessment provides the data to help the clinician identify, measure, and classify behavior problems. It also helps the clinician identify variables that affect and are correlated with a patient's behavior problems. Ultimately, the data acquired in behavioral assessment aid in the selection of intervention goals, the design of intervention programs, and the evaluation of intervention effects.

This chapter focuses on the assessment of adult behavior problems to aid clinical decision making.[2] The first two sections review the status and history of behavioral assessment. The subsequent section presents the objectives and conceptual and methodological assumptions of the behavioral assessment paradigm. Methods of behavioral assessment are then considered. The final sections examine behavioral assessment and clinical judgment. Current trends in behavioral assessment and differences between behavioral and nonbehavioral assessment paradigms are discussed in all sections.

THE STATUS OF BEHAVIORAL ASSESSMENT

Behavioral assessment is a rapidly evolving area of research and application. Indices of its status include the proliferation of books, published articles, symposia, and presentations at scientific conventions that focus on behavioral assessment. Although no books on behavioral assessment were published before the mid-1970s, several volumes were published in the latter part of the 1970s and the 1980s (Barlow, 1981; Bellack & Hersen, 1988; Ciminero, Calhoun, & Adams, 1977, 1986; Cone & Hawkins, 1977; Haynes, 1978; Keefe, Kopel, & Gordon, 1978; Kratochwill & Shapiro, 1988; Mash & Terdal, 1981, 1988; Nay, 1979; Nelson & Hayes, 1986; Ollendick & Hersen, 1983). Several behavioral assessment books have been prepared in the 1990s (Haynes & O'Brien, 1999; Hersen & Bellack, 1998; Ollendick & Hersen, 1993), including one in Spain (Fernandez-Ballestros, 1994).

Two behavioral assessment journals were introduced in 1979—*Behavioral Assessment* and the *Journal of Psychopathology and Behavioral Assessment.* In 1992 Pergamon Press stopped publishing *Behavioral Assessment* and added a "Behavioral Assessment Section" to *Behaviour Research and Therapy.*

Further evidence of the status of behavioral assessment comes from the fact that many graduate-level courses in psychology, education, and rehabilitation focus on behavioral assessment (see "focus on graduate training" series in *The Behavior Therapist*). Piotrowski and Zalewski (1993) reported that behavioral assessment was a required course in 80 Ph.D. and Psy.D clinical psychology doctoral training programs. Over one-third of the program directors expected the emphasis on behavioral assessment to increase in the future (7% expected the emphasis to decrease).[3]

Behavioral assessment methods and concepts are also applied across diverse disciplines, including clinical psychology (Hersen, Kazdin, & Bellack, 1991), behavior analysis (see *Journal of Applied Behavior Analysis*), and behavioral medicine and health psychology (Haynes & Wu-Holt, 1995). The paradigm has also been applied in social work and psychiatry (see *Journal of Behavioral Therapy* and *Experimental Psychiatry*), cognitive psychology (Linscott & DiGiuseppe, 1998), community psychology (see *American Journal of Community Psychology*), rehabilitation and developmental psychology (Kail & Wickes-Nelson, 1993), developmental disorders (Repp & Singh, 1990), and pediatric medicine (Karoly, 1988). Behavioral assessment is also well represented in international journals and books (e.g., *Dutch Journal of Behavior Therapy, European Journal of Psychological Assessment, Psichologia Conductual* [Behavioral Psychology].

A SHORT HISTORY OF BEHAVIORAL ASSESSMENT

The history of behavioral assessment reflects the diversity of its methods (see historical overviews of behavioral paradigms in Alexander & Selesnick, 1966; Kazdin, 1978; Haynes & O' Brien, 1999; McReynolds, 1986; Nelson, 1983). Naturalistic and analogue observations were used in early Pavlovian, Watsonian, and other experimental psychological studies and can be traced to Hellenic and Egyptian eras. Observation, as a method for scientific inquiry, has been adopted and refined by behavior analysts. Methodological refinements to behavioral observation and other assessment procedures have also come from ethology, social psychology, developmental psychology, and experimental psychology (e.g., Bott, 1928; Goodenough, 1928; Parten, 1932; Hutt & Hutt, 1970).

The statistical analysis of the time-series data, often acquired in behavioral assessment (Suen & Ary, 1989), has been influenced by multiple disciplines (Collins & Horn, 1991). Similarly, many advances in the methods of behavioral assessment follow advances in computer technology and in the technology for ambulatory monitoring (e.g., Tryon, 1996b).

Self-report methods of behavioral assessment, such as questionnaires and interviews, have been adapted from disciplines such as educational, developmental, organizational, and personality psychology. The content and focus of many self-report instruments have been refined to increase their methodological and conceptual congruence with the behavioral assessment paradigm. Sometimes, traditional self-report instruments have been adopted by behavioral assessors without refinement (Guevremont & Spiegle, 1990). Questionnaires and interviews that are congruent with the behavioral assessment paradigm focus on narrowly defined variables (e.g., specific behaviors and thoughts). Behavioral, compared to nonbehavioral, questionnaires and interviews also yield data on less inferential variables and often address environmental sources of variance.

Behavioral intervention methods and foci strongly influence behavioral assessment methods. Although behavioral interventions with adult behavior problems were reported in the 1950s and earlier (Kazdin, 1978; e.g., Wolpe, 1958), they were not applied extensively until the 1960s (Bachrach, 1962; Bandura, 1969; Ullmann & Krasner, 1965). These interventions emphasized the manipulation of the patient's motoric and cognitive responses in specific situations. The focus and methods of these interventions required assessment procedures that differed procedurally and conceptually from those used in traditional clinical interventions. The evaluation of the immediate, intermediate, and ultimate outcomes in behavior therapy mandated the use of precisely focused measurement procedures that were sensitive to changes in multiple response modalities in the natural environment and across time.

There has been a reciprocal influence between advances in behavioral assessment concepts and methods. For example, the methods and foci of the behavioral assessment paradigm have been affected by research on stimulus-control factors in sleep problems (Youkilis & Bootzin, 1981), cognitive factors in phobic disorders (Taylor & Agras, 1981), the physiological mechanisms associated with many behavior problems (Gatchel & Blanchard, 1993), behavior chains in child behavior problems (Voeltz & Evans, 1982), temporally noncontiguous events in marital distress (Margolin, 1981), multiple and interactive causal factors for behavior problems (Haynes, 1992; Kazdin, & Kagan, 1994), systems factors in most behavior problems (Kanfer, 1985), the multiple response modes often characteristic of complex behavior and behavior problems (Lang, 1995), the situational specificity in many behavior problems (McFall, 1982), and the dynamic and nonlinear aspects of behavior problems (Burton, 1994; Heiby, 1995).

One impetus for the development of behavioral assessment has been a dissatisfaction with traditional clinical assessment instruments and their underlying assumptions (McFall, 1986; McReynolds, 1986). Traditional assessment methods, such as projective techniques and global personality trait questionnaires, do not provide data that are sufficiently specific or that reflect the conditional nature of behavior problems. In addition, traditional assessment instruments do not provide data on the multiple response modes of behavior problems. The aggregated, global nature of many constructs measured in traditional clinical assessment rendered traditional instruments insufficiently sensitive to changes across time or situations and insufficiently amenable to individualized assessment. Often, the constructs measured were permeated with untestable psychodynamic causal connotations with limited clinical utility.

The failure of clinical psychology to evolve more powerful conceptual models and intervention strategies was attributed, in part, to emphases on unobservable and highly inferential intrapsychic processes and causal factors. These emphases were manifested in a reliance on verbal, insight-oriented psychotherapy, the dependence on psychodynamic assumptions (as in those underlying the extant diagnostic systems, DSM I; DSM II), and the use of assessment instruments to infer unobservable, causally imbued "traits" of persons (Wolman, 1978). Assessment instruments of questionable psychometric qualities were used to provide indices of highly inferential, unobservable, and situationally insensitive intrapsychic phenomena.

THE MULTIPLE OBJECTIVES OF BEHAVIORAL ASSESSMENT

Psychological assessment paradigms differ in their objectives—the purposes for which they are applied. For example, neuropsychological assessment is often used to estimate functional impairment and cognitive functioning for intervention planning. Intellectual (and cognitive) assessment is often used to make decisions about the best educational environment for a child.

The behavioral assessment paradigm can have many objectives. The objectives of behavioral assessment in clinical settings can include: (a) the identification of intervention target behaviors (patient behavioral excesses and deficits), (b) the identification of immediate, intermediate, and ultimate intervention goals, (c) the identification of behaviors that are positive alternatives to unwanted target behaviors, (d) the identification of causal and moderating variables for target behaviors and goals, (e) the development of a functional analysis, (f) the design of intervention strategies, (g) the evaluation of ongoing intervention strategies, (h) the facilitation of positive client-therapist interactions, (i) diagnosis, (j) the identification of therapy process variables that affect treatment outcome, and (k) the measurement of patient satisfaction and compliance.

The multiple objectives of behavioral assessment renders it a *functional approach to assessment.* That is, the strategies of behavioral assessment (and the assessment targets, sampling parameters, etc.) are determined by the objectives of the assessment for each assessment occasion and the clinical judgments that the assessment data are intended to affect. For example, in a smoking treatment program (Shiffman, 1993), self-report questionnaires of "motives" for smoking cessation may be useful when the objective of assessment is to design individualized interventions for patients. However, self-monitoring of smoking (or of serum conicotine and thiocyanate) may be more useful than self-report questionnaires if the objective of assessment is treatment outcome evaluation. The importance of various sources of measurement error (e.g., reactive effects of assessment, biases associated with retrospective reports), and under-

Table 19.1. The Multiple Objectives of Behavioral Assessment

To identify, specify, and measure patient behavior problems
To identify positive alternatives to patients' behavior problems
To identify, specify and measure patients' goals and behavioral assets
To identify causal and noncausal functional relationships for behavior problems and goals
To develop a functional analysis,
To design individualized treatment programs
To evaluate the immediate, intermediate and ultimate main effects of behavioral intervention
To evaluate intervention side-effects (e.g., effects on extended social systems, generalization across behaviors, situations, persons; negative side-effects)
To facilitate positive assessor-patient interactions, informed consent, and positive patient expectancies regarding treatment
To facilitate basic research in behavior analysis, learning, psychopathology, cognitive psychology, developmental psychology, and social psychology
Deemphasized goals: psychiatric diagnosis, "personality" description

lying assumptions of assessment (e.g., the importance of situational factors) also vary across the objectives of assessment (Haynes & Waialae, 1994). Table 1 outlines the objectives of behavioral assessment; several of which are discussed below.

To Specify Patient Behavior Problems

Patients frequently report multiple and vaguely specified behavior problems (Persons, 1989; Walitzer & Connors, 1994). Consequently, two purposes of assessment are to specify the patient's behavior problems and to select those upon which to focus initial intervention efforts (Hawkins, 1986; Nezu & Nezu, 1993; see the mini-series on target behavior selection in *Behavioral Assessment,* 1985, 7, 1-78). Whenever possible, behavior problems are specified in measurable, lower-level (i.e., more molecular, precise) units.

Initial treatment targets are often selected from multiple behavior problems on the basis of their frequency, duration, and magnitude. Other criteria for problem behavior selection include the degree to which the behavior problem affects the patient's quality of life (Evans, 1993), the goals of the patient, the degree of danger to the patient or others presented by the behavior problem, and the probability of successful intervention.

Estimates of *shared variance* affect target behavior selection. Behavior problems differ in the degree to which they covary with (i.e., belongs to the same *empirical or functional response class*; Alessi, 1988; Mash & Terdal, 1988), or serve as a causal variable for, other behavior problems. For example, a patient's marital distress might initially be targeted for treatment if it triggered substance use, depression, or anxiety episodes. Alternatively, depression might initially be targeted if it triggered marital distress and other behavior problems. A child's "noncompliance" could be targeted because it covaries with "stealing." Thus, behaviors whose modification are likely to have the greatest positive influence (i.e., the greatest *magnitude of effect*) on other behaviors are often the most appropriate targets for intervention (see discussion on "functional analysis" later in this chapter). This is the "centrality" of particular behavior problems to the patient's other problems in living (Haynes, 1994; "*keystone behaviors,*" Nelson, 1988)

Estimates of the relative importance of behavior problems can be unstable. These estimates often change across assessment sessions (Mash & Hunsley, 1993). Importance estimates can also change as a result of new information about the patient, changes in the patient's environment, treatment, or from the reactive effects of assessment.

To Specify Patient Positive Goals and Behavioral Skills

The identification of behavioral goals for a patient is an important and sometimes underemphasized objective of behavioral assessment (Karoly, 1993). A goal-oriented approach to assessment encourages time-series measurement of goal attainment, the use of positive reinforce-

ment contingencies, and the identification of behavioral skills that would enable the client to adapt flexibly to changing environmental contexts. A goal-oriented approach can also enhance the acceptability of behavioral interventions to patients, reduce the need to assess low-frequency behaviors, and help reduce problem behaviors (Haynes, 1978). Goal setting also requires the active participation of patients in the assessment-treatment process. Many recently published studies have emphasized the utility of goal setting in behavior therapy (e.g., Halford, Sanders, & Behrens, 1994; James, Thorn, & Williams, 1993; Kahle & Kelley, 1994). As with problem behavior identification, goals should be narrowly defined, precisely specified, and measurable. The specification of immediate and intermediate in addition to ultimate goals in treatment can help in the detection of failing treatment (Mash & Hunsley, 1993).[4]

Another objective of behavioral assessment is the identification of desirable alternatives to undesirable problem behaviors (Goldfried, 1982; Hawkins, 1986). This objective can involve the identification of behaviors that are incompatible with, and therefore reduce the probability of, problem behaviors (e.g., positive communication behaviors that reduce the probability of self-injurious behaviors in adults with developmental disabilities; Durand & Carr, 1991).[5] The undesirable target behavior and the positive alternative may be members of the same response class,[6] in that they may have similar effects on, or similar functional relationships with, environmental variables. For many patients, positive skills (e.g., recreational skills, verbal communication skills) can be strengthened to substitute for undesirable behaviors.

To Identify Causal and Noncausal Functional Relationships for Behavior Problems and Goals

The primary goal of many behavioral intervention programs is to modify the causal variables that affect a patient's behavior problems or goal attainment (Haynes, Spain, & Oliveira, 1993; see "Conceptual and Methodological Foundations" section of this chapter for a discussion of "causality").[7] A major goal of behavioral assessment is to identify the causal factors for a patient's depression episodes, self-injurious behaviors, headaches, classroom inattention, or marital violence so that intervention programs can be designed for their modification.

Models of causality in the behavioral assessment paradigm can be complex (Haynes, 1992; Kazdin & Kagan, 1994). Besides an emphasis on the importance of temporally contiguous antecedent and consequent environmental events, contemporaneous causal models of behavior problems emphasize multiple causal variables. The dimensions (e.g., magnitude, chronicity) of causal variables can differ between patients and vary across time. There can be important differences between patients in the causal factors affecting their behavior problems. Finally, noncontemporaneous causes and systems-level causal variables (e.g., family interactions, occupational stressors) can be important.

Haynes and colleagues (1993) noted that there are several methods of estimating causal relationships for a patient's behavior problems. All involve estimating the *conditional probabilities* for behavior problems or the *magnitude of shared variance* between behavior problems and hypothesized controlling variables. Methods of inferring causal relationships include:

1. Rational derivation from psychopathology research,
2. Use of causally focused self-report instruments (e.g., those that request information about antecedent and consequent events for a behavior problem; functional analytic interviews),
3. Use of *causal marker variables* (i.e., cost-efficient indices of causal relationships, such as laboratory psychophysiology measures of autonomic arousal to estimate the effect of naturally occurring stressors in a patient's blood pressure elevations),
4. *Multivariate time-series regression assessment*—the frequent measurement of behavior problems and hypothesized causal variables and using time- and cross-lagged correlations to infer causal relationships,
5. *Manipulation* of the values of a hypothesized causal variable while values of the behavior problem are measured.

To Develop a Functional Analysis

A functional analysis is "the identification of important, controllable, causal functional relation-

ships applicable to specified behaviors for an individual" (Haynes et al., 1993). The functional analysis is the integration of assessment results into a clinical case conceptualization about a client's problems and all variables functionally related to those problems. A functional analysis includes (a) a client's problem behaviors, (b) multiple interacting behavioral, cognitive, and physiological causal and moderating factors; (c) the effects of behavior problems, (d) noncausal functional relationships, (e) a client's assets and deficits, and (f) situational source of variance and other moderating variables. Beyond identifying the pattern of relationships between these multiple causal factors, the functional analysis also includes estimates of the strength of the relationships and the temporal sequence of various causal factors. The functional analysis is congruent with the idiographic emphasis in behavioral assessment. It is often the primary determinant of problem behavior selection and intervention design. It is also the most complex and least researched aspect of treatment-related clinical judgment. The functional analysis is discussed in greater detail in a subsequent section of this chapter.

To Design Behavioral Intervention Strategies

Therapy paradigms differ in the degree to which they emphasize the importance of preintervention assessment for the design of clinical intervention programs. The degree to which a therapy paradigm emphasizes preintervention assessment depends on several factors. First, it depends on the degree of presumed individual differences in the characteristics and determinants of behavior problems, both within and between classes of behavior disorders (e.g., the degree to which persons with a diagnosis of "Major Depressive Disorder" differ in the causes and symptoms of depression and the degree to which they differ from persons with other diagnoses). The importance of preintervention assessment also depends on the diversity of presumed mechanisms of change in therapy. For example, if the primary mechanism of change is presumed to be "a supportive client-therapist relationship," preintervention assessment would have little impact on intervention strategy. Finally, it depends on the diversity of available intervention strategies for a particular behavior problem—if all patients will receive the same type of treatment, pretreatment assessment will have little affect on treatment decisions.

Behavior therapy is unique in several ways that aggregate to enhance the importance of preintervention assessment (Acierno et al., 1994; Cone, 1986; Persons, 1992). Behavior therapy recognizes multiple possible causal factors for behavior problems and between-person differences in the causes of behavior problems. Behavior therapy also acknowledges multiple mechanisms of treatment-related change and includes many intervention strategies.

The design of intervention strategies for a patient is strongly influenced by the functional analysis for that patient. However, intervention decisions are also influenced by other factors: (a) the goals of the patient, (b) the skills and resources of the therapist, (c) the availability and relative cost effectiveness of interventions, (d) the potential side-effects of interventions, (e) the patient's social supports and other social and personal moderating variables (e.g., cognitive functioning, physical impairment), and (f) the relative acceptability and social validity of available interventions (Haynes, 1986).

To Evaluate the Effects of Behavioral Intervention

Behavioral assessment is frequently used to evaluate intervention effects. The systematic, quantitative evaluation of intervention effects has four benefits: (a) It provides data on the validity of the functional analysis upon which the intervention is based, (b) It facilitates the detection and modification of failing intervention programs, (c) It advances clinical science by adding to the database on intervention effects, and (d) It is congruent with an increasing emphasis on treatment accountability.

Evaluation of intervention outcome is consistent with the behavioral assessment paradigm's emphasis on hypothesis-testing and time-series assessment. Behavior therapists often begin treatments based on a hypothesized functional analysis of a patient. The validity of the functional analysis is estimated by the degree to which behaviors change in predicted ways, presuming that treatments are validly implemented

The power and clinical utility of intervention outcome evaluation is strengthened to the degree that measurement occurs in a time-series format

(multiple measures across time), involves multiple instruments, and focuses on multiple modes of behavior problems. Inferences about treatment outcome are also strengthened when assessment focuses on multiple behavior problems, multiple modes and parameters of behavior problems, possible side-effects of treatment (i.e., effects of intervention other than those that involve the primary target behaviors), and generalized effects of treatment across situations, behaviors, and persons. It is also strengthened by the use of data from multiple sources (e.g., spouses, staff members, patient), the measurement of lower-level variables, and the use of validated assessment instruments. Inferences about mechanisms underlying treatment effects can also be strengthened by the measurement of independent variables (e.g., compliance with treatment prescriptions, behaviors of change agents) and systems-level behaviors.

Time-series assessment across the course of clinical intervention aids in the early detection and modification of failing treatments. The early detection of failing interventions may be facilitated by precisely measuring approximations to immediate, intermediate, and final treatment goals (Mash & Hunsley, 1993). Principles of intervention outcome evaluation also depend on the research design within which the treatments are applied. Principles of intervention outcome evaluation have been discussed in greater detail in several books (Kazdin, 1992; Hersen & Bellack, 1998; Kratochwill & Levin, 1992).

Additional Objectives of Behavioral Assessment

Most of the objectives of behavioral assessment presented in this chapter are relevant to the design and evaluation of behavioral interventions. Behavioral assessment can have several other objectives. One goal of behavioral assessment is to maintain a positive client-assessor ambience. Patients' attitudes toward the assessor and the assessment-intervention process, and patients' beliefs regarding the underlying rationale for the assessment-treatment process and expected benefits can affect their cooperation, the amount and validity of assessment data acquired, and the probability of successful intervention. In an aversive or conflictive patient-assessor assessment environment, a patient may stop cooperating, thus rendering useless any data acquired.

Congruent with the emphasis on a positive professional ambience, one objective of behavioral assessment is to *facilitate positive patient-therapist interactions*—an often overlooked objective of the assessment process. The major vehicle for establishing a positive patient-assessor relationship is the assessment interview (see discussions in Haynes & Chavez, 1983; Hersen & Turner, 1994; Sarwer & Sayers, 1998; Turk & Salovey, 1988; Turkat, 1986). Although the variables affecting the relationship ambience remain to be articulated, many authors have stressed that the assessor should strive for positive, supportive, empathic, respectful, and informative interactions.

Assessment and treatment should proceed only with the *informed consent* of the client or responsible person (e.g., McConaghy, 1998; Wincze & Carey, 1991). Patients should understand the rationale, methods, and expectations of the assessment process (e.g., how many hours of interviews, the focus of the interviews, the content and duration of role-playing).[8]

Summary

Behavioral assessment has multiple objectives, which vary across assessment occasions. Behavioral assessment is a functional approach to assessment in that the objectives extant for a particular assessment occasion determine the best methods and focus of assessment. Most of the assessment objectives discussed here pertain to the development of a functional analysis and the evaluation of treatment outcome. Other objectives include the acquisition of data for differential diagnosis, basic research in psychology, the collection of epidemiologic data, and for assessment instrument development and validation.

CONCEPTUAL AND METHODOLOGICAL FOUNDATIONS OF BEHAVIORAL ASSESSMENT

The objectives, methods, and foci of psychological assessment paradigms are influenced by several related sets of assumptions regarding: (a) the characteristics of behavior problems, (b) the best measurement strategies in clinical assessment, and (c) the causal relationships for behavior problems and goals. The conceptual and methodological foundations of behavioral assessment are outlined

Table 19.2. Conceptual Foundations of the Behavioral Assessment Paradigm

Assumptions About The Characteristics of Behavior Problems
Behavior problems have *multiple response modes* (motoric, verbal, physiological, cognitive) that sometimes evidence *low levels of covariation*
Behavior problems and their modes have *multiple parameters* (onset, duration, magnitude, rate)
There are *between-person differences* in the importance of the individual modes and parameters of a behavior problem
There can be complex *interrelationships* (e.g., noncausal covarying, bidirectional causal) among multiple behavior problems
Behavior problems are *conditional*; they can vary systematically across situations, time, states, and contexts
Assumptions About the Causal Relationships for Behavior Problems and Goals
Multiple causality—Behavior is affected by multiple, interacting causal factors
There are *multiple parameters* of causal variables (onset, duration, magnitude, rate), which can differ in causal effects
There can be important *between-person differences* in the strength of causal factors for the same behavior problems
Causal relationships are *dynamic* (they can change across time)
Social/environmental response contingencies and antecedent stimuli can be particularly important causal factors
Reciprocal determinism: Person-environment interactions are important causal mechanisms (i.e., *bidirectional causality*; reciprocal causation)
Contemporaneous causal factors may be more important, or have greater clinical utility, than historical causal factors
Systems-level factors (family, work, marital, other interpersonal factors) may serve important causal functions for a patient
There can be significant variance in causal factors *across situations*
Causal factors may have *differential effects* on the different parameters and modes of behavior problems
Casual relationships can demonstrate *nonlinear* (e.g., plateau, critical level) functions

in Tables 2 and 3.[9] Several of these assumptions and their impact on methods of assessment are discussed in the sections that follow.

Assumptions About the Characteristics of Behavior Problems

Five assumptions about the characteristics of behavior problems were outlined in a previous section and strongly affect behavioral assessment strategies: (a) Behavior problems can have motoric, cognitive, and physiological modes; (b) Behavior problem modes can have multiple parameters (onset, duration, magnitude, rate); (c) There are between-person differences in the importance of the individual modes and parameters of a behavior problem; (d) There can be complex interrelationships among multiple behavior problems for a patient; and (e) Behavior problems are conditional

Many studies have documented the *multiple response modes* of adult behavior problems. For example, anxiety disorders may involve overt avoidance or escape from anxiety-arousing situations, subjective distress, physiological arousal, catastrophic thoughts, and worry (Last & Hersen, 1988). While some persons show a high degree of covariance among multiple modes of a behavior problem, others show low levels of covariation. For example, some patients with an anxiety disorder experience considerable subjective distress without avoiding anxiety provoking situations and other patients evidence significant physiological arousal without any catastrophic thoughts or negative beliefs. Such differences are suggested by low-to-moderate levels of covariation between many modes of many behavior problems in nomothetic and time-series research (i.e., the modes can be *discordant* for groups of persons and for one person across time).[10] In addition, different response modes of a behavior problem may demonstrate different time courses, be affected by different causal factors, respond differently to the same causal factor, and respond differently to an intervention.

The multiple response modes of behavior problems have two important implications for assessment. First, a diagnosis, (e.g., a DSM-IV

diagnosis; APA, 1994), is insufficient to draw inferences regarding the most important response modes for a patient. Second, data regarding one mode of a behavior problem for a patient is insufficient to draw inferences regarding another mode. Consequently, *multimodal assessment* is an important objective of behavioral assessment (Carey, Lantinga, & Krauss, 1994; Turk & Melzack, 1992; Weiss & Heyman, 1990).

Behavior problem modes can have *multiple parameters* or *dimensions* (such as onset, duration, magnitude, and rate). Multiple parameters are important for two reasons. First, there are between-patient differences in the relative importance of different behavior problem parameters. For example, patients with the same depressive disorder diagnosis may report frequent but brief depressive symptoms or infrequent but long-lasting depressive symptoms.

Second, the effects of causal variables can vary across the parameters of a behavior disorder. For example, Barnett and Gotlib (1988) suggested that learned helplessness beliefs are more likely to affect the duration and magnitude of depressive behaviors rather than the onset of depressive behaviors. Also, a variable that triggers a patient's paranoid delusions may not affect the duration or magnitude of the delusions (Haynes, 1986).

Between-person differences in the importance of behavior problem parameters mandate careful specification and measurement of multiple parameters. Different interventions are likely to be selected and the effects of interventions are likely to differ, depending on which parameter is most important for a patient. Consequently, measures of behavior problems that aggregate across parameters (e.g., a single measure of "depression" or "anxiety" that includes many elements of those constructs) will often be insufficiently precise for the development of a functional analysis and for the design of intervention programs.

Most patients present *multiple, interdependent behavior problems.* For example, high incidences of comorbidity are often found for drug use (Regier et al., 1990), panic disorders (Craske, & Waikar, 1994), depression (Persons & Fresco, 1998), and posttraumatic stress disorder (Figley, 1979). Each of a patient's behavior problems can be the result of different causal variables. Additionally, multiple behavior problems can have both causal and noncausal functional relationships. For example, marital distress can be both an effect and a cause of depression for some patients (Beach, Sandeen, & O'Leary, 1990).

The fact that patients frequently report multiple behavior problems explains the emphasis in behavioral assessment on identifying the form and strength of relationships among the behavior problems, and estimating the relative importance of the behavior problems.[11] These inferences are important components of the functional analysis and affect decisions about where to focus interventions. The identification of functional response classes is also an important objective because behavioral interventions often attempt to substitute less problematic for more problematic behaviors in the same response class (e.g., teaching relaxation skills as a means of reducing anxiety-associated binge eating).

The assumption that behavior problems are often *conditional* is an important element of the behavioral assessment paradigm. Many studies have indicated that the parameters of behavior problems can vary across settings and as a function of antecedent and discriminative stimuli (Gatchel, 1993; Glass, 1993; Ollendick & Hersen, 1993). The conditional nature of behavior problems is important because identifying environmental or personal (e.g., physiological state, thoughts) sources of behavior variance can lead to the identification of causal variables and causal mechanisms for those behaviors. Consider the potential impact on causal inferences and intervention strategies of identifying specific triggers of a patient's asthma episodes (Creer & Bender, 1993), the specific social situations that precipitate a patient's panic episodes (Craske & Waikar, 1994), or the antecedents of marital violence in a family (O'Leary, Vivian, & Malone, 1992).

The potential for cross-situation variability in the parameters of behavior problems strengthens the inference that aggregated measures of a behavior problem are often insufficiently precise for a functional analysis. Assessment instruments should allow the assessor to examine the conditional probabilities of behavior problems or the magnitude of shared variance between the behavior problem and multiple situational factors. Assessment methods such as functionally oriented structured interviews, self-monitoring, situation-specific questionnaires, and observation are all conducive to gathering data about the conditional nature of behavior problems.

The parameters and other characteristics of behavior problems are *dynamic* (i.e., they change

over time).[12] The magnitude, frequency, duration, and form of a patient's depression or headaches, a psychiatric patient's delusions, the blood pressure of a hypertensive patient, the pain reports of a chronic pain patient, the activity rate of hyperactive children, the sleep patterns of an insomniac, or the caloric intake of an anorectic can change in important ways across time. Sensitive measurement of dynamic variables requires the frequent measurement of lower-level variables, using time-sampling assessment strategies. Recent books by Collins and Horn (1991), Heatherton and Weinberger (1994), Kazdin (1992), and Kratochwill and Levin (1992) discuss strategies and issues in the measurement of dynamic variables.

Assumptions About the Causes of Behavior Disorders

Assumptions about the causes of behavior problems vary across psychological assessment paradigms and greatly affect their methods and foci. Causal assumptions are particularly influential in the behavioral assessment paradigm because behavioral interventions are often designed to modify the variables hypothesized to control variance in the targeted behavior problem. Therefore, the identification of potential causal variables is a primary objective in pretreatment behavioral assessment.

There are several underlying and empirically based assumptions about the nature of causal relationships associated with the behavioral assessment paradigm. These assumptions are outlined in Table 2 and are discussed in greater depth in Haynes (1992) and Haynes and Wu-Holt (1995). First, a behavior problem can result from multiple permutations of multiple causal variables. Also, multiple causal factors can act concurrently, interactively, and additively (Kazdin & Kagan, 1994). Multivariate, interactive, and aggregated causal factors have been proposed for schizophrenia, chronic pain, sleep disorders, paranoia, personality disorders, child abuse, and many other behavior disorders (see reviews in Gatchel & Blanchard, 1993; Sutker & Adams, 1993). For example, sexual dysfunctions (e.g., male erectile dysfunctions, dyspareunia) can result from many permutations of physiological dysfunctions (e.g., diabetes, hormonal dysfunctions), attention processes, environmental contexts, relationship distress, and conditioned fear reactions.

A causal variable may also affect a behavior problem through *multiple paths*. For example, chronic life stressors may impair a patient's immune system functioning (Asterita, 1985), through increased drug use, dietary changes, reduction of lymphocyte levels, reduced production of interferon, and sleep disruption. Similarly, social isolation may increase the risk of depression through many paths: by restricting the potential sources of social reinforcement, by increasing dependency on reinforcement from a few persons, and by reducing the rate of socially mediated reinforcers. Social isolation may also prevent the development of social support networks to buffer the effects of negative life events, reduce physical activity level, and increase the chance and duration of negative ruminations.

The array of operating causal variables and mechanisms can *differ across patients* with the same behavior problem. For example, the self-injurious behaviors of one developmentally disabled patient can result from positive social reinforcement while the identical behavior for another patient can result from negative social reinforcement (e.g., escape from a parent or staff member), escape from aversive tasks or demands (e.g., escape from a classroom situation), or self-reinforcement from the behavior (Iwata et al., 1994).

Causal variables, relationships, and mechanisms associated with a patient's behavior problem can be, dynamic and nonstationary (Haynes, Blaine, & Meyer, 1995). Causal relationships can change across time in several ways: (a) New causal variables may appear (e.g., new health problems of a family member; meeting a new friend); (b) The magnitude and direction of effect of a causal variable can change (e.g., a decrease over time in sleep disruption caused by a traumatic event); (c) A causal variable may disappear (e.g., a coercive supervisor is transferred); (d) The temporal parameters or form of a causal relationships may change (e.g., Burish, Carey, Krozely, & Greco, 1987, found that the latency for and duration of anticipatory nausea reactions associated with chemotherapy for cancer patients decreased for many patients as the number of chemotherapy sessions increased); and (e) Changes may occur in moderating variables (e.g., a change in a patient's expectancies about the beneficial effects of alcohol may change the probability that a patient will drink alcohol in response to a life stressor; Smith, 1994).[13]

Emphases on multivariate, idiosyncratic, and dynamic causal models have several implications for behavioral assessment strategies. First, a broad-spectrum pretreatment assessment is necessary to identify the causal variables that operate for a particular patient. Specification of the behavior problem, or a psychiatric diagnosis, can point to potential causal factors but will often be insufficient for a functional analysis because there are many possible causal factors for most behavior disorders. Second, assessors should avoid "premature" or "biased" presumptions of causal relationships for a patient (Haynes, 1994; Nezu & Nezu, 1989; Turk & Salovey, 1988)—the assessor should collect data, then draw inferences about possible causal factors. Third, multiple causal variables operating for a particular patient are most likely to be identified though the use of multi-modal, multi-method, and multi-source assessment. Fourth, it is important to identify *causal mechanisms*: The assessor should frequently ask "How (in what manner? In what way?) does an identified causal variable affect the behavior problem?" (see also, discussion in Shadish, 1996). Fifth, potential causal factors should be measured frequently, using time-sampling assessment strategies. Finally, interactions among causal factors should be assessed (e.g., interactions between life stressors and causal attributions, reinforcement contingencies and task demands).

The behavioral assessment paradigm also stresses the importance of environmental causal factors—more precisely, the importance of behavior-environment interactions (McFall & McDonel, 1986). Many studies have indicated that variance in many behavior problems can partially be accounted for by variance in person-environment interactions, response contingencies, situational and antecedent stimulus factors, and other aspects of learning (see discussions in Eysenck & Martin, 1987; O'Donohue & Krasner, 1995). Spouse, parent, teacher, friend, and staff responses can greatly affect self-injurious behaviors, social anxiety, mood, pain behaviors, substance use, medication compliance, delusional talk, aggressive behaviors, causal attributions, sleep, and many other behaviors.

An important element of environmental causation is *reciprocal determinism* (i.e., *bidirectional causality, reciprocal causation*; Bandura, 1981): The occurrence, type, form, magnitude, and duration of a client's behavior problems can be influenced by environmental and other events which are, in turn, affected by the behavior of the client (McFall & McDonel, 1986). For example, a client's depressive behaviors may cause others to withdraw with the client, further strengthening the client's depressive behaviors. A patient's drug use may place him or her in social situations that increase the chance that he or she will ingest drugs. Parents and children exert reciprocal influences on each other that may exacerbate family conflict. A patient may behave in ways that trigger negative social reactions that, in turn, trigger paranoid thoughts and mood. The assumption of reciprocal determinism promotes a view of a client as an active participant in his or her environment and an active participant in the assessment-therapy process. An important goal of assessment is to identify ways that the patient may be contributing to his or her behavior problems or can contribute to goal attainment in therapy.

The concept of reciprocal determinism renders arbitrary the application of labels such as "dependent variable," "causal variable," "independent variable," and "target behavior." Both variables in a bidirectional relationship can be considered either a target behavior or a causal variable. Label selection depends more on the intent of the assessor than on the functional relationships involving the labeled variables.

A *behavioral skills* focus is an important consequence of an emphasis on reciprocal determinism. It is assumed that a patient's behavioral repertoire (e.g., excesses, deficits, topography, content, timing) affects the probability, type, or degree of behavior disorders. For example, a behavior skills assessment might focus on specific skills that are necessary for a socially isolated individual to form more frequent and satisfying friendships (similar to a task analysis). One set of skills often targeted by behavioral assessors is *cognitions*—a patient's beliefs, expectancies, and other thoughts regarding his or her capabilities in specific situations (e.g., Linscott & DiGiuseppe, 1998). Another, mentioned earlier, is *adaptive flexibility*—behavior repertoires that allow adaptive functioning to a variety of and changing environments.

The behavioral assessment paradigm emphasizes the importance of *contemporaneous*, more than historical, *causal factors*. Behavioral assessors presume that a greater and more clinically useful proportion of variance in the parameters of behavior problems is attributable to recent or contemporaneous, rather than historical, behavior-environmental variables. Several lines of evidence highlight the importance of contemporaneous

causal factors: (a) laboratory evidence in applied behavior analysis showing that the manipulation of immediate response contingencies can dramatically affect the rates of many behaviors; (b) the amenability of temporally contiguous events to manipulation; and (c) the demonstrated effectiveness of several clinical interventions based on the manipulation of current environmental events (e.g., desensitization, role-playing, exposure therapies). For example, while early learning undoubtedly contributes to the development of paranoid ideation (Haynes, 1986), contemporaneous causal variables for paranoid ideation (e.g., a restricted social network, social skills deficits, negative scanning, or failure to consider alternative explanations for ambiguous events) are more amenable to intervention.

The emphasis on contemporaneous and bidirectional behavior-environment interactions dictates an emphasis on particular methods of assessment. For example, naturalistic observation, analogue observation, and self-monitoring are well suited to measuring contemporaneous, reciprocal dyadic interactions. Additionally, in behaviorally oriented interviews and questionnaires patients are often asked about current behavior-environment interactions (Jensen & Haynes, 1986; Sarwer & Sayers, 1998).

The emphasis on extant environmental events and behaviors does not preclude a causal role for genetic, physiological, or early learning experiences. There is strong evidence that genetic/organic factors have an important influence on many behavior problems (see review in Asterita, 1985). Historical learning experiences or events, particularly traumatic ones, have also been shown to be strong risk factors for many behavior problems and to mediate the response to contemporaneous environmental risk factors (e.g., Yoshikawa, 1994, for delinquency behaviors). There are also differences among behavioral assessors in their emphasis on historical data. Joseph Wolpe, for example, emphasized the importance of gathering a complete clinical history for patients before therapy (Wolpe & Turkat, 1985).

As noted in preceding sections, one assumption of the behavioral assessment paradigm is that behavior problems are *conditional*: an important proportion of the variance in behavior can often be accounted for by variance in *situational stimuli* (e.g., discrete and compound antecedent stimuli, contexts, discriminative stimuli for differential reinforcement contingencies).

A situational model of behavior problems contrasts with personality trait models, which emphasize cross-situational consistency of behavior.

Personalty trait concepts can sometimes be useful because they point to an array of possible causal factors. However, they are often excessively molar, poorly defined, faddish, and associated with superfluous connotations (e.g., "codependency," "hardiness"). Personality trait concepts also sometimes confound explanatory and descriptive constructs. They can be difficult to measure and can be incompatible with an idiographic approach to assessment. Finally, personality trait concepts are often imbued with unwarranted causal properties and seldom identify the conditional probabilities and dynamic aspects of behavior (Haynes, Uchigakiuchi, et al., 1993; see discussions in Heatherton & Weinberger, 1994).

Situational and trait models of behavior variance are not necessarily incompatible. If we want to predict or change a patient's behavior it usually helps to know something about the robust behaviors of the person (e.g., unconditional probabilities of specific behaviors) and something about the situational factors and conditional behaviors (e.g., conditional probabilities of specific behaviors). The interactional perspective is a welcomed refinement of exclusively trait models (see discussions by Mischel, 1968; McFall & McDonel, 1986).

The issue of cross-situational consistency of behavior is further complicated because the degree of situational control varies across behaviors, individuals, and situations. The *person x situation interactive model* of behavior variance suggests that relative cross-situational behavior stability can, but may not necessarily, occur. Therefore, the assessor must evaluate the conditional nature of important behaviors, although methods for the selection or classification of situations have not been developed (Schlundt & McFall, 1987). The assessor must identify the specific conditions associated with differential probabilities or varying magnitudes and durations of panic episodes, headaches, alcohol ingestion, sleep disruption, subjective pain, intrusive thoughts, relapse, delusions, medication compliance, and spouse battering.

Although the behavioral assessment paradigm emphasizes immediate stimulus-response associations,[15] *extended social systems* can also affect behavior problems, affect the patient's treatment progress, and are an important assessment focus (Nezu & Nezu, 1993). A patient's behavior prob-

Table 19.3. Methodological Foundations of the Behavioral Assessment Paradigm

The Strategies of Behavioral Assessment
Idiographic assessment; a focus on the client's specific goals, individual behavior problems often used in conjunction with nomothetic assessment instruments
The use of *time-series measurement strategies* (as opposed to single-point or pre-post measurement strategies)
An *hypothesis-testing* approach to assessment and treatment
Assessment of *multiple variables* and use of *multiple measures*
Multisource assessment—use of *multiple methods* and *multiple informants*
Measurement of variables across *multiple situations*
The use of assessment instruments that *valid* for the client and assessment goal
Informed consent and active participation by clients
The Focus of Behavioral Assessment Strategies
Less inferential more specific constructs and measures
Observable behavior (as opposed to hypothesized intrapsychic events)
The measurement of behavior in the *natural environment*
The measurement of *multiple social systems* (e.g., martial, family, work) relevant to the client
An emphasis on *behavior-environment interactions,* particularly response contingencies
Proximal, temporally contiguous events

lems and goals cannot be viewed independently from the context of his or her social environment. For example, a couple's financial status, level and type of social support from family members, culture and ethnic identity, and extramarital relationships can all affect their marital interactions and level of marital distress (Stuart, 1980). Similarly, a parent's family and work environment can affect the way in which he or she attends to and interacts with a child (Wahler & Dumas, 1989) and conflictive interpersonal relationships can impede the treatment of a patient's panic episodes (Craske & Waikar, 1994).

The emphasis of the behavioral assessment paradigm on extended social systems is also consistent with *chaos theory* and *dynamical modeling*. It may be difficult to develop powerful predictive models by measuring behavior independent of the complex dynamical systems in which the behavior is imbedded. Consequently, one task of the assessor is to evaluate the degree to which extended community, family, marital, and other interpersonal factors affect a patient's behavior problems, goals, and treatment effects. The bi-directional causal relationships and covariance among behavior problems mandate an assessment focus on *behavioral systems*. Assessment efforts cannot be confined to individual elements extracted from a complex array of interacting variables. Later sections of this chapter discuss the use of the clinical pathogenesis map and the functional analytic causal models to help the clinician in integrating such complex data.

Summary

The objectives, methods, and foci of the behavioral assessment paradigm are influenced by assumptions about the characteristics and causes of behavior problems and assumptions about the best strategies for clinical assessment. The behavioral assessment paradigm includes several assumptions about the characteristics of behavior problems: (a) Behavior problems can have multiple modes and parameters; (b) There are between-person differences in the importance of the individual modes and parameters of the same behavior problem; (c) There can be different forms of relationships among multiple behavior problems for a patient; and (d) Behavior problems can vary across situations and time.

The behavioral assessment paradigm includes several assumptions about the causes of behav-

ior problems: (a) There are multiple possible causes for most behavior problems; (b) A patient's behavior problem may result from multiple causal factors that act independently, interactively, or additively; (c) A causal variable may affect a behavior problem through multiple paths; (d) Causal variables and mechanisms can differ across patients with the same behavior problem; (e) Causal relationships are unstable; (e) Many behavior problems can be the result of environmental and reciprocal determinism; (f) Contemporaneous causal factors can be more important, or more clinically useful, than historical causal factors; and (g) Important causal relationships can reside in extended social systems and in situational and contextual factors.

MEASUREMENT STRATEGIES AND THE METHODS OF BEHAVIORAL ASSESSMENT

Measurement Strategies

Many methodological elements of the behavioral assessment paradigm were introduced in the previous sections of this chapter. These and other elements are delineated in Table 3. In this section I draw attention to two of these elements: (a) an emphasis on empirical hypothesis-testing and (b) the use of time-series assessment strategies.

During preintervention assessment, the assessor forms many hypotheses (i.e., clinical judgments) about the patient. The assessor forms hypotheses about the relative importance of and relationships among the patient's behavior problems and goals, the variables that influence problems and goals, and the causal role of systems-level factors. The assessor also develops judgments about the best methods of intervention based on the case formulation (e.g., Eels, 1996; Nezu et al., 1996; O'Brien & Haynes, 1995; Persons & Bertagnolli, 1994). These hypotheses, most of which are components of the functional analysis, are tested and refined as assessment continues and during subsequent intervention process.

Hypothesis development, evaluation, and refinement require an empirically based, skeptical attitude toward the results of assessment. Clinical judgments are intrinsically subjective and can be evaluated and refined in several ways: (a) by careful specification and measurement of dependent and independent variables, (b) by using minimally inferential constructs and assessment instruments of known psychometric properties, (c) by using multiple sources of information, (d) by careful control of measurement conditions, and (e) by a receptiveness to disconfirmatory data.

Time-series measurement (i.e., the frequent (e.g., > 40) measurement of independent and dependent variables across time) is a powerful strategy for estimating causal relationships in behavior problems, particularly with dynamic variables of individual patients in their natural environment. Time-series measurement can provide an estimate of causal relationships for a client's behavior problems, such as the factors associated with the frequent onset of a patient's depressed mood or the factors associated with the duration of a patient's delusional statements. It is also the best strategy for tracking the time course of unstable behavior problems, goal-directed behaviors, and intervention effects. Finally, time-series measurement is an essential element in interrupted time-series designs, such as the A-B-A-B, multiple baseline, or changing criterion designs (Kazdin, 1992). These designs are useful for strengthening the internal validity of inferences about treatment effects and mechanisms.

The renewed emphasis in applied psychology on the intensive longitudinal study of patients under controlled conditions using precise lower-level measures has been a major contribution of behavioral assessment, behavior therapy, and behavior analysis paradigms. This empirical approach to assessment has accentuated the importance of professional accountability and is suited to the evaluation of service delivery and intervention outcome.

An overzealous adoption of *methodological empiricism* in assessment, however, can have negative ramifications for a construct system. Excessively molecular measures can involve trivial quantification of trivial variables with little social or practical importance. Exaggerated attempts at quantification can demean the behavioral assessment paradigm and contribute to the belief that it often focuses on trivial events, removed from their contexts. The growing recognition of the practical and clinical importance of causal relationships and treatment effects is reflected in recently published articles on *clinical significance* (e.g., Jacobson & Truax, 1991).

An excessive reliance on quantification can also reduce the creativity of psychological inquiry and impair the evolution of an assessment paradigm.

Our understanding of functional relationships among behaviors and environmental events is elementary. A paradigm that is in an early stage of scientific development requires adherents to be open to new concepts and relationships. Although scientific empiricism is the only method of evaluating hypotheses, and while the close examination of data can promote new hypotheses, many ideas are generated from qualitative observations of phenomena. By supplementing quantitative with qualitative analyses, behavior assessors can generate creative and potentially useful clinically hypothesis.

While acknowledging an important role for qualitatively based inferences, the importance of empiricism in behavioral assessment cannot be understated. Psychological construct systems must ultimately be based on quantitatively based methods of inquiry. Gestalt, transactional, person-centered, Adlerian, and most psychoanalytic construct systems have remained essentially unchanged for decades because they are defined by rigidly invoked assumptions about causality, rather than by methods of inquiry that encourage examination and refinement of hypotheses. In contrast, the rapid evolution of behavioral paradigms, the enhanced power and utility of behavioral assessment methods, and the expanding array of available behavioral intervention strategies can be attributed to an emphasis on a set of methods for studying behavior, rather than an emphasis on a prescribed set of concepts about the causes of behavior and the best methods of modifying behavior.

The Domain of Methods of Behavioral Assessment

Although the conceptual and methodological foundations of behavioral assessment have been well articulated by many assessment scholars, the methods used by behavioral assessors in clinical situations are becoming more inclusive; the boundary between behavioral and nonbehavioral assessment methods is becomingly increasingly fuzzy (e.g., Haynes & O'Brien, 1999; Hersen & Bellack, 1998). Many recent behaviorally oriented books and articles discuss assessment instruments that have not been traditionally associated with a behavioral construct system. Furthermore, some of these assessment instruments are not congruent with the conceptual elements of the behavioral assessment paradigm. For example, many cognitive assessment instruments are nomothetically developed and trait-based (they do not address the conditional nature of expectancies, beliefs, cognitive processes; Linscott & DiGiuseppe, 1998). Assessment methods recently used by behavior analysts include neuropsychological assessment, thought listing, videotape reconstruction, sociometric status evaluation, a projective method involving interpretation of patients' stories, trait-based personality tests such as locus of-control scales and the MMPI, aggregated mood scales, historically focused interviews, and tests of academic achievement (see Hersen & Bellack, 1998).

The permeable boundary between behavioral and nonbehavioral assessment methods has several roots. First, behavioral assessors are more frequently focusing on variables (e.g., beliefs, expectancies, mood) excluded from earlier, operantly influenced behavioral paradigms. This expanded focus has required the use of an expanded array of assessment instruments, many of which provide indices of molar, aggregated, and more inferential constructs. Second, there has been moderation in the tendency of behavioral assessors to automatically reject any assessment procedure identified with traditional clinical psychology. This former bias, often warranted, has been replaced with a more reasoned appraisal of the applicability and psychometric qualities of traditional assessment instruments. Third, an early exclusive emphasis on situational control of behavior has been replaced by the person x situation interactional model discussed earlier. This conceptual refinement has led to the use of some trait-based assessment instruments that are insensitive to situational sources of variance. Fourth, although many behavioral assessment methods are powerful, their clinical utility is often hindered by cost-effectiveness considerations.

A possible, and more troubling source of the inclusiveness in behavioral assessment methods is that some behavioral assessors are insufficiently educated in the conceptual and methodological aspects of the behavioral assessment paradigm. For example, behavioral assessors often do not acknowledge the assumptions inherent in the use of an assessment instrument that provides an aggregated "score" of a construct with multiple facets (e.g., "depression," "extroversion"), or of a construct that displays important between-situation variance. Also, many norm-referenced assessment instruments are applied without consideration of the degree to which they are appropriate or useful for assess-

ing the client. Similarly, there are frequently unacknowledged inferential problems in the use of an assessment instrument that provides a global or indirect measure of a construct (e.g., measures of "irrational beliefs" or "cognitive distortion") or in use of an assessment instrument with psychometric properties that have not been established for the target population or purposes to which it is applied (Cone, 1998; Haynes & O'Brien, 1999; Silva, 1993). Finally, behavioral assessors sometimes do not acknowledge the conceptual difficulties in interpreting data from assessment instruments developed from conceptual frameworks that are incompatible with a behavioral construct system. Insufficient conceptual sophistication can ultimately threaten the evolution, power, and empirical rigor of the behavioral assessment paradigm.

Behavioral assessment methods are discussed below in four overlapping categories: (a) behavioral observation, (b) self-monitoring, (c) self-report instruments, and (d) psychophysiology. The specific strategies, conceptual foundations, utility, psychometric properties, and contribution to clinical judgment of each method will be discussed. More extensive discussions of behavioral assessment methods and instruments can be found in books by Haynes & O'Brien, 1999; Hersen and Bellack (1998), Mash and Terdal (1997), Ollendick and Hersen (1993), and Shapiro and Kratochwill (1988).

Behavioral Observation

Behavioral observation is the assessment method most strongly associated with the behavioral assessment paradigm. It involves the acquisition of quantitative, time-series data on molecular, well-defined behaviors and environmental events (Foster & Cone, 1986; Foster, Bell-Dolan, & Burge, 1988; Mash & Hunsley, 1990; Tryon, 1998b). Two observation strategies are discussed below: behavioral observation in the natural environment and behavioral observation in analogue environments.

Behavioral Observation in the Natural Environment

The assessment method that is most congruent with the underlying assumptions of the behavioral assessment paradigm is observation in the patient's natural environment using nonparticipant observers (observers who are not normally part of the natural environment). Observation in natural environments has been used in the assessment of marital and family interactions in the home, student and teacher behaviors in schools, pain and other health-related behaviors at home and in medical centers, eating and drinking in a variety of settings, autistic, self-injurious, delusional, and hallucinatory behaviors in inpatient institutions, and community behaviors (e.g., driving, littering), among many others. It is used most often for treatment outcome evaluation but has also been used to gather data for functional analysis and in the basic behavioral and social sciences. Observation in the natural environment is least applicable for the assessment of very low-frequency behaviors (e.g., stealing), covert behaviors (e.g., mood), and socially sensitive (reactive) behaviors (e.g., sexual behaviors, marital violence).

Typically, one or two observers enter the patient's natural environment several times on a predetermined schedule and record the occurrence of preselected and carefully defined behaviors. Each observation session is usually divided into smaller *time-sampling* periods (e.g., 10-, 15-, or 30-second periods).[16] The observers may record occurrence or nonoccurrence of specified patient behaviors and other events (e.g., pain-referenced verbalizations or physical activity in chronic pain) that occur during all or part of the sampling interval (i.e., *whole interval* versus *partial interval* sampling). Observers may also record event durations, chains of events (e.g., sequential interaction between a depressed patient and family members), behaviors that are occurring at predetermined sampling points in time (e.g., momentary time sampling; such as the behavior being emitted by a psychiatric inpatient at the end of serial 30-second intervals). Alternatively, observers may rate behaviors on a predetermined scale (e.g., "aggressiveness," "social skill").

Behavioral data can also be obtained from videotapes and audiotapes from the natural environment and the acquisition of behavioral data can be facilitated with many instruments (Tryon, 1991). Computerization has facilitated the acquisition and analysis of behavior in real time and the tracking of response durations and latencies (Tryon, 1996b).

Observers sample several events from a large array of potential target events (i.e., *behavior sampling*). Patient behaviors and other events are

selected for observation because they are: (a) causal or correlated variables for the patient's behavior problems and goals (e.g., compliments and insults emitted during distressed marital interaction), (b) patient problem behaviors, (e.g., frequency of social interaction by a depressed psychiatric inpatient), (c) possible side-effects and generalization of intervention (e.g., peer interactions for a child in a timeout program), (d) behavior goals or positive alternatives to undesirable behaviors (e.g., appropriate social interactions by a delusional patient), (e) high-rate covariates of low rate problems (e.g., classroom compliance by an aggressive adolescent), (f) high-risk events (e.g., aggressive behaviors), and (g) immediate, intermediate, and final outcomes of treatment (Mash & Hunsley, 1990). Always, target events are carefully selected and defined before observation.

Although observers often focus on only one individual at a time, the interaction between two or more individuals can be tracked by coding sequences of behavioral exchanges. A few persons may be selected for observation from a larger group (e.g., a classroom or psychiatric unit) sequentially or randomly.

The accuracy and validity of observation data are affected by several factors. The degree to which observers are trained and their method of observation affect the validity of derived data. Observers must be carefully trained to a criterion level of accuracy prior to observing patients. To reduce the probability of inaccuracy, bias, drift, and other observer errors, inter-observer agreement should be evaluated frequently and randomly. Retraining should be initiated when inter-observer agreement indices fall below an acceptable level (e.g., .8). The composition of observer teams should be changed periodically, and observer knowledge of the patient's status (e.g., pre- or posttreatment) should be minimized.

In unrestricted environments such as a home, some constraints are often placed on the behavior of the individuals to be observed. For example, family members might be requested to remain within two rooms, to postpone phone conversations, and to avoid TV and visits from friends while being observed at home. While such constraints compromise the generalizability of the obtained data, they increase the time efficiency of the observation process.

Several types of data can be derived from observation measures. The most frequent type is the rate of targeted behaviors (usually, the percentage of sampling intervals in which a behavior occurs). More helpful for developing functional analyses with patients, observation data can be subjected to *sequential analysis* and for the calculations of *conditional probabilities* (the probability that a specific behavior will occur given the occurrence of other specified behaviors, events, or situations). For example, observation of family interaction in the home can provide data on negative reciprocity—the relative probability that one family member will respond negatively following a negative (in comparison to a nonnegative) response by another family member.

Qualitative observation of patients in their natural environment (as is commonly used in ethnography and cultural anthropology) can also be a rich source of clinical hypotheses concerning problem behaviors, behavior skill deficits, response classes, behavior chains, and causal variables. As such, qualitative observation can often enhance the content validity of the functional analyses and suggest additional assessment foci. Because of its subjective nature, qualitative observation should not be the primary source of data for the functional analysis.

Observation in the natural environment is a powerful method of assessment. The obtained data are useful for a functional analysis of patient behavior problems, basic psychology research, and intervention outcome evaluation. However, there are several potential sources of inferential error, which include: (a) variance in the environmental contexts in which observation occurs (data should be obtained in contexts of greatest clinical relevance), (b) observer inaccuracy, bias, and drift, (c) errors in behavior sampling (e.g., failure to include important behaviors in a coding system), (d) errors in the time-sampling parameters (e.g., frequency and duration of sampling periods are not matched to the temporal dimensions of the observed events), (e) code complexity that challenges the abilities of observers, (f) insufficient definitional precision of codes, and (g) errors associated with the coders (e.g., attention, training, bias). All sources of error are threats to the validity inferences drawn from the obtained data.

A major source of error in all assessment procedures, but particularly in observation methods, is *reactivity* (Foster et al., 1988; Haynes & Horn, 1982). An assessment process is reactive when it is associated with changes in the behaviors of the individuals involved in the observation process. That is, staff, spouses, and parents may behave differently when observers are present than when they

are not present. Therefore, reactivity is a threat to the *external validity* or *situational and temporal generalizability* of the acquired data and limits the inferences that can be drawn. In the cases of highly socially sensitive behaviors (e.g., sexual or antisocial behaviors), observation in the natural environment may be sufficiently reactive to preclude its use.

Another strategy of observation in the natural environment is participant observation (i.e., informant reports). Participant observation is observation in the natural environment using observers who are normally part of the patient's natural environment.[17] Examples include: (a) observations of positive and negative family interactions in the home by parents, spouses, and children, (b) observation of social behaviors and unusual speech of psychiatric inpatients by nurses, (c) observation of students' academic and disruptive behaviors by teachers, (d) observation of a patient's depressive and sexual behaviors by his or her spouse, (e) observation of social behaviors by participants on a date, (f) observation of a patient's pain behaviors by family members, and (g) observation of children's health behaviors (e.g., asthma episodes, headaches) by parents, and observation by nurses of sleep behaviors of hospitalized patients.

Event- and time-sampling methods used by participant observers can be similar to those used by nonparticipant observers. However, participant observers are usually not as well trained, focus on a more restricted range of target events, and use simpler methods of time-sampling. For example, a staff member on a psychiatric unit might monitor the frequency of social initiations by a patient during short mealtime periods or on the hour. Because of the training that would be required, participant observers seldom record complex sequences of events.

Participant observation has several advantages. It is amenable to idiographic assessment and is inexpensive and applicable with a wide range of problem behaviors, populations, and environmental events. Participant observation can be a cost-efficient method of gathering data on patients in their natural environment, particularly of low frequency or highly sensitive events (e.g., seizures, aggressive and panic episodes, and antisocial behaviors).

Participant observation has been the object of very little research but there are several threats to the validity of data acquired using this method. In addition to all the sources of error when using nonparticipant observers, the acquired data can reflect observer biases and selective attention, previous experience with the patient, and many other difficult-to-identify sources of observation error. The validity of participant observation is probably affected by the same variables that affect nonparticipant observation—observer training, the degree of specification of observed events, the specification of sampling and recording strategies, and the judgments to which the data are applied.

Participant observation can have reactive effects. The reactive effects are likely to be less than those associated with nonparticipant observation because participant observation involves less change in the natural environment of the patient. However, the method of recording, the behaviors recorded, and the relationship between the observer and patient may affect the degree of reactivity. There are many situations in which participant observation might be expected to alter the monitored behavior or to affect the social interaction between the observer and target (e.g., an individual monitoring the sexual or eating behavior of a spouse).

In sum, participant observation may be a very useful method of acquiring clinically useful data on patients in the natural environment and it is a frequently recommend assessment strategy. However, additional research is needed to identify the sources of measurement error, the clinical judgment for which it can be most helpful, and the methods of enhancing its accuracy and validity.

Another potentially useful but infrequently used method of behavioral assessment in the natural environment is *critical event sampling*—the recording (e.g., tape, video) of interactions in problematic situations in the client's natural environment (Tryon, 1998b). For example, tape recorders can be self-actuated by a distressed marital couple during verbal altercations at home, dinner conversations can be recorded, or a socially anxious individual can record conversations while on a date. Interactions during the critical situations are later analyzed by the assessor. Although this method has undergone little psychometric evaluation, it is another cost-efficient method of acquiring data on patients in their natural environment.

Analogue Observation

Analogue observation is a powerful, clinically useful, idiographically amenable, and underused

assessment method. It involves assessment elements that are similar to the natural environment of the patient on some but not other dimensions. Analogue and natural environment assessment may differ in the participants, social stimuli, settings, and required behaviors. Patient behavior in analogue assessment is presumed to correlate with their behavior in the natural environment. For example, to evaluate possible problem-solving difficulties at home, a distressed marital couple might be requested to discuss a problem in their relationship while being observed in a clinic from a one-way mirror. One type of analogue observation used in the evaluation of social skills is the role play, in which a patient in a simulated social situation responds to social stimuli typical of those encountered in the natural environment. A psychiatric patient or socially isolated student might be observed in a clinic waiting room while attempting to initiate and maintain a conversation with a confederate-stranger. Typically, a scene is described to the patient and the confederate provides carefully controlled prompts and responses. Another type of analogue observation is the behavior avoidance test, in which patients are asked to approach a feared object or situation (e.g., entering a crowded cafeteria with another person).

Analogue observation is amenable to the measurement of multiple response modes. Self-report measures of subjective discomfort, self-monitoring, observational, and psychophysiological measures can be obtained. Time- and behavior-sampling methods in analogue observation can be similar to those used in observation in the natural environment.

Analogue observation has been used with a variety of behavior problems, including social anxiety and avoidance, social skills deficits, self-injurious behaviors, dental anxiety, stuttering, heterosexual anxiety, alcohol ingestion, pain behaviors, panic episodes, cigarette refusal skills, parent-child interaction, cognitive processes, marital interaction, speech anxiety, animal phobias, test anxiety, pain behaviors, and eating patterns.

Analogue observation can be a cost-efficient method of observational assessment and an important supplement to self-report data. The assessment situation is arranged to increase the probability that clinically important variables and functional relationships will occur. It is particularly useful for observing events that occur at a low rate in the environment (e.g., for observing social behaviors of an isolated psychiatric patient; for observing problem solving in a noncommunicative marital dyad). Because the physical environment and social stimuli are more carefully controlled than in naturalistic observation, behavioral variance attributable to situational stimuli is reduced, although external validity may concomitantly be reduced.

Several sources of variability and inferential error have been identified in analogue observation (e.g., Kern, 1991; Torgrud & Holborn, 1992). These sources include: (a) instructions to participants, (b) situational stimuli, (c) reactive effects, (d) demand factors, and (e) errors associated with the observers, time and behavior sampling, or other aspects of the data acquisition process. The primary disadvantage to analogue observation is that it is an *indirect* measure of the individual's behavior in the natural environment. Although many studies have demonstrated significant discriminant and criterion-related validity for analogue observation, the results of other studies have been less supportive. Given the importance of situational sources of variance for many behavior problems, analogue assessment can be expected to generate data that are valid in many ways (e.g., discriminating between distressed and nondistressed marital couples), but may not reflect accurately the rate of behavior in the natural environment.[18] The clinical utility of analogue observation and the degree to which data from analogue settings are generalizable to natural settings are likely to vary across subjects, target behaviors, settings, and observation methods.

Analogue observation is particularly useful for the functional analysis of behavior problems when used in conjunction with systematic *manipulation* of hypothesized controlling variables. For example, attention and demand factors can be systematically presented and withdrawn contingent on the self-injurious behavior of developmentally disabled individuals to help identify the factors maintaining those behaviors (e.g., Iwata et al., 1994).

Self-Monitoring

Self-monitoring is a self-report assessment method; it is discussed in a separate section because of its usefulness in clinical assessment and in time-series assessment strategies. Self-monitoring involves systematic self-observation and recording of parameters (e.g., occurrence, intensity) of specified behaviors and environmental

events (Bornstein, Hamilton, & Bornstein, 1986; Gardner & Cole, 1988; Shapiro, 1984, see special section in *Psychological Assessment,* December, 1999). Typically, the events to be recorded by the client are first specified by the client and assessor. A recording form is selected and the patient monitors the selected events for a designated number of days. Recording usually occurs immediately after the targeted event.

Time-sampling parameters vary with the temporal characteristics of the behavior. For low-rate behaviors (e.g., seizures, migraine headaches) patients may record every occurrence of the behavior and contiguous events. For high-rate or continuous behaviors (e.g., frequent tics, blood pressure, mood), patients may record only in specified periods or situations. Self-monitoring has been used in the assessment of many clinically important events, such as eating patterns of obese or bulimic persons, certain types of thoughts (e.g., self-criticisms), smoking, bruxism, blood pressure, heart rate, caffeine intake, fuel conservation, startle responses, deviant sexual behavior and urges, Raynaud's symptoms, hair pulling, self-care behaviors, nausea associated with chemotherapy, arthritic and other chronic pain, alcohol and drug intake, exercise, panic episodes, social anxiety and behaviors, marital interactions, study time and other academic behaviors, seizures, sleeping patterns, and nightmares.

Several types of data can be acquired on multiple response modes through self-monitoring. Clients can monitor overt motor behavior, verbal behavior, occurrence of environmental events associated with their behavior, physiological responses, thoughts, topographical aspects of behavior problem (e.g., location of headaches, multimodal aspects of panic episodes), and affective responses. Durations and intensities, as well as frequencies, can also be monitored. The patient can monitor several behaviors, antecedent events, and consequent events (e.g., situations in which binge eating occurs and social reactions to attempts at social initiation). Thus, data from self-monitoring can aid the development of the functional analysis.

Self-monitoring has many positive attributes. It is applicable to a wide range of behavior problems and is amenable to idiographic assessment. Self-monitoring is inexpensive and takes little client and clinician time. It can be used to gather data on functional relationships in the natural environment and is suitable for time-series assessment and the derivation of quantitative indices of multiple response modalities. Self-monitoring is applicable with many populations—children, inpatients, parents and teachers, and developmental disabled individuals. Events that are not amenable to direct observation by participant and nonparticipant observers (e.g., low-frequency events; events that might be affected by the presence of observers) may be more amenable to assessment with self-monitoring. The use of computer technology in self-monitoring (e.g., Shiffman, 1993) facilitates the acquisition of real-time data, enables more sophisticated analyses of complex functional relationships, and increases its clinical utility.

As with other assessment methods, self-monitoring is subject to several general deficiencies and idiosyncratic sources of error. Perhaps the most significant of those are *observer errors and bias.* Data obtained from self-monitoring can be affected by the expectancies and selective attention of the patient, the social valence of the target behaviors, the contingencies associated with the acquired data, the client's abilities to track and record behaviors, and difficulties associated with compliance for extended periods of self-monitoring. Sometimes, these client-associated constraints may be so great as to compromise the utility of the data.

Other sources of error variance include the degree to which the client has been trained in self-monitoring procedures, the degree to which target events have been specified, time-sampling and recording parameters, contingencies associated with self-monitoring or the submission of the acquired data to the assessor, reactions from persons in the client's social environment to the self-recording procedures, characteristics (e.g., rate, duration) of the targeted behaviors, and valence of the target behavior. One particularly powerful source of inferential error is *reactivity* (Bornstein et al., 1986). The reactive effects of self-monitoring are frequently so great that self-monitoring is sometimes used as a method of treatment with patients (e.g., self-monitoring caloric intake by obese individuals, cigarette smoking by persons in a smoking programs).

Psychophysiological Assessment

A review of treatment studies (Haynes, Falkin, & Sexton-Radek, 1989) noted a dramatic increase in the use of psychophysiological assessment in behavior therapy outcome studies in the last 30

years. Four factors that contribute to this trend are: (a) an increased focus on physiological components of behavior problems (e.g., physiological components of anxiety disorders), (b) an increased involvement by behavior therapists in the analysis and treatment of medical-psychological disorders (e.g., cancer, pain, cardiovascular disorders), (c) an increased use of intervention procedures (e.g., relaxation training, desensitization) designed to modify physiological processes, and (d) advances in measurement technology (e.g., ambulatory monitoring and computer technology).

The focus of behavioral assessors upon physiological as well as cognitive and motoric components of behavior problems has encouraged the adoption of psychophysiological measurement methods, particular electromyographic, cardiovascular, EEG, and electrodermal measures. Behavior problems such as obsessive-compulsive behavior problems, panic and other anxiety disorders, depression, substance abuse, sleeping difficulties, and trauma symptoms have multiple components, including autonomically and centrally mediated physiological responses. As noted earlier in this chapter, physiological, cognitive, and motoric components of a behavior problem frequently do not covary significantly and may have different covariates and causal variables. Consequently, for many behavior problems, assessment of all components is necessary for a valid description, functional analysis, and intervention outcome evaluation.

Psychophysiological measurement is a powerful and clinically useful assessment method for many clients. It is congruent with the emphasis in behavioral assessment on the acquisition of precisely specified variables within a time-series format. Excellent overviews of measurement methods, instrumentation and technological innovations, clinical applications, and sources of measurement error can be found in Andreassi (1995) and Cacioppo and Tassinary (1990).

Self-Report Methods in Behavioral Assessment

Two behavioral assessment methods—interviews and questionnaires—have been adopted from other applied psychological disciplines. A complete discussion of these methods is beyond the domain of this chapter, but I will note several differences in format and content between behavioral and traditional self-report methods. These differences parallel the contrasts in the assumptions underlying the assessment paradigms. More extensive discussions of self-report measurement methods and their psychometric properties, are provided by Anastasi (1988), Nunnally and Bernstein (1994), and Sarwer and Sayers (1998).

Self-report measures have been viewed skeptically by many behavioral assessors, particularly behavior analysts. It has been presumed that the probabilities of client biases, memory errors, and other errors associated with subjective reports are sufficient to prohibit their use in behavioral assessment.[19] Many questionnaires, for example, rely on retrospective recall, generate aggregated indices of traits, focus on global and poorly defined constructs with fuzzy boundaries, do not tap the conditional nature of behavior, and require nomothetically based inferences. Despite these constraints, interviews and questionnaires are among the most frequently used assessment methods used by behavior therapists (e.g., Piotrowski & Zalewski, 1993).

The interview is probably the most frequently used assessment instrument. Almost every behavioral intervention involves pre-intervention verbal interaction with the patient or significant individuals (e.g., teachers, staff, parents) from the patient's environment. Interviews are an important assessment method because they can be used for multiple purposes. The interview is an important source of data on the patient's interactions with his or her environment and contributes strongly to the functional analysis. The pretreatment assessment interview is also used to screen patients for therapy, evaluate and enhance patients' motivations for further assessment and intervention, select additional assessment strategies, inform patients about the assessment-intervention process, establish a positive relationship between the behavior analyst and patient, gather information about causal relationships, and gather historical information.

Behavioral and nonbehavioral assessment interviews differ in their content and format. Compared to nonbehavioral interviews, behavioral interviews are often more: (a) structured, (b) focused on overt behavior and behavior-environment interactions, (c) attentive to situational sources of behavioral variance, (d) focused on current rather than histor-

ical behaviors and determinants, (e) quantitative in orientation, and (f) focused on precise definition of molecular events. The systems perspective of behavioral assessment is most apparent in the interview. Assessment foci often include the client's extended social network, the social and contingency systems of caregivers (e.g., incentives in operation for staff at a psychiatric hospital), sequelae and systems effects potentially associated with intervention (such as changes in family interactions or occupational patterns), and patterns of relationships among multiple behavior problems.

The behavioral assessment interview is becoming a more frequent topic of technological advancement and psychometric evaluation (Hersen & Turner, 1994; Sarwer & Sayers, 1998). Computerization is reducing some sources of error in the interview process and increasing the clinical utility of structured interviews. Other innovations, such as Timeline Followback (Sobell, Toneatto, & Sobell, 1994) may also increase the accuracy of the data derived in interviews.[20]

The questionnaire (e.g., self-report questionnaires, problem inventories, rating scales) is probably the second most frequently used method in behavioral assessment. Questionnaires have been used in the assessment of almost all adult behavior disorders. Many questionnaires used by behavioral assessors (e.g., depression and anxiety scales, marital satisfaction scales) are identical to those used in traditional psychological assessment. Although some questionnaires provide clinically useful data, many have been adopted without sufficient attention to their methods of construction, psychometric properties, and underlying assumptions. They are often insensitive to the conditional nature of the targeted behavior (i.e., they measure "traits") and provide aggregated and molar indices of a multifaceted "syndrome" or "disorder" (Haynes, Uchigakiuchi, et al., 1993). They are sometimes helpful for initial screening or as a nonspecific index of program outcome but seldom have utility for most of the other objectives of behavioral assessment outlined in Table 1.

Some questionnaires are more consistent with assumptions of the behavioral assessment paradigm. These target specific adult behavior problems such as social skills deficits, obsessive-compulsive behaviors, fears and phobias, anger, somatic symptoms, specific areas of marital distress, specific expectancies (e.g., regarding the consequences of drinking alcohol), specific responses to life stressors, physiological dysfunctions, and thoughts associated with specific events. Most focus on more specific and lower-level behaviors and events and attend to situational determinants of behavior. However, their development and application have sometimes violated standard psychometric principles (see The special issue on "Research Methods in Psychological Assessment" in *Psychological Assessment*, 1995, vol. 7). Many were developed in ways inconsistent with principles of instrument development and refinement (Haynes et al., 1995), were not subjected to internal consistency or factor analyses, and did not undergo multi-method validity evaluation. Such deficiencies reduce confidence in the inferences that can be derived from their resultant scores.

When properly developed, evaluated, and applied, self-report questionnaires (and participant report questionnaires) can be an efficient and useful source of data. They are inexpensive to administer and score, have face validity for patients, and their analysis and interpretation can be simplified through computer administration and scoring. They can also be designed to yield data on functional relationships of variables at a clinically useful level of specificity. However, because of reporting biases and other sources of error, self-report methods should be used in conjunction with other methods.

Summary

Many assessment methods are congruent with the behavioral assessment paradigm. The methodological elements of the behavioral assessment paradigm include: (a) an emphasis on empirical hypothesis-testing, (b) the use of time-series assessment strategies, (c) a focus on lower-level, less inferential variables, (d) a focus on functional and conditional aspects of behavior, (e) an emphasis on obtaining data in the natural environment of the patient, and (f) a focus on contemporaneous behavior-environment and behavior-behavior functional relationships. Assessment methods, and different instruments within each method, vary along dimensions of reliability, construct validity, power, applicability, cost-efficiency, incremental utility for clinical decision making, and sources of error. The psychometric and clinical utility characteristics of each method also vary across patient behavior problems, goals, assessment settings, and populations. For example, psychophysiological assessment may be more useful for clinical deci-

sion making for cardiovascular disorders than for childhood conduct disorders.

Considering the alternative assessment methods (e.g., nonbehavioral interviews, objective and projective tests), behavioral assessment is, despite its deficiencies and cost, a powerful assessment paradigm. It provides the clinician and clinical researcher with a set of methods amenable to the multi-method and multi-modal assessment of most adult disorders, in most settings, and for most clinical judgment purposes.

BEHAVIORAL ASSESSMENT AND CLINICAL CASE CONCEPTUALIZATION

One important component of clinical assessment is the *clinical case conceptualization*—the synthesis of assessment information about a patient, usually for the purpose of treatment design. A clinical case conceptualization is a hypothesized model of a patient's behavior problems, the causes, correlates, and effects of those problems; and treatment goals. Different terms have been used to describe this component of behavioral assessment. Recently used terms include "*clinical pathogenesis map*" (Nezu & Nezu, 1989, 1993; Nezu, Nezu, Friedman, & Haynes, 1997), "*case formulation*" (Persons, 1989), and "*functional analysis*" (Haynes & O'Brien, 1990; Haynes et al., 1993; O'Brien & Haynes, 1995) (see discussion in Haynes & O'Brien, 1999). I prefer the term "functional analysis"[21] because it reflects the emphasis of the behavioral assessment paradigm on identifying important functional relationships associated with a patient's behavior problems and goals. However, the elements of the "functional analysis" are congruent with those proposed by Nezu and Nezu, Persons, and others.

The functional analysis is important in behavior therapy because behavioral treatments can differ across patients with the same behavior problem. Behavioral interventions are often designed to modify variables that account for variance in (i.e., trigger, maintain, dampen-exacerbate, or mediate) problem behaviors and goals (Haynes, Spain, & Oliveira, 1993). As noted earlier in this chapter, identical behavior problems can often result from different permutations of multiple causal variables and, consequently, warrant different treatment strategies.

The functional analysis integrates multiple lower-level judgments about a patient that are fundamental to the design of behavioral intervention programs. Lower-level judgments include those regarding a patient's *behavior problems and goals*—their importance (e.g., severity, degree of risk associated with), interrelationships, and sequelae. The functional analysis also includes many judgments about the *causal variables* that affect a patient's behavior problems and goals—their modifiability, functional form (e.g., causal, noncausal, unidirectional-bidirectional), magnitude of effect, and interrelationships.

Component clinical judgments in the functional analysis, and treatment decisions based on them, are subject to many sources of error (Nezu & Nezu, 1989; Turk & Salovey, 1988). Inferential error can be reduced to the degree that the clinical judgments that compose the functional analysis are based on data from multiple sources from valid assessment instruments. Data from previously published studies (e.g., Persons & Fresco, 1996) can also provide information that is useful for the construction of the functional analysis.

Inferential errors may also be reduced by integrating obtained data into a taxonomy or organizational structure, such as the eight diagnostic types of contingencies suggested by Tryon (1996a). Such a taxonomy also can help the behavioral assessor decide which variables should be assessed to best facilitate construction of a functional analysis and treatment decisions.

The functional analysis is a hypothesized and dynamic model of the client. It reflects the clinician's current judgments about a client (which can change with additional data). The functional analysis can also change as a result of treatment and from naturally occurring changes in causal variables. The functional analysis is also conditional in other ways: it may accurately reflect causal variables for a behavior problem in one setting and not in another. The functional analysis emphasizes the identification of variables and relationships that are important for treatment design: *important and controllable functional* relationships, *unidirectional and bidirectional causal relationships*, the *strength of causal relationships,* and the *degree of modifiability of causal variables.*

Despite its important role in behavior therapy, the functional analysis is limited in several ways (Haynes, 1996). First, the methods for selecting the best assessment instruments to develop a func-

tional analysis for a particular patient and behavior problem have not been identified. Second, the methods for integrating data from behavioral assessment into a functional analysis have not been developed. Third, the incremental utility and cost effectiveness of the functional analysis have yet to be established for many behavior problems. It may be most useful with complex cases or when brief therapies are ineffective (Sobell et al., 1994). Fourth, the specific clinical judgments that are most useful in treatment design have not been identified.

Two methods have been proposed to help the clinician in integrating and expressing the complex information contained in a functional analysis: *Clinical Pathogenesis Maps* (Nezu et al., 1996) *and Functional Analytic Clinical Case Models* (*FACCMs*) (Haynes et al., 1993). Both involve graphic depictions of elements of the clinical case conceptualization to promote less intuitive intervention decisions. They graphically illustrate hypotheses about a patient's behavior problems and goals and their relative importance, interrelationships, sequela and the strength, modifiability, and direction of action of causal variables. The FACCM allows the clinician to estimate formally or informally the relative magnitude of effect of a particular treatment focus, given the clinician's hypotheses about the patient.

SUMMARY

The behavioral assessment is a powerful and evolving psychological assessment paradigm. It is the subject of many books, published articles, symposia, and presentations at scientific conventions. Behavioral assessment methods are often used in clinical practice and are taught in many Ph.D. programs. The impetus for behavioral assessment's development comes from behavior therapy and dissatisfaction with traditional trait-based self-report and projective assessment methods.

Behavioral assessment can have many objectives, including the identification of intervention target behaviors and treatment goals, the identification of causal variables, the development of a functional analysis and intervention strategies, and the evaluation of ongoing intervention strategies. The identification of shared variance is a supraordinate goal. Behavioral assessment is a functional approach in that the methods of assessment depend on the objectives of assessment.

The behavioral assessment paradigm includes several assumptions about the characteristics and causes of behavior problems. Behavior problems can have multiple modes and parameters, which can differ among individuals and across situations and time. There are multiple possible causes and causal paths for most behavior problems, which can also differ across individuals and time. Causal factors often act independently, interactively, and additively. The behavioral assessment paradigm emphasizes contemporaneous, environmental, reciprocal determinism. Finally, extended social systems and situational factors are presumed to have important causal functions.

The behavioral assessment paradigm includes many methods of assessment. These include observation, self-monitoring, psychophysiology, and self-report. Clinical judgments are most likely to be valid and useful when using assessment methods congruent with an emphasis on empirical hypothesis-testing, time-series assessment strategies, lower-level variables, the functional and conditional aspects of behavior, data obtained in the natural environment, and contemporaneous behavior-environment and behavior-behavior functional relationships.

In spite of its cost-efficiency deficiencies, behavioral assessment includes powerful assessment methods. It provides the clinician and clinical researcher with a set of methods amenable to the multi-method and multi-modal assessment of most adult disorders, in most settings, and for most clinical judgment purposes. The behavioral assessment paradigm provides a coherent set of principles and methods to guide clinical judgment and research.

NOTES

1. A psychological assessment paradigm includes a coherent set of principles, values, assumptions, and methods. It includes assumptions about the relative importance of behavior problems, the causal variables that affect behavior, the mechanisms of causal action, the importance and role of assessment, and the best methods of assessment. A psychological assessment paradigm also includes guidelines for problem solving, decision-making strategies, and data interpretation. (Haynes, 1996a).

2. A "behavior problem" refers to specific behavior (motoric, verbal, cognitive, physiological) excess or deficit. Several covarying behavior problems are sometimes considered as a behavior "disorder." Component behaviors of a disorder are often controlled by different variables (or controlled to different degrees by the same variable) and may demonstrate only mild to moderate covariation (Haynes, 1992).

3. Other required assessment courses were: intellectual assessment (94%), objective personality assessment (89%), and projective assessment (85%). Eight percent and 45 percent of the program directors predicted an increase and decrease, respectively, in the emphasis on projective assessment in the future.

4. Karoly (1993) suggested that a problem for many patients can be traced to conflicts between multiple and incompatible goals.

5. Kanfer (1985) referred to this class of alternative target behaviors as those that are "instrumental for altering the current problem situation toward a more effective future state" (p. 12).

6. A *functional response class* is a set of covarying behaviors, which may differ in form, that are under the control of the same contingencies—a set of behaviors which have the same "function" (e.g., completion of a homework assignment, raising a hand in class, and talking to classmates may all be maintained by teacher attention).

7. Because of epistemological complexities associated with the idea of "causation," behavioral assessors often emphasize "functional" rather than "causal" relationships (see discussion in Haynes, 1992; Haynes & O'Brien, 1990). However, inferences of causation underlie all behavioral interventions. Causal relationships are a subset of functional relationships. Two variables have a causal relationship when: (a) they covary (i.e., when they have a functional relationship), (b) the causal variable reliably precedes its effect, (c) there is a logical mechanism for the causal relationship (i.e., a logical connection), and (d) alternative explanations for the observed covariance can be excluded (Asher, 1976; Haynes, 1992). Causal variables may be original, triggering, moderating, or maintaining. Furthermore, causal variables need not be necessary, sufficient, exclusive, important, or modifiable.

8. There are many assessment occasions in which fully informed consent is difficult to obtain. For example, informed consent principles are challenged when assessing paranoid individuals, severely developmentally disabled persons, and in some forensic assessment cases.

9. Subparadigms of behavior therapy (e.g., behavior analysis, cognitive-behavior therapy) differ in the weight they place on individual elements in this table. Tables are adopted from Haynes (1996a, 1996b). The selection of elements for Tables 2 and 3 were influenced by Bandura (1969), Barrios (1988); Bellack and Hersen (1988); Bornstein, Bornstein, and Dawson (1984); Ciminero (1986); Cone (1988), Eysenck (1986), Hersen and Bellack (1996), Johnston and Pennypacker (1993), Kratochwill and Shapiro (1988), Mash and Terdal (1988), Nelson and Hayes (1986), O'Donohue and Krasner (1995), Ollendick and Hersen (1984, 1993), Strosahl and Linehan (1986), and Tryon (1985).

10. The apparent level of covariation among response modes is affected by the manner in which they are measured. Different time-sampling parameters and assessment instruments will result in different inferences about covariation among multiple response modes.

11. The "importance" of a behavior problem is a complex clinical inference based on judgments about its rate and magnitude, probability of harm, degree of impact on the patient's quality of life, and its causal relationship to other behavior problems.

12. The dynamic attributes of behavior and events have long been recognized. The Greek philosopher Heraclitus, around 500 BC, noted "Everything flows, nothing stays" (Daintith, 1994).

13. In causal models of behavior disorders, a moderating variable is one that changes the relationship between two other variables. For example, "social support" would be a moderating variable if it affected the probability that an environmental disaster would be associated with PTSD symptoms.

14. Behavior problems can also be conditional for physiological states. For example, the probability of aggressive or delusional behaviors can covary with alcohol intoxication and medication intake.

15. There has been an emphasis in behavioral assessment on a SORC (stimulus, organism, response, contingency) model (e.g., Goldfried, 1982) to guide assessment. The SORC model has served to contrast behavioral with traditional conceptual systems, to emphasize the multifaceted qualities and determinants of behavior problems

and to emphasize the importance of contiguous antecedent and consequent events.

16. The number of observation sessions, their length, and the time-sampling intervals in behavioral observation are influenced by the rate, variability, cyclicity, and slope of the observed behaviors. Suen and Ary (1989) discuss many of these principles. Observation also involves situation sampling, in which observers sample behaviors in situations associated with the highest probability of clinically significant behaviors and interactions (i.e., *high-risk situations*).

17. In ethnography and cultural anthropology, the term "participant observation" usually refers to qualitative observation by external observers.

18. Inferences of "validity" of an assessment instrument depend on the purposes for which the data is used. For example, data from the analogue assessment of specific parent-child interactions may not represent the rate of those interactions in the natural environment but can discriminate between distressed and nondistressed families (Haynes & Wai'alae, 1994).

19. Vaillant (1977) observed that "It is all too common for caterpillars to become butterflies and then to maintain that in their youth they had been little butterflies."

20. Clients provide retrospective estimates of substance use with the visual aid of a calendar for up to 12 months. Several aids are used to enhance accuracy: a daily calendar, key dates, long periods in which they abstained or were continuously drunk, and discreet events.

21. "Functional analysis" has also been used to refer to the systematic manipulation of hypothesized controlling variables in applied and experimental behavior analysis (e.g., Iwata et al., 1994). "Functional analysis" in algebra and calculus refers to the study of linear and nonlinear functions (Borowski & Borwein, 1991) and is consistent with its use in this chapter.

AUTHOR NOTE

Dorothy Chin, Elaine Heiby, Edward Kubany, Dave Richard, and Warren Tryon made helpful comments on an earlier version of this chapter.

REFERENCES

Acierno, R., Hersen, M., & Ammerman, R. T. (1994). Overview of the issues of the issues in prescriptive treatment. In M. Hersen & R. T. Ammerman (Eds), *Handbook of prescriptive treatments for adults.* New York: Plenum.

Alessi, G. (1988). Direct observation methods for emotional/behavior problems. In E. S. Shapiro & T. R. Kratochwill (Eds.), *Behavioral assessment in schools, Conceptual foundations and practical applications* (pp. 14–75). New York: The Guilford Press.

Alexander, F. G., & Selesnick, S. T. (1966). *The history of psychiatry: An evaluation of psychiatric thought and practice from prehistoric times to the present.* New York: Harper & Row.

American Psychiatric Association (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan.

Andreassi, J. L. (1995). *Psychophysiology: Human behavior and physiological response,* (3rd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Asher, H. B. (1976). *Causal modeling.* Beverly Hills: Sage Publications.

Asterita, M. F. (1985). *The physiology of stress.* New York: Human sciences press, Inc.

Bachrach, A. J. (Ed.) (1962). *Experimental foundations of clinical psychology.* New York: Basic Books.

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

Bandura, A. (1981). In search of pure unidirectional determinants. *Behavior Therapy, 12,* 315–328.

Barlow, D. H. (1981). *Behavioral assessment of adult disorders.* New York: Guilford.

Barnett, P. A., & Gotlib, I. H. (1988). Psychosocial functioning and depression: Distinguishing among antecedents, concomitants, and consequences. *Psychological Bulletin, 104,* 97–126.

Barrios, B. A. (1988). On the changing nature of behavioral assessment. In A. S. Bellack & M. Hersen (Eds.), *Behavioral assessment, a practical handbook* (pp. 3–41). New York: Pergamon Press.

Beach, S., Sandeen, E., & O'Leary, K. D. (1990). *Depression in marriage.* New York: Guilford Press.

Bellack, A. S., & Hersen, M. (1988). *Behavioral assessment, a practical handbook.* New York: Pergamon Press.

Bornstein, P. H., Bornstein, M. T., & Dawson, D. (1984). Integrated assessment and treatment. In T.H. Ollendick & M. Hersen (Eds.), *Child behavioral assessment: Principles and procedures* (pp. 223–243). New York: Pergamon.

Bornstein, P. H., Hamilton, S. B., & Bornstein, M. T. (1986). Self-monitoring procedures. In A. R. Ciminero, C. S. Calhoun, & H. E. Adams (Eds.), *Handbook of behavioral assessment* (pp 176–222). New York: John Wiley & Sons.

Borowski, E. J., & Borwein, J. M. (1991). *The Harper Collins dictionary, Mathematics.* New York: Harper Collins, Publishers.

Bott, H. (1928). Observations of play activity in nursery school. *Genetic Psychology Monographs, 4,* 44–88.

Burish, T. G., Carey, M. P., Krozely, M. G., & Greco, F. A. (1987). conditioned side effects induced by cancer chemotherapy: Prevention through behavioral treatment. *Journal of Consulting and Clinical Psychology, 55,* 42–48.

Burton, S. (1994). Chaos, self-organization, and psychology. *American Psychologist, 49,* 4–14. In M. Hersen, & A. S. Bellack, (Eds.) *Behavioral assessment: A practical handbook* (4th ed.). Boston: Allyn & Bacon.

Carey, M. P., Lantinga, L. J., & Krauss, D. J. (1994) Male erective disorder. In M. Hersen & R. T. Ammerman (Eds), Handbook of prescriptive treatments for adults.(pp 347-367). New York: Plenum.

Cacioppo, J. T., & Tassinary, L. G. (1990). *Principles and psychophysiology, Physical, social, and inferential elements.* New York; Cambridge University Press.

Ciminero, A. R., Calhoun, K. S., & Adams, H. E. (1977). *Handbook of behavioral assessment.* New York: Wiley.

Collins, L. M., & Horn, J. L. (Eds.) (1991). *Best methods for the analysis of change.* Washington, DC: American Psychological Association.

Cone, J. D. (1988). Psychometric considerations and the multiple models of behavioral assessment. In Bellack, A. S. & Hersen, M. (Eds.), *Behavioral assessment, A practical handbook* (pp. 42–66). New York: Pergamon Press.

Cone, J. D. (1998). Psychometric considerations: concepts, contents and methods. In M. Hersen & A. S. Bellack, (Eds.) *Behavioral Assessment: A practical handbook* (4th ed.). Boston: Allyn & Bacon.

Cone, J. D., & Hawkins, R. P. (Eds.). (1977). *Behavioral assessment: New directions in clinical psychology.* New York: Brunner/Mazel.

Craske, M. G., & Waikar, S. V. (1994). Panic disorder. In M. Hersen & R. T. Ammerman (Eds), *Handbook of prescriptive treatments for adults* (pp 135–155). New York: Plenum.

Creer, T. L., & Bender, B. G. (1993). Asthma. In R. J. Gatchel & E. B. Blanchard (Eds.), P*sychophysiological disorders, research and clinical applications* (pp. 151–204) Washington, DC: American Psychological Association.

Daintith, J. (1994). *Bloomsbury treasury of quotations,* London: Bloomsbury Publishing Plc.

Durand, V. M., & Carr, E. G. (1991). Functional communication training to reduce challenging behavior: Maintenance and application in new settings. *Journal of Applied Behavior Analyses, 24,* 251–264.

Eels, T. (1996). *Handbook of psychotherapy case formulation.* New York: Guilford Press.

Evans, I. (1993). Constructional perspectives in clinical assessment. *Psychological Assessment, 5,* 264–272.

Eysenck, H. J. (1986). A critique of contemporary classification and diagnosis. In T. Millon & G. L. Klerman (Eds.), *Contemporary directions in psychopathology, Toward the DSM-IV* (pp 73–98). New York: Guilford Press.

Eysenck, H. J., & Martin, I. (1987). *Theoretical foundations of behavior therapy.* New York: Plenum Press.

Fernández-Ballestros, R. (Ed.) (1994). *Evaluacion Conductual Hoy Behavioral assessment today.* Madrid: Ediciones Piramide.

Figley, C. R. (Ed.) (1979). T*rauma and its wake: Volume 1: The study of post-traumatic stress disorder.* New York: Brunner/Mazel.

Foster, S. L., Bell-Dolan, D. J., & Burge, D. A. (1988). Behavioral observation. In A. S. Bellack & M. Hersen (Eds.), *Behavioral assessment, A practical handbook* (pp. 119–60). New York: Pergamon Press.

Foster, S. L., & Cone, J. D. (1986). Design and use of direct observation systems. In A. R. Ciminero, C. S. Calhoun, & H. E. Adams (Eds.), *Handbook of behavioral assessment* (pp. 253–324). New York: John Wiley & Sons.

Gatchel, R. J. (1993). Psychophysiological disorders: Past and present perspectives. In R. J. Gatchel & E. B. Blanchard (Eds.), *Psychophysiological disorders, research and clinical applications* (pp. 1–22) Washington, DC: American Psychological Association.

Gatchel, R. J., & Blanchard, E. B. (1993). *Psychophysiological disorders, research and clinical*

applications. Washington, DC: American Psychological Association.

Gardner, W. I., & Cole, C. L. (1988). Self-monitoring procedures. In E. S. Shapiro & T. R. Kratochwill (Eds.), *Behavioral assessment in schools, Conceptual foundations and practical applications* (pp. 206–246). New York: The Guilford Press.

Glass, C. (1993). A little more about cognitive assessment. *Journal of Counseling and Development, 71,* 546–548.

Goldfried, M. R. (1982). Behavioral Assessment, an overview. In A. S. Bellack, M. Hersen, & A. E. Kazdin (Eds.), *International handbook of behavior modification and therapy* (pp. 81–107). New York: Plenum Press.

Goodenough, F. L. (1928). Measuring behavior traits by means of repeated short samples. *Journal of Juvenile Research, 12,* 230–235.

Guevremont, D. C., & Spiegler, M. D. (1990, November). *What do behavior therapists really do? A survey of the clinical practice of AABT members.* Paper presented at the 24th Annual Convention of the Association for Advancement of Behavior Therapy, San Francisco, CA.

Halford, W. K., Sanders, M. R., & Behrens, B. C. (1994). Self-regulation in behavioral couples' therapy. Behavior Therapy, 25, 431–452.

Hawkins, R. P. (1986). Selection of target behaviors. In R. O. Nelson & S. C. Hayes (Eds.), *Conceptual foundations of behavioral assessment* (pp. 331–383). New York: Guilford Press.

Haynes, S. N. (1986). The design of intervention programs. In R. O. Nelson & S. Hayes (Eds.), *Conceptual foundations of behavioral assessment* (pp. 386–429). New York: Guildford Press.

Haynes, S. N. (1992). *Models of causality in psychopathology: Toward synthetic, dynamic and nonlinear models of causality in psychopathology.* Des Moines, IA: Ayllon & Bacon.

Haynes, S. N. (1994). Clinical judgment and the design of behavioral intervention programs: Estimating the magnitudes of intervention effects. *Psichologia Conductual. 2,* 165–184.

Haynes, S. N. (1998b). The changing nature of behavioral assessment. In M. Hersen & A. Bellack (Eds.), *Behavioral assessment, A practical guide* (4th ed.). Boston: Allyn & Bacon.

Haynes, S. N. (1998c). The assessment-treatment relationship in behavior therapy: The role of the functional analysis. *The European Journal of Psychological Assessment, 14,* 26–34.

Haynes, S. N. & Chavez, R. (1983). The interview in the assessment of marital distress. In E. E. Filsinger (Ed.), *A sourcebook of marriage and family assessment.* Beverly Hills: Sage Publications.

Haynes, S. N., Blaine, D., & Meyer, K. (1995). Dynamical Models for Psychological Assessment: Phase-Space Functions. *Psychological Assessment, 7,* 17–24.

Haynes, S. N., Falkin, S., & Sexton-Radek, K. (1989). Psychophysiological measurement in behavior therapy. In G. Turpin (Ed.), *Handbook of clinical psychophysiology* London: John Wiley and Sons.

Haynes, S. N., & Horn, W. F. (1982). Reactive effects of behavioral observation. *Behavioral Assessment, 4,* 369–385.

Haynes, S. N., & O'Brien, W. O. (1990). The functional analysis in behavior therapy. *Clinical Psychology Review, 10,* 649–668.

Haynes, S. N., & O'Brien, W. O. (1999). Behavioral assessment. New York: Kluwor/Plenum.

Haynes, S. N., Spain, H., & Oliviera, J. (1993). Identifying causal relationships in clinical assessment. *Psychological Assessment, 5,* 281–291.

Haynes, S. N., Uchigakiuchi, P., Meyer, K., Orimoto, Blaine, D., & O'Brien, W. O. (1993). Functional analytic causal models and the design of treatment programs: Concepts and clinical applications with childhood behavior problems. *European Journal of Psychological Assessment, 9,* 189–205.

Haynes, S., N., & Waialae, K. (1994). Psychometric foundations of behavioral assessment. In R. Fernández-Ballestros, (Ed.), *Evaluacion Conductual Hoy (Behavioral Assessment Today).* Madrid: Ediciones Piramide.

Haynes, S. N., & Wu-Holt, P. (1998). Methods of assessment in health psychology. In M. E. Simon (Ed.), *Handbook of health psychology* (pp. 420–444). Madrid, Sigma

Heatherton, T. F., & Weinberger, J. L. (Eds.) (1994). *Can personality change.* Washington, DC: American Psychological Association.

Heiby, E. M. (1995). Chaos theory, nonlinear dynamic models, and psychological assessment. *Psychological Assessment, 7,* 5–9.

Hersen, M., & Bellack, A. S. (Eds.) (1998). *Behavioral assessment: A practical handbook* (4th ed.). Boston: Allyn & Bacon.

Hersen, M., Kazdin, A., & Bellack, A. (Eds.) (1991). *The clinical psychology handbook.* New York: Pergamon Press.

Hersen, M., & Turner, S. M. (Eds.) (1994). *Diagnostic interviewing* (2nd ed). New York: Plenum Press.

Hutt, S. J., & Hutt, C. (1970). *Direct observation and measurement of behavior.* Springfield, IL: Charles C. Thomas.

Iwata, B. A. (and 14 other authors) (1994). The functions of self-injurious behavior: An experimental-epidemiological analysis. *Journal of Applied Behavior Analysis, 27,* 215–240.

Jacobson, N. S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59,* 12–19.

James, L. D., Thorn, B. E., & Williams, D. A. (1993). Goal specification in cognitive-behavioral therapy for chronic headache pain. *Behavior Therapy, 24,* 305–320.

Jensen, B. J., & Haynes, S. N. (1986). Self-report questionnaires. In A. R. Ciminero, C. S. Calhoun, & H. E. Adams (Eds.), *Handbook of behavioral assessment* (pp 150–175). New York: John Wiley & Sons.

Johnston, J. M., & Pennypacker, H. S. (1993). *Strategies and tactics of behavioral research, second edition.* Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Kahle, A. L., & Kelley, M. L. (1994). Children's homework problems: A comparison of goal setting and parent training. *Behavior Therapy, 25,* 275–290.

Kail, R. V., & Wickes-Nelson, R. (1993). *Developmental psychology.* New York: Prentice Hall.

Kanfer, F. H. (1985). Target selection for clinical change programs. *Behavioral Assessment, 7,* 7–20.

Karoly, P. (Ed.) (1988). *Handbook of Child health assessment.* New York: John Wiley and Sons.

Karoly, P. (1993). Goal systems: An organizing framework for clinical assessment and treatment planning. *Psychological Assessment., 5,* 273–280

Kazdin, A. E. (1978). *History of behavior modification.* Baltimore: University Park Press.

Kazdin, A. (1992). *Research design in clinical psychology.* New York: Macmillan.

Kazdin, A. E., & Kagan, J. (1994). Models of dysfunction in developmental psychopathology. *Clinical Psychology: Science and Practice, l,* 35–52.

Keefe, F. L., Kopel, S. A., & Gordon, S. B. (1978). *A practical guide to behavioral assessment.* New York: Springer.

Kern, J. M. (1991). An evaluation of a novel role-play methodology: The standardized idiographic approach. *Behavior Therapy, 22,* 13–29.

Kratochwill, T. R., & Levin, J. R. (1992). *Single-case research design and analysis; New directions for psychology and education.* Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Kratochwill, T. R., & Shapiro, E. S. (1988). Introduction: Conceptual foundations of behavioral assessment. In E. S. Shapiro & T. R. Kratochwill (Eds.), *Behavioral assessment in schools, Conceptual foundations and practical applications* (pp. 1–13). New York: The Guilford Press.

Lang, P. J. (1995). The emotion probe: Studies of motivation and attention. *American Psychologist, 50,* 519–525.

Last, C. G., & Hersen, M. (Eds.) (1988). *Handbook of anxiety disorders.* New York: Pergamon.

Linscott, J., & DiGiuseppe, R. (1996). Cognitive assessment. In M. Hersen, & A. S. Bellack (Eds.), *Behavioral assessment: A practical handbook* (4th ed.). Boston: Allyn & Bacon.

Malec, J. F., & Lemsky, C. (1996). Behavioral assessment in medical rehabilitation: Traditional and consensual approaches. In L. Cushman & M. Scherer (Eds.). Psychological assessment in medical rehabilitation (pp 199–236). Washington: American Psychological Association Press.

Margolin, G. (1981). Practical applications of behavioral marital assessment. In E. E. Filsinger & R. A. Lewis (Eds.), *Assessing marriage: New behavioral approaches.* Beverly Hills: Sage Publications.

Mash, E. J., & Hunsley, J. (1990). Behavioral assessment: A contemporary approach. In A. S. Bellack, M. Hersen, & A. E. Kazdin (Eds.), *International handbook of behavior modification and therapy* (2nd ed.) (pp. 87–106). New York: Plenum.

Mash, E. J., & Hunsley, J. (1993). Assessment considerations in the identification of failing psychotherapy: Bringing the negatives out of the darkroom. *Psychological Assessment, 5,* 292–301.

Mash, E. J., & Terdal, L. G. (1981). *Behavioral assessment of childhood disorders.* New York: Guildford Press.

Mash, E. J., & Terdal, L. G. (1997). *Behavioral assessment of childhood disorders.* New York: Guildford Press.

McConaghy, N. (1998). Assessment of sexual dysfunction and deviation. In M. Hersen & A. S. Bellack (Eds.), *Behavioral assessment: A*

practical handbook (4th ed.). Boston: Allyn & Bacon.

McFall, R. M. (1982). A review and reformulation of the concept of social skills. *Behavioral Assessment, 4,* 1–33.

McFall, R. M. (1986). Theory and method in assessment: The vital link. *Behavioral Assessment, 8,* 3–10.

McFall, R. M., & McDonel, E. (1986). The continuing search for units of analysis in psychology: Beyond persons, situations and their interactions. In R. O. Nelson & S. C. Hayes (Eds.), *Conceptual foundations of behavioral assessment* (pp. 201–241). New York: Guilford Press.

McReynolds, P. (1986). History of assessment in clinical and educational settings. In R. O. Nelson & S. C. Hayes (Eds.), *Conceptual foundations of behavioral assessment* (pp. 42–80). New York: Guilford.

Mischel, W. (1968). *Personality and assessment.* New York: Wiley.

Nay, W. F. (1979). *Multimethod clinical assessment.* New York: Gardner Press, Inc.

Nelson, R. O. (1983). Behavioral assessment: Past, present, and future. *Behavioral Assessment, 5,* 195–206.

Nelson, R. O. (1988). Relationships between assessment and treatment within a behavioral perspective. *Journal of Psychopathology and Behavioral Assessment, 10,* 155–169.

Nelson, R. O., & Hayes, S. C. (1986). *Conceptual foundations of behavioral assessment.* New York: Guilford Press.

Nezu, A. M., & Nezu, C. M. (1989). *Clinical decision making in behavior therapy: A problem solving perspective.* Champaign, IL: Research Press.

Nezu, A. M., & Nezu, C. M. (1993). Identifying and selecting target problems for clinical interventions: A problem-solving model. *Psychological Assessment, 5,* 254–263.

Nezu, A., Nezu, C., Friedman, & Haynes, S. N. (1997). Case formulation in behavior therapy. T. D. Eells (Ed.), *Handbook of psychotherapy case formulation.* New York: Guilford.

Nunnally, J. C., & Burnstein, I. H. (1994). *Psychometric theory, 3rd ed.* New York: McGraw-Hill, Inc.

O'Brien, W. H., & Haynes, S. N. (1995). A functional analytic approach to the conceptualization, assessment and treatment of a child with frequent migraine headaches. In *Session, l,* 65–80.

O'Donohue, W., & Krasner, L. (1995). *Theories of behavior therapy.* Washington, DC: American Psychological Association.

O'Leary, K. D., Vivian, D., & Malone, J. (1992). Assessment of physical aggression against women in marriage: The need for multimodal assessment. *Behavioral Assessment, 14,* 5–14.

Ollendick, T. H. & Hersen, M. (1984). *Child behavioral assessment, principles and procedures.* New York: Elmsford, NY: Pergamon Press.

Ollendick, T. H., & Hersen, M. (1993). *Handbook of child and adolescent assessment.* Boston: Allyn & Bacon.

Parten, M. B. (1932). Social participation among preschool children. *Journal of Abnormal and Social Psychology, 27,* 243–269.

Persons, J. B. (1989). *Cognitive therapy in practice: A case formulation approach.* New York: Norton.

Persons, J. B. (1992). A case formulation approach to cognitive-behavior therapy: Application to panic disorder. *Psychiatric Annals, 22,* 470–473.

Persons, J. B., & Bertagnolli, A. (1994). Cognitive-behavioural treatment of multiple-problem patients: Application to personality disorders. *Clinical Psychology and psychotherapy, l,* 279–285.

Persons, J. B., & Fresco, D. M. (1998). Assessment of depression. In M. Hersen & A. S. Bellack (Eds.), *Behavioral assessment: A practical handbook* (4th ed.). Boston: Allyn & Bacon.

Piotrowski, C., & Zalewski, C. (1993). Training in psychodiagnostic testing in APA-approved PsyD and PhD Clinical psychology programs. *Journal of Personality Assessment, 61,* 394–405.

Regier, D. A., Farmer, M. E., Rae, D. S., Locke, B. Z., Keith, S. J., Judd, L. L., & Goodwin, F. K. (1990). Comorbidity of mental disorders with alcohol and other drug abuse. *Journal of the American Medical Association, 264,* 251–2518.

Repp, A. C., & Singh, N. N. (1990). (Eds.), *Perspectives on the use of nonaversive and aversive interventions for persons with developmental disabilities.* Sycamore, IL: Sycamore.

Sarwer, D., & Sayers, S. L. (1998). Behavioral interviewing. In M. Hersen & A. S. Bellack (Eds.), *Behavioral assessment: A practical handbook* (4th ed.). Boston: Allyn & Bacon.

Schulte, D. (1992). Criteria of treatment selection in behaviour therapy. *European Journal of Psychological Assessment, 8,* 157–162.

Schlundt, D. G., & McFall, R. M. (1987). Classifying social situations: A comparison of five methods. *Behavioral Assessment, 9,* 21–42.

Shadish, W. R. (1996). Meta-analysis and the exploration of causal mediating processes: A primer of examples, methods, and issues. *Psychological Methods, l,* 47–65.

Shapiro, E. W., & Kratochwill, T. R. (Eds.) (1988). *Behavioral assessment in schools, Conceptual foundations and practical applications.* New York: The Guilford Press.

Shapiro, E. S., (1984). Self-Monitoring. In T. H. Ollendick & M. Hersen (Eds.), *Child behavioral assessment: Principles and procedures.* Elmsford, NY: Pergamon Press.

Shiffman, S. (1993). Assessing smoking patterns and motives. *Journal of Consulting and Clinical Psychology, 61,* 732–742.

Silva, F. (1993). *Psychometric foundations and behavioral assessment.* Newbury Park, CA: Sage Publications, Inc.

Simon, H. A. (1971). Spurious correlation: A causal interpretation. In H. M. Blalock (Ed.), *Causal models in the social sciences* (pp. 5–17). Chicago: Aldine Atherton.

Smith, G. T. (1994). Psychological expectancy as mediator of vulnerability to alcoholism. *Annals of the New York Academy of Sciences, 708,* 165–171.

Sobell, L. C., Toneatto, T., Sobell, M. B. (1994). Behavioral assessment and treatment planning for alcohol, Tobacco, and other drug problems: Current status with an emphasis on clinical applications. *Behavior Therapy, 25,* 523–532.

Strosahl, K. D., & Linehan, M. M. (1986). Basic issues in behavioral assessment. In A. Ciminero, K. S. Calhoun, & H. E. Adams (Eds.), *Handbook of behavioral assessment* (pp 12–46). New York: Wiley.

Stuart, R. B. (1980). *Helping couples change.* New York: Guilford Press.

Suen, H. K., & Ary, D. (1989). *Analyzing quantitative observation data.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Sutker, P. B., & Adams, H. E. (Eds.) (1993). *Comprehensive handbook of psychopathology* (pp. 47–56). New York: Plenum Press.

Taylor, C. B., & Agras, S. (1981). Assessment of phobia. In D. H. Barlow (Ed.), *Behavioral assessment of adult disorders.* New York: Guilford Press.

Torgrud, L. J., & Holborn, S. W. (1992). Developing Externally Valid role-play for Assessment of social skills: A behavior analytic perspective. *Behavioral Assessment, 14,* 245–277.

Tryon, W. W. (1991). *Activity measurement in psychology and medicine.* New York: Plenum.

Tryon, W. (1998a). Observing contingencies: Taxonomy and methods. *Clinical Psychology Review* (in press).

Tryon, W. W. (1996b). Behavioral observation. In M. Hersen & A. S. Bellack. *Behavioral Assessment: A practical handbook* (4th ed.). Boston: Allyn & Bacon.

Turk, D. C., & Melzack, R. (Eds.) (1992). *Handbook of pain assessment.* New York: The Guilford Press.

Turk, D. C., & Salovey, P. (Eds.) (1988). *Reasoning, inference, and judgment in clinical psychology.* New York: The Free Press.

Turkat, I. (1986). The behavioral interview. In A. Ciminero, K. S. Calhoun, & H. E. Adams (Eds.), *Handbook of behavioral assessment* (pp. 109–149). New York: Wiley.

Ullmann, L. P., & Krasner, L. (1965). *Case studies in behavior modification.* New York: Holt, Rinehart & Winston.

Vaillant, G. E. (1977). *Adaptation to life.* Boston: Little, Brown.

Voeltz, L. M., & Evans, I. M. (1982). The assessment of behavioral interrelationships in child behavior therapy. *Behavioral Assessment, 4,* 131–165.

Wahler, R. G., & Dumas, J. E. (1989). Attentional problems in dysfunctional mother-child interactions: An interbehavioral model. *Psychological Bulletin, 105,*116–130.

Walitzer, K. S., & Connors, G. J. (1994). Psychoactive substance use disorders. In M. Hersen & R. T. Ammerman (Eds), *Handbook of prescriptive treatments for adults.* (pp 53–71). New York: Plenum.

Weiss, R. L., & Heyman, R. E. (1990). Observation of marital interaction. In F. D. Fincham & T. N. Bradury (Eds.), *The psychology of marriage: Basic issues and applications* (pp. 87–117). New York: Guilford.

Wincze, J. P., & Carey, M. P. (1991). *Sexual dysfunctions: Guide for assessment and treatment.* New York: Guilford.

Wolman, B. B. (1978). *Clinical diagnosis of mental disorders.* New York: Plenum.

Wolpe, J. (1958). *Psychotherapy by reciprocal inhibition.* Stanford, CA: Stanford University Press.

Wolpe, J., & Turkat, I. D. (1985). Behavioral Formulation of clinical cases. In I. Turkat (Ed.), *Behavioral case formulation* (pp. 213–144). New York: Plenum Press

Yoshikawa, H. (1994). Prevention as cumulative protection: Effects of early family support and education on chronic delinquency and its risks. *Psychological Bulletin, 115,* 28–54.

Youkilis, H. D., & Bootzin, R. R. (1981). A psychophysiological perspective of the etiology and treatment of insomnia. In S. N. Haynes & L. R. Gannon (Eds.), *Psychosomatic disorders: A psychophysiological approach to etiology and treatment.* New York: Praeger.

PART IX

SPECIAL TOPICS AND APPLICATIONS

This Page Intentionally Left Blank

CHAPTER 20

TESTING AND INDUSTRIAL APPLICATION

Robert D. Gatewood
Robert Perloff
Evelyn Perloff

INTRODUCTION

To a large extent performance of an organization is dependent upon performance of its individual members. Because individual performance is a function of ability, motivation, and situational factors (Borman, 1991), the many organizational programs that affect any of these three variables (e.g., selection, job design, compensation, training, performance measurement) are important to the overall well-being of the organization. They, therefore, must be designed and implemented appropriately. Because each of these organizational programs use data about characteristics of employees, a key component of their appropriate design and implementation is the assessment of individuals. For example, pay-for-performance compensation systems are based upon assessment of individual/group work performance. While these organizational programs have used data from the assessment of numerous characteristics of individuals, measurement of five types of characteristics has been dominant in these programs: (a) knowledge, skills, and abilities; (b) personality; (c) physiological attributes; (d) attitudes; and (e) job performance. That is, data from at least one of these five types of characteristics are used in nearly every program which is intended to increase employee performance.

The purpose of this chapter is to discuss assessment of these five types of characteristics. In doing this, we will describe the major types of instruments that are used for these assessments and summarize the research as to their use within organizations. Because of equal employment opportunity laws, significant differences in the scores of demographic groups on these assessment instruments may be justification for a legal review of the assessment procedures. For this reason we will also comment on the legal implications of using these devices. As an introduction we will briefly describe the five types of characteristics and some of the organizational programs for which the data about these characteristics are used.

Knowledge, skills, and abilities (KSAs) refer, respectively, to the amount of factual information known by an individual, his or her conduct of job-specific activities such as the operation of a particular piece of machinery, and his or her conduct of generalized job activities such as statistical analysis and verbal expression. These characteristics have been demonstrated to be strongly related to the job performance of employees (Gatewood & Feild, 1998). Therefore, assessment of KSAs is essential to programs such as recruitment, selection, job design, training, career development, and some skill-based compensation.

Personality, which has been defined and discussed in previous chapters, refers to characteristics such as thoughts, feelings, and behaviors that combine distinctly in each individual (Allport, 1961). Personality data have been included in several organizational programs. For example, these data have been used in selection since the development of the Army Beta test during World War I. Personality data have also become increasingly important in the implementation of job design and work process programs, such as autonomous work teams and employee involvement. Because these two, and other closely related, programs emphasize ongoing work interactions among team members, it is commonly thought that there must be compatibility among team members in several characteristics, including personality, in order for a team to work effectively. Additionally, personality data have also been used in leadership training and career counseling.

Assessments of physiological attributes, such as hearing, vision, strength, and coordination have been used for both selection and design of work equipment and processes. Work attitudes, such as job satisfaction and organizational commitment, are employees' cognitive and/or affective responses to aspects of the work environment (Hulin, 1991). Therefore, attitude data have been obtained in order to measure employees' reactions to a wide variety of organizational programs, such as promotion opportunities, compensation, benefits, and task activities. Also, measured change in attitudes has been the subject of several types of training programs, especially those that address the management of a culturally diverse workforce. Finally, job performance data, which are measures of various aspects of work activities or output, are extensively used in selection, promotion, training, and compensation programs.

ASSESSMENT OF KNOWLEDGE, SKILLS, AND ABILITIES

Achievement and aptitude tests, application forms, interviews, and performance tests are the assessment devices which have been used the most often to measure KSAs.

Achievement and Aptitude Tests

Achievement and aptitude tests are paper-and-pencil, usually group administered tests that emphasize factual information and its use in problem solving. The distinction between achievement and aptitude is made primarily on the basis of the use of test results (Anastasi, 1982). Achievement test results are used to assess present levels of knowledge, while aptitude tests are used to predict future performance of individuals. In reality, the two types of tests are almost identical in terms of content.

Cognitive Tests

One of the most extensively used achievement/aptitude measures within organizations involves cognitive ability tests, which are discussed in other chapters of this text as intelligence tests. Tests of this type usually tap a variety of characteristics, such as memory span, numerical fluency, general reasoning, verbal comprehension, spatial orientation, perceptual relations, and logical evaluation. One popular example of this type of test is *The Wonderlic Personnel Test* which is a 12-minute, multiple-choice instrument that consists of 50 items and has been printed in at least 14 different forms. Although cognitive ability tests have been used in selection for over 80 years, frequency of their use has diminished since the early 1970s as the result of Supreme Court decisions regarding discrimination in selection. In widely publicized cases, courts found that defendant organizations, some of whom used the *Wonderlic* as a standard part of their selection programs, were guilty of race discrimination. The basis for these cases was that the use of cognitive ability tests inevitably results in "adverse impact," that is, a significantly larger percentage of minority applicants being rejected for employment than nonminority applicants. ("Adverse impact" upon a group, usually a minority group, occurs in selection when the selection rate for applicants of any one demographic group is less than 80% of the selection rate of the demographic group with the highest selection rate. For example, if 30 of 50 [60%] male applicants are selected and if the applicant pool for females is 40, adverse impact would occur if fewer than 19 females are selected [40 x .60 x .80 = 19.2].) In these early court cases, the organizations using the cognitive ability tests did not have adequate evi-

Table 20.1. Validity Generalization Coefficients of Cognitive Ability and Job Performance for Selected Jobs

JOB	TEST TYPE	CORRECTED VALIDITY COEFFICIENT
Computer programmer[a]	Figure analogies	.46
	Arithmetic reasoning	.57
	Total score all tests	.73
First-line supervisor[b]	General mental ability	.64
	Mechanical comprehension	.48
	Spatial ability	.43
Computing and account-recording clerks[c]	General mental ability	.49
	Verbal ability	.41
	Quantitative ability	.52
	Reasoning ability	.63
	Perceptual speed	.50
	Memory	.42
Operator (petroleum industry)[d]	General intelligence	.26
	Arithmetic reasoning	.26
Police and detectives[e]	Quantitative ability	.26
	Reasoning	.17
	Spatial/mechanical ability	.17

Notes: [a]Schmidt, Gast-Rosenbery, & Hunter (1980); [b]Schmidt, Hunter, Pearlman & Shane (1979); [c]Pearlman, Schmidt, & Hunter (1980); [d]Schmidt, Hunter, & Caplan (1981); [e] Hirsh, Northrup, & Schmidt (1986).

dence to support their use in selection and, subsequently, failed to justify the adverse impact that resulted.

However, more recent research on cognitive ability tests has consistently demonstrated that these tests are empirically related to job performance and, therefore, should be legally defensible in selection programs. One example of such research is Project A, which was a seven-year study to develop a selection system appropriate for all entry-level positions in the U.S. Army (Campbell, 1990). Project A collected data from over 4,000 incumbents and developed six domains of predictor instruments. Among these domains were general cognitive ability and spatial ability. Data analyses indicated that these two cognitive ability domains were more strongly related to job performance measures than were the tests in the other four domains. General cognitive ability correlated .63 and .65 with the two performance measures that most directly measured task performance, (e.g., core technical proficiency and general task proficiency). Spacial ability correlated .56 and .63 with the same two measures.

Other evidence of the validity of cognitive ability measures has been produced by a series of validity generalization studies. The principle of validity generalization is that individual-company studies that have computed the correlation between measures of cognitive ability and job performance are hampered by a number of methodological limitations that artificially affect the resulting validity coefficient. Among the most serious of these methodological limitations are small sample size, unreliability of the predictor and criterion measures, and restriction in range on both measures. Validity generalization studies correct each of these methodological limitations in each of a group of previously conducted individual-company validation studies which have used the same predictor and criterion measures. After these corrections, a corrected validity coefficient is calculated which combines data across all of the previous studies. Table 20.1 contains results of validity generalization studies that have been conducted for cognitive ability tests for specific jobs. As can be seen, several different cognitive ability tests have been found to be significantly related to performance with corrected validity coefficients ranging from .30 to .64, the latter being quite high for a single selection measure.

A second type of validity generalization study has been conducted by combining data from studies of different jobs within the same occupation. Hunter (1986) examined the validity of cognitive ability for nine occupations. He determined corrected validity coefficients for cognitive abilities to be .61 for salespersons, .54 for clericals, .53 for managers, and .48 for service workers. Corrected coefficients for the remaining five occupations

Table 20.2. Validity Generalization Coefficients as a Function of Job Complexity

JOB FAMILIES	VALIDITY PERFORMANCE	VALIDITY TRAINING
General job families		
high complexity	.58	.50
medium complexity	.51	.57
low complexity	.40	.54

Note: Hunter (1986, pp. 340-362)

ranged from .28 to .46. These results indicate that cognitive ability is related to job performance for dissimilar jobs within an occupation and also for many occupations.

A third type of validity generalization study which further emphasized the validity of cognitive ability tests combined jobs by complexity level. Complexity level is a very broad grouping variable which measures the amount of decision making and information processing required by the job. Results, presented in Table 20.2, found that both job performance and performance in training programs were strongly related to cognitive ability. Moreover, the magnitude of the validity coefficient (Validity Performance in Table 20.2) increased as job complexity increased. The most likely explanation of these findings is that all jobs involve some information processing and problem-solving components. These components increase as jobs and occupations become more complex. Because cognitive ability tests are, in large part, measures of information processing and problem solving, they will be related to performance, especially in complex jobs and occupations. Use of cognitive ability tests for selection is expected to increase in the future as assembly-type manufacturing jobs decrease and service and technology jobs increase.

As mentioned previously, a critical issue with use of cognitive ability tests in organizations is the adverse impact which accompanies these tests. At one time it was thought that differences in mean scores between minority and nonminority demographic groups on these tests were a function of "cultural bias" (in terms of test content which was unrelated to job performance). However, research has found that differences in test performance are, in fact, consistently related to differences in job performance among individuals. Cognitive ability tests are equally valid for almost all demographic groups. For example, one study examined 781 pairs of validity coefficients for African-American and white groups (Hunter, Schmidt, & Hunter, 1979). A graph of the pattern of these coefficients was drawn for each group. The two curves were almost identical, meaning that the cognitive ability tests acted in the same manner for both African-American and white groups. These findings and those of related studies have led to the conclusion that cognitive ability tests have minimal cultural bias (that is, little, if any, non-job-performance related differences) and can be validity used for selection across demographic groups.

A related study compared the use of 18 types of predictor instruments for selection (Reilly & Warech, 1991) on the basis of validity, adverse impact, feasibility (cost of development and use), and fairness to the applicant. Results determined that cognitive ability tests were ranked very highly among the 18 test types in both validity and adverse impact (differences in test scores). Additionally, large differences were demonstrated among the test types in feasibility. Generally, cognitive ability tests were more feasible to use than other types. Analyses of fairness to the applicant included evaluations of false rejection rates, perceived relevance by the applicant of test material to the job, and potential of improvement for employability for the rejected applicant. Cognitive ability tests performed well on the first two criteria and poorly on the third. Overall, this study concluded that cognitive ability tests can be recommended because of their validity and low cost. However, other tests do demonstrate less adverse impact and are more fair to applicants, although each has major deficiencies in terms of cost and ease of development and use. (An important point should be made here with respect to adverse impact. As mentioned previously, this term signifies test score differences between demographic groups. However, these test score differences are legally justifiable if they are related to differences in job performance. That is, if test differences and job performance differences occur in the same pattern, the test can legally be used for selection. This is the case in the use of cognitive ability

tests in selection. They cause adverse impact, but this adverse impact is related to differences in job performance.)

Mechanical Ability Tests

A second widely used aptitude/achievement instrument includes mechanical tests, which measure either (a) knowledge about or skill in using tools, machines, and electrical equipment; or (b) verbal and mathematical ability to follow directions or make calculations concerning the use of mechanical tools. Within this general description, a wide variety of tests is available. These tests can have a broad array of topics (general mechanical ability) or be narrowly focused on one specific topic (e.g., electronics or welding).

An often-used general mechanical test is the *Bennet Mechanical Comprehension Test* (Gatewood & Feild, 1998). The items of this test contain objects that are almost universally familiar in American culture: airplanes, carts, steps, pulleys, seesaws, gears, and so on. The questions measure the respondent's ability to perceive and understand the relationship of physical forces and mechanical elements in practical situations. There have been six different forms of this test plus a Spanish-language version. Each form has approximately 68 items and no time limit for administration. Reported reliabilities, in the .80s, are quite favorable. Studies have correlated scores on the *Bennett* with other ability tests and have found correlations with verbal and mathematical ability and spatial visualization. Another frequently used general mechanical ability test is the *MacQuarrie Test for Mechanical Ability*. This is also a paper-and-pencil test that requires about 30 minutes to administer. It contains seven subtests: tracing, tapping, dotting, copying, location, blocks, and pursuit.

Examples of specific-topic tests are the *Purdue Trade Tests*, which include the *Test for Electricians*, *Trade Information Test in Welding*, and *Trade Information Test in Engine Lathe Operation*. Each one is a multiple-choice test of technical knowledge about tools and operations in a specific subject matter. Another type of mechanical ability test involves performance tests, such as *The Hand-Tool Dexterity Test* which measures manipulative skill with materials and equipment which is important in factory jobs and industrial apprentice training.

Mechanical ability tests have proven to be valid in selection for several specific jobs. For example, Ghiselli (1973), as part of a larger study, reviewed the use of spatial and mechanical tests as well as motor abilities' tests in selection for eight occupations. Spatial and mechanical tests were valid for use with managerial, service, vehicle, trades and crafts, and industrial occupations. Motor ability tests were useful for service, vehicle, trades and crafts, and industrial occupations. Similarly, researchers in the Project A study found that perceptual-psychomotor ability correlated .53 with core technical proficiency and .57 with general task proficiency. This type of test is also useful for diagnosing deficiencies in workers in their knowledge of mechanical principles and techniques.

Clerical Ability Tests

Traditionally, clerical jobs have been designed with a large number of tasks which focused on bookkeeping, typing, filing, and recordkeeping. To a large extent, these tasks require the employee to extensively check or copy words and numbers and to create and maintain orderly systems for the keeping of files and reports. Therefore, clerical tests have predominantly measured perceptual speed and accuracy in the processing of verbal and numeric data. Perhaps the most widely used clerical test is the *Minnesota Clerical Test,* which has two separately timed and scored subtests: number checking and name checking. Each subtest has 200 items, consisting of a pair of numbers or a pair of names. The respondent is to compare the pair and place a check on a line between the two entries of the pair if these two entries are identical. Entries in the numbers subtest range from three through 12 digits, while the entries in the names subtest range from seven through 16 letters. The score on the *Minnesota Clerical Test* is the number right minus the number wrong. Reliability has been estimated at .90 for parallel forms and .85 for test-retest. Studies have demonstrated satisfactory validity coefficients for clerical, managerial, protective, trades and crafts, and industrial occupations (Ghiselli, 1973).

Application Forms

Application-form information is frequently used by organizations in the initial step of selec-

tion programs to make general assessments of job-related KSAs. However, because many application forms ask superficial information about education, previous work history, and avocational activities, they have not demonstrated adequate validity in assessing KSAs. Therefore, research has identified specialized assessment forms that may be used in this initial stage of selection. One such form is Training and Experience (T&E) Evaluations (Ash, Johnson, Levine, & McDaniel, 1989). Although there are several different types of T&E forms, each type usually contains statements of specific, important job tasks which have been identified through job analysis. The applicant is asked to respond to each of these task statements by indicating his or her work experiences and/or training which relate to these tasks. In addition, the applicant must provide names of persons who can be contacted to verify this experience or training. This information is then graded by an organizational specialist using a scoring key. Such a scoring key usually takes into account the importance of each task for overall job performance, the extensiveness of previous experience and training, and the depth of reported knowledge.

The Weighted Application Blank (WAB) is a second type of useful application form (England, 1971). A WAB is actually a technique for scoring application forms rather than a separate assessment form. This procedure is composed of the following steps: (a) choosing a criterion, (e.g., job tenure); (b) drawing large samples of high- and low-criterion groups from the organization's employment files (e.g., individuals with more than one year of tenure and individuals who terminated employment before one year); (c) selecting application-blank items (writing items based on application information which can be expressed as questions with multiple-choice response alternatives); (d) scoring the application blank of each individual in the sample on the items selected in the previous step; (e) determining item weights (using statistical analyses to identify those items for which high-criterion respondents answered differently from low-criterion respondents); (e) applying weighted items to another sample used as a holdout group. These steps result in a scoring key which can be applied to the application forms of future applicants with the resulting score indicating whether or not the applicant should be invited back for further employment assessment (Mumford & Owens, 1987).

Biographical Data forms are a third type of application assessment. These forms usually contain several hundred multiple-choice items which ask about the applicant's previous life experiences involving social, educational, work, and family interactions. Scoring keys are developed through multivariate statistical analyses using factor analyses, clustering, and discriminant function analysis. The result is the specification of a scoring key in which applicants are selected on the basis of the similarity between their self-reported previous life experiences and those of successful employees of the organization. Biographical data inventories are written based upon hypotheses as to which life experiences may be theoretically related to success in the activities of the job of interest. Items are most often written on the following topics: (a) habits and attitudes, (b) health, (c) human relationships, (d) financial characteristics, (e) parental home, childhood, teen years, (f) personal attributes, (g) present home, spouse, and children, (h) recreation, hobbies, and interests, (I) school and education, (j) self impressions, (k) values, opinions, and preferences, and (l) work.

All three of these assessment devices have demonstrated acceptable reliability (usually above .80) and validity. For example, different studies have determined validity coefficients of .45 for T&E forms (McDaniel, Schmidt, & Hunter, 1988) and .37 for Biographical Data (Reilly & Chao, 1982). Brown's (1978) study found that a WAB used in the insurance industry was able to predict both production and tenure during a 45-year time period.

Selection Interview

The interview has been one of the most often used selection devices, because it can be applied to applicants of all job groups. However, periodic reviews of the use of this device have found that the interview frequently is characterized by limited reliability and validity in the assessment of KSAs (Schmitt, 1976). These deficiencies have, in turn, led to extensive research which has tried to determine which factors have impeded both the interview process and the interviewer's decision making (Dipboye, 1992).

One source of difficulty with the interview has been its use in assessing a wide variety of characteristics of applicants (Gatewood & Feild, 1998). Among these characteristics are job knowledge, personality traits, future work motivation, adjust-

ment to incumbent workers, verbal ability, and career development potential. While such a diversity of characteristics is theoretically possible to measure, reviews have agreed that there is evidence to support the assessment of only a limited number of these. Ulrich and Trumbo (1965) concluded that "... the interviewer is all too frequently asked to do the impossible because of limitations on the time, information, or both. ... When the task was limited ... acceptable validity was achieved" (p. 114). Three main types of KSAs have been identified as appropriately measured in the interview: job knowledge, personal relations (sociability, verbal fluency, conflict resolution, etc.), and work habits (dependability, stability of performance of tasks, ability to coordinate simultaneous projects, etc.) (Schmitt, 1976).

Several factors, which are only peripherally related to the KSAs being assessed, have been shown to influence the evaluation of the interviewer. Physical attractiveness of the applicant and personal liking of the applicant by the interviewer are two such factors (Keenan, 1977). A variety of nonverbal behaviors of the applicant, such as eye contact, head movement, smiling, hand motions, and general body posture has also been identified (Dipboye, 1992). The disproportionate influence of any negative information on the overall evaluation has also been found (Rowe, 1963). Related to this is the finding that the interviewer frequently makes an overall assessment of the applicant within the first few minutes of the interview (Ulrich & Trumbo, 1965). In addition, the contrast effect, associated with previous applicants (Valenzi & Andrews, 1973), is among many information-processing factors that have been studied (Dreher & Sackett, 1983).

Because of the frequency of use of the selection interview and its traditionally low reliability and validity, it has been the subject of several equal-opportunity court cases. These cases have yielded opinions about specific features of the interview which are presumed to be sources of adverse impact: having all male and/or white interviewers, not using a structured or written interview format, not having stated criteria for employment decisions, and not using uniformity in applying selection criteria (Ledvinka & Scarpello, 1991). Also, research has determined that females are often given lower evaluations than comparable males when the jobs of interest are those frequently thought of as "male" jobs (Haetner, 1977).

As a result of the previously mentioned research and court decisions, there have been several suggestions to ensure that the interview is both reliable and valid. One of the most effective of these suggestions has been to use behaviorally based interview questions. One technique for developing such questions is the situational interview. The intent of this technique is to identify specific activities which represent important job tasks and use this information to form questions that ask an applicant how he or she would behave in the situation (Latham, Saari, Pursell, & Campion, 1980). Various studies of the situational interview have reported validity coefficients ranging from .30 to .46 and interrater reliability estimates between .76 and .87 (Harris, 1989). A second technique is the Behavioral Descriptive Interview (Janz, 1982). This type of interview is similar to the situational interview in its identification of job activities. However, resulting questions are different in that they ask about general performance behaviors rather than specific job situations. They also contain probe questions. Validity coefficients of .54 and .48 have been reported for the Behavioral Descriptive Interview (Harris, 1989).

A second, related suggestion for improvement has been to structure the interview. This consists of such steps as training all interviewers in a particular style of interviewing, identifying KSAs to be evaluated in each interview, specifying the main questions to be asked of all applicants, and developing a formal scoring system and decision rules. Meta-analytic studies have determined corrected validity coefficients of .62, .47, and .49 for the structured interview, indicating that it is roughly equivalent to cognitive ability tests for selection purposes (Gatewood & Feild, 1998).

Performance Tests

Performance tests are assessment devices that present testing situations that closely resemble actual job tasks and require the individual to complete some activity under structured testing conditions. In addition to their use in selection, these tests have also been extensively used in training in order to diagnose an individual's present KSAs and identify any deficiencies. The two most often used types of performance tests are work samples and assessment centers.

Work Sample Tests

Work sample tests have been used for the selection and training of clerical staff, technicians, and skilled crafts. Specific tests require such activities as typing, welding, wiring connections, cutting and nailing boards, and mixing chemicals. For example, Robinson (1981) describes a work sample test used for construction superintendents. An architect was retained to identify common architectural errors in blueprints and to incorporate these errors into the drawings of buildings which had actually been built by the company. Applicants were asked to review the blueprints and to mark the location of the errors with a felt-tipped pen on copies of the drawings.

Evaluations of the use of work sample tests have been very positive and have identified a number of benefits. Gordon and Kleiman (1976) compared correlations of work sample and paper-and-pencil ability tests with success in a police training course using three different groups of subjects. For each group, correlations of the work sample test with success were superior to those correlations using the ability tests. A meta-analytic study found a corrected validity coefficient of .54 between work samples and job performance (Hunter & Hunter, 1984). Additionally, Schmidt, Greenthol, Hunter, Berner, and Seaton (1977) compared scores of minority and nonminority applicants who were applying for positions as metal trade apprentices on both written tests and work samples. A difference in scores between the two groups was identified for only one of three work sample tests but for all five of the written tests. Finally, Cascio and Phillips (1979) found that work samples served as realistic job previews. That is, work sample tests provided a representative preview of the actual job, including positive and negative aspects. As a result, some applicants finding the work sample to be unsuitable, removed themselves from the selection program. The net effect was to reduce the turnover rate among new hires, resulting in large savings for the organization.

Assessment Centers

An assessment center (AC) is a standardized evaluation of multiple behaviors using multiple assessment devices, many of which must be simulations (Task Force on Assessment Center Guidelines, 1989). These multiple behaviors are referred to as dimensions in the AC. Several trained assessors observe the candidates' performance during each of these assessment devices, meet to discuss each applicant's performance, and use these data to develop ratings for each dimension as well as an overall rating of goodness of performance.

Two of the most often used simulations in the AC are the In-Basket and the Leaderless Group Discussion (LGD). The In-Basket is a written simulation, of up to 30 memos, that is intended to replicate administrative tasks. In responding, the assessee writes the specific actions that he or she would take in reference to each memo, even naming individuals to be involved in these actions. Typically, three hours are allotted for completion. This device is frequently used to assess KSAs of planning, organizing, ability to delegate, decisiveness, independence, and initiative. The LGD is designed to represent those managerial behaviors that require the interaction of small groups of individuals to solve a problem. In the LGD, participants are tested in groups of six and seated around a conference table. Assessors are seated at various places around the participants to observe and record their behavior. The problem presented to the six-person group commonly is an allocation-of-resource dilemma in which there are more demands for the resource than there is supply. Each participant is asked to play a specific role and is provided background information. The group is usually given 1.5 hours to complete the task. Commonly, KSAs such as oral communication, tolerance for stress, persuasiveness, and adaptability are measured.

A meta-analysis (a statistical analysis for combining data across multiple samples) of ACs reported corrected validity coefficients of .53 for predicting career development and .36 for predicting immediate job performance (Gaugler, Rosenthal, Thornton, & Bentson, 1987). These results indicate that ACs measure KSAs that are related to long-term success within organizations. Another positive feature of ACs is their generally favorable support by courts in alleged discrimination cases. For example, in *The Richmond Black Police Officers Association v. the City of Richmond* the use of an AC for selection for supervisory positions in both the police and the fire departments was upheld. Evidence indicated that while the paper-and-pencil tests which were used in selection had severe adverse impact, the AC had no such effect and compensated for the adverse impact of the written tests.

PERSONALITY ASSESSMENT

There are several indications that personality should be a significant factor in job performance. First, job analysis has determined that many positions, such as receptionist, salesperson, and manager, require the incumbent to interact with others during most of the task activities. Other jobs, such as air traffic controller, law enforcement officer, and high school principal, require the incumbent to cope with many sources of stress. As a result, many job analysis methods (e.g., the Position Analysis Questionnaire) (McCormick & Jeanneret, 1988) include personality dimensions among the KSAs that are routinely identified. Second, empirical studies have noted differences in personality between successful and unsuccessful job performers. For example, Grimsley and Jarrett (1975) concluded that differences were evident between successful and unsuccessful managers in drive, energy, social adjustment, self-confidence, social aggressiveness, and emotional stability. Similarly, a longitudinal study of managers at AT&T found personality differences between those managers who had been promoted to middle-management positions and those who remained at lower managerial positions during an eight-year period (Bray, Campbell, & Grant, 1979).

Despite these indications, various reviews of the validity of personality assessments in predicting job performance have found very low empirical relationships between the two variables. Guion and Gottier (1965) stated that the number of significant findings in their review of validity coefficients was barely above the chance level of occurrence. They partially attributed these poor findings to methodological deficiencies in many of the validity studies. Similarly, Kinslinger (1966) found that the frequency of methodological shortcomings in studies using projective techniques prevented any positive judgment of their validity. Finally, Schmitt, Gooding, Noe, and Kirsch (1984) conducted a meta-analysis which determined a corrected validity coefficient of only .15 for self-report personality questionnaires.

However, in opposition to these reviews, Hogan (1991) has provided evidence strongly supporting the relationship of personality data to performance and argues that, to a large extent, previous findings were due to the use of inappropriate assessment devices and the lack of methodological rigor in the studies. Recent work has developed three theoretical and empirical arguments for the relationship of personality and job performance. First, there is a growing agreement among I/O psychologists and HR specialists that personality characteristics can be grouped into the broad personality dimensions referred to as the Big Five. In the past, researchers used a variety of personality traits in selection and, consequently, a variety of instruments, many of which focused on narrow personality traits. Agreement on the use of the Big Five provides consistency among researchers in terms of the constructs that are to be studied. Second, there is evidence that studies which are methodologically sound often yield significant, uncorrected validity coefficients of .30 and higher. Such coefficients are generally comparable to those of other, well accepted, selection instruments. Third, personality data have been found to be uncorrelated with cognitive ability and, therefore, can be additional predictors of job performance. In other words, personality data add unique variance to the prediction of job performance.

Self-Report Inventories

One of the most popular methods of assessing personality is through self-report questionnaires. These questionnaires present the respondent with a statement of feelings about an object, person, or activity and request the respondent to indicate agreement, neutrality, or disagreement with the statement. Several such measures of the Big Five have been developed. Of these multiple measures, we will describe the measure of the Big Five developed by Barrick and Mount (1991) because it has been used in several organizationally based studies. Their *Personality Characteristics Inventory* has 200 multiple-choice items derived from empirical research of several self-report personality inventories. Each of the 200 questions has three possible responses, "agree," "?," and "disagree." The inventory takes approximately 45 minutes to complete. The first dimension is Extraversion. Traits used to form this dimension include (from the positive pole) being sociable, gregarious, assertive, talkative, and active. The second dimension is Emotional Stability, characterized (from the negative pole) as being emotional, tense, insecure, nervous, excitable, apprehensive, and easily upset. The third dimension is Agreeableness, and is made up (positive pole) of being courteous, flexible, trusting, good-natured, cooperative, forgiving, softhearted, and tolerant. Conscientiousness is the

fourth dimension and is composed (positive pole) of being responsible, organized, dependable, planful, willing to achieve, and persevering. The fifth dimension is Openness to Experience (sometimes referred to as Intellect or Culture in other work concerning the Big Five). This factor includes (positive pole) being imaginative, cultured, curious, intelligent, artistically sensitive, original, and broad-minded.

Among their several studies Barrick and Mount (1991) conducted a meta-analysis of the use of the Big Five in selection. Using data from 117 validity studies grouped across five occupations and three measures of job performance, they drew the following conclusions:

1. Conscientiousness is a valid predictor of job performance for all five occupational groups (corrected r of .20 to .23).
2. Extraversion and Emotional Stability were valid predictors for some, but not all, of the five occupational groups.
3. Agreeableness and Openness to Experience showed minimal validity, although these results may be partially attributable to the small number of studies which examined these two dimensions.
4. Conscientiousness was related to each of the three performance measures used. The other four personality factors were not related to all measures of performance.

The main conclusions from this work are that: (a) the factor of Conscientiousness appears to be the most important personality factor to be used in explaining job performance; (b) not all of the Big Five factors are related to job performance, perhaps explaining the weak empirical validity determined from previous reviews; and (c) personality, unlike cognitive ability, is not related to job performance in almost all jobs. Rather, specific factors of the Big Five are differentially related to specific job performance criteria within specific occupations.

Related studies have provided further information about the role of personality in work performance. Barrick, Mount, and Strauss (1993) found that sales representatives high in conscientiousness were more likely to set goals and were more likely to be committed to goals, which in turn was associated with greater sales volume and higher supervisory ratings of performance. Contrary to hypotheses, extraversion was not related to performance for these sales representatives. These findings indicate that goal-directed behavior is a stronger determinant of sales performance than is the ability to interact easily with others. A related study (Mount, Barrick, & Strauss, 1994) found that supervisor and coworkers' ratings of conscientiousness and extraversion were valid predictors of performance ratings for a group of sales personnel, indicating that coworkers recognize the importance of personality factors in job performance. Finally, two other studies have provided some evidence that personality data may add validity above that associated with cognitive ability, although neither study used all the dimensions of the Big Five (Day & Silverman, 1989; Wright, Kacmar, McMahan, & Deleeuw, 1992).

Projective Instruments

Although self-report personality questionnaires are used extensively in organizations, there are instances of the use of projective instruments, especially in studies that involve high-level managers. Two instruments that have been used in such work are the *Thematic Apperception Test* (TAT) and the *Miner Sentence Completion Scale* (MSCS). In the TAT, the respondent is asked to tell a story about each of 19 cards that depict one or more individuals in a variety of ambiguous situations. It is assumed that the content of the individual's stories about these cards will reveal unconscious desires, inner tendencies, attitudes, and conflicts. The cards are administered individually in two one-hour sessions. The most often used scoring system for the TAT was developed by Murray and analyzes the hero (leading character in each story), the needs of the hero (such as achievement, order, and aggression), press (the pressures operating on the hero), and themes (the interplay among needs, press, and resolution of conflict) (Murray, 1943).

The MSCS was developed for the assessment of motives that are manifested in managerial work. The respondent is presented with 40 sentence fragments and asked to complete each (Miner, 1977). These items load on seven scales. *Authority figures* provides a measure of the subject's capacity to meet role requirements in relationships with a superior. *Competitive Games* and *Competitive Situations* both focus on occupational or work-related competition. The *Assertive Role* generally reflects

confidence in one's ability to perform well and a wish to participate in activities. *Imposing Wishes* refers to controlling or directing the behavior of others. The *Standing Out from the Group* scale measures the desire to assume a somewhat deviant position as compared with subordinates. Finally, *Routine Administrative Functions* indicates the desire to meet job requirements of day-to-day administrative work.

Cornelius (1983) reviewed the use of projective tests in organizations and concluded that those tests which have defined scoring systems, especially the TAT and the MSCS, have demonstrated reliability of .80 or higher. He also reported that 10 of 14 studies appearing in academic journals between 1960 and 1981 reported significant validity coefficients between scores on projective measures of personality and job performance.

ASSESSMENT OF PHYSIOLOGICAL RESPONSES

There are three uses for the assessment of specific physiological characteristics of individuals in work situations: (a) to enhance selection decisions, (b) to assist in the diagnosis of poor performance of current employees, and (c) to provide information for medical insurance. There are four common types of assessment of physiological responses: vision, hearing, strength and coordination, and drug testing.

Vision Testing

Visual sensitivity includes several separate functions. For industrial work, the most important are color discrimination, near acuity at reading (13 to 16 inches), far acuity (usually measured at 20 feet), depth perception, and muscular balance of the eyes (phoria). The most common measure of vision is the Snellen Chart, which contains rows of letters of gradually decreasing size. It is intended to measure only far acuity. Accuracy in reading the letters of this chart has been found to be affected by many factors in a normal employment testing situation: amount of illumination, distance from chest, position of examinee's head, and so on. For this reason, more accurate and complete visual measures are taken by using specially designed instruments such as the Ortho-Rater, the AO Sight Screener, and the Keystone Telebinocular. These instruments provide measures of all the visual characteristics mentioned previously.

Hearing Testing

The most important aspect of hearing for industrial work is auditory acuity—the faintest sound that the individual can just barely hear. The most reliable measurement of this sound involves electronic audiometers. With these instruments, one ear at a time is tested. During testing, the subject receives the sound through a headphone pressed against the ear. The examiner increases the decibel level of the transmitted sound until the subject indicates that sound has been heard. This sound threshold is then remeasured by starting with a clearly audible sound and decreasing the decibel level until the subject reports no hearing. At each sound-wave frequency, the subject's hearing loss in decibels can be determined from the audiometer dial. This dial has been calibrated with a standard of "normal hearing" for the population. These normal hearing levels have been determined through testing a large, representative sample of people.

Strength and Coordination

There are two major taxonomies of physical abilities which identify those characteristics that are necessary for carrying out work assignments. The first is the *Ability Requirements Scale* which measures the following nine physical abilities (Fleischman & Mumford, 1988):

1. static strength—maximum force that can be exerted against external objects. Tested by lifting weights.
2. dynamic strength—muscular endurance in exerting force continuously. Tested by pull-ups.
3. explosive strength—ability to mobilize energy effectively for bursts of muscular effort. Tested by sprints or jumps.
4. trunk strength—limited dynamic strength specific to trunk muscles. Tested by leg-lifts or sit-ups.
5. extent flexibility—ability to flex or stretch trunk and back muscles. Tested by twist and touch test.

6. dynamic flexibility—ability to make repeated, rapid, flexing trunk movements. Tested by repeated rapid bending over and touching floor.
7. gross body coordination—ability to coordinate action of several parts of body while body is in motion. Tested by cable jump test.
8. gross body equilibrium—ability to maintain balance with nonvisual cues. Tested by rail-walk test.
9. stamina—capacity to sustain maximum effort requiring cardiovascular exertion. Tested by 600-yard run-walk.

The following validity coefficients for performance on specific jobs are among the results of studies which examined the use of these scales: pipeline workers (.63); correctional officers (.64); warehouse workers (.39); electrical workers (.53); and enlisted army personnel (.87). All coefficients are the product of a battery of two to four physical abilities correlated with job performance (Fleischman & Mumford, 1988).

The second taxonomy is the product of extensive work of Joyce Hogan who combined two lines of research. The first was data about physical requirements that were derived from job analysis and the second was data based on physical ability tests already developed for selection. Factor analyses were performed on several sets of data from these two sources. Results consistently identified three factors of physical abilities (Hogan, 1991). The first, muscular strength, is the ability to apply or resist force through muscular contraction. Within this factor are three, more specific, dimensions: muscular tension, muscular power, and muscular endurance. The second factor is cardiovascular endurance, which refers to the capacity to sustain gross muscular activity over prolonged periods. It is aerobic capacity and general systemic fitness involving the large muscles. The third factor, movement quality, concerns characteristics that contribute to skilled performance and also contains three dimensions: flexibility, balance, and muscular integration.

Specialists using results from these or other physical ability instruments must be especially concerned by the legal issues associated with testing three groups of individuals: females, disabled, and older workers. Adverse impact for scores on physical ability tests is common for each group. Frequently, males will score higher than females on such tests, nondisabled persons higher than disabled, and younger persons higher than older. The essential issue, therefore, is that the tests must clearly be linked to critical job tasks, which require these physical abilities in their completion. However, even this principle is complicated by the question of whether the tasks can be modified to reduce or eliminate physical demands. If such modifications can be made, the use of the physical ability test which is appropriate for the original tasks may be unwarranted when used with the modified job activities.

Drug Testing

According to some estimates, drug abuse costs organizations over $30 billion annually through absenteeism, mistakes, damage, injury, and sick leave. In response, organizations are increasing the use of drug testing for both applicants and incumbent workers. The most common test is *The Immunoassay Test,* which attempts to determine whether drugs are in a person's system on the basis of the reaction of the urine specimen to certain antibodies created by the immune systems of laboratory animals. This test can detect both the presence and absence and the amount of drugs in a person's system. Another, more precise, test is the *Gas Chromatography/Mass Spectometry Test*, which can separate complex mixtures of drugs and other substances into their pure parts. When a mixture, such as an extract of drugs from urine, is injected into the testing instrument, each drug will move through the instrument in gas form at a different speed. When a particular drug reaches the end of the instrument and enters the mass spectrometer, it is separated from other drugs and is in a pure form. Therefore, this test can identify a specific drug from a group of similar drugs.

Although these tests are quite reliable and valid in their identification of drugs in an individual, there are several legal issues accompanying their use. For one, there is no definite relationship between the presence of drugs and work performance. The test does not indicate how much drug was ingested, when it was ingested, or its effect on physical processes and cognitive functions. Second is the issue of invasion of privacy. By their very nature, drug assessment procedures are intrusive upon the individual. When such intrusiveness is coupled with the tests' inability to measure decreases in work performance, one argument against drug testing is that unless there is evidence

of the individual's inability to perform work safely and efficiently, employers are invading the off-work time of employees and are attempting to regulate their private lives. For this reason it is important that testing be linked to workplace problems such as accidents, theft, absenteeism, and sabotage. Data that indicate the existence of such problems before testing can serve as evidence that the company is pursuing a legitimate self-interest.

ASSESSMENT OF WORK ATTITUDES

As mentioned previously, data about employees' attitudes are used to assess strengths and weaknesses in both the organization's programs and its physical facilities. Scarpello and Vandenberg (1992) have summarized guidelines for the construction and use of attitude-assessment instruments which gather data for these purposes. Because space limitations do not permit us to discuss each of the large number of specific attitudes which have been assessed, we will only discuss two of the most often studied attitudes: job satisfaction and organizational commitment.

Job Satisfaction

Locke (1976) defined job satisfaction "...as a pleasurable or positive emotional state resulting from the appraisal of one's job or job experiences" (p. 1300). While this definition is generally accepted by both researchers and practitioners, there is no certainty that the measurement of satisfaction reflects this definition, according to Organ and Near (1985). They point out that job satisfaction is typically measured with instruments modeled after attitude scales. Because psychologists have viewed attitudes as an assemblage of cognitive, emotional, and action tendencies, factors other than emotional states have entered into the measurement of satisfaction. After reviewing these instruments, Organ and Near (1985) concluded that "...the items on most job attitude scales tend to focus on the job itself or the factors of the job (e.g., supervision, the task, pay), not on the feelings of the respondent" (p. 244). The wording and format of items on many job-satisfaction surveys actually require a cognitive evaluation of the work situation. For example, the respondent may be asked to judge various aspects of the job relative to how much of this aspect there should be in an ideal job or how much of this aspect he or she expected from the job.

Recent research has provided some explanation of the differences in the affective and cognitive facets of job satisfaction. Olson and Zanna (1993) discuss that, while the view of attitudes as cognitions, affective reactions, and behaviors provides a useful framework for examining attitudes, not every attitude must contain each of these three components. Rather, specific attitudes toward the same object may be based on either cognitive, affective, or behavioral antecedents. For example, Edwards (1990) induced subjects to form either affect-based or cognition-based attitudes through the experimental manipulation of varying the order of reading about and tasting a soft drink. Millar and Millar (1990) also classified subjects as possessing either affect-based or cognitive-based attitudes toward the same object. It seems, therefore, that job satisfaction can be either an affective and/or cognitive attitude, each of which is based upon different antecedents in the workplace. For example, Judge (1993) found support for the proposition that an individual's disposition to be satisfied with everyday life events moderates the relationship between affective job satisfaction and organizational outcomes. That is, the more positive the disposition of the individual, the stronger was the relationship between job dissatisfaction (affective attitude) and turnover.

The instruments used to measure job satisfaction are generally classified into two groups: measures of overall satisfaction and measures of satisfaction of specific job facets. Both types typically employ Likert-type items as measuring devices. Cook, Hepworth, Wall, and Warr (1981) reviewed and illustrated many of the currently used major instruments in the following manner. Measures of overall satisfaction differ in terms of the number of items and the content of questions; items range from four to 38 while the content varies from questions about a worker's emotional reactions toward the job as a whole, to cognitive reactions of organizational or supervisory functioning, to evaluations of specific intrinsic and extrinsic features of jobs. The score obtained from measures of overall satisfaction is obtained by summing the scores of all individual items. Reported reliabilities, (generally internal consistency), are usually at least .80. Most measures also report criterion and construct validity in the form of relationships with other organizational variables.

Table 20.3.

SCALES OF JOB SATISFACTION FACET MEASURES	
supervision	social needs
company as a whole	autonomy
nature of work	personal growth
extent of work	esteem needs
coworkers	subordinates
working conditions	intrinsic rewards
pay	extrinsic rewards
promotions	friend's attitudes
security	family attitudes

Note: Cook, Hepworth, Wall, & Warr (1981).

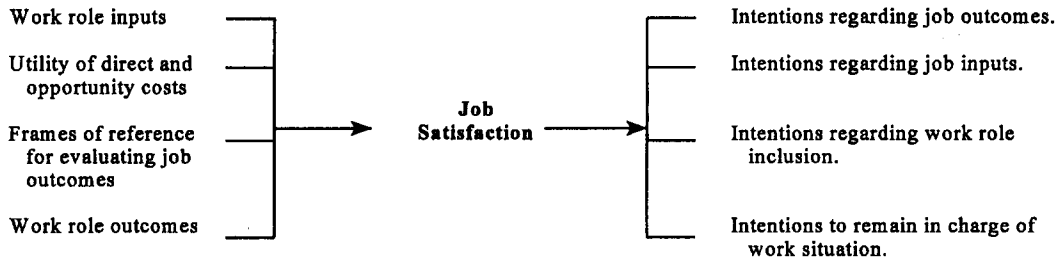


Figure 20.1. Model of Antecedents and consequences of Job Satisfaction

A variety of topics (see Table 20.3) is measured in job facet satisfaction, with each instrument being composed of a different combination of facets. The facets that are included most often in these questionnaires are satisfaction with pay, promotions, supervision, and job content. As with the overall satisfaction measures, most of the reported reliabilities for facet measures are internal consistency estimates, which most often are at least .80. Validity data for measures of facet satisfaction are usually reported in terms of correlations with other satisfaction measures and correlations among the various scales of the instrument.

Much research has examined the organizational correlates of job satisfaction. Roznowski and Hulin (1992) argue that, while assessment of cognitive ability is the most important information that an organization can have about individuals prior to organization entry, measures of job satisfaction are similarly important after-entry data. This is because job satisfaction has been demonstrated to influence variables, such as employee attendance at work, decisions to leave an organization, decisions to retire, general behavioral syndromes reflecting pro-organizational orientations (organizational citizenship), attempts to change work situations by voting for union representation, and psychological withdrawal behaviors.

Figure 20.1 is a variation of a model proposed by Roznowski and Hulin to explain the factors which influence job satisfaction and the variables which are affected, in turn, by satisfaction. Briefly, external employment and economic factors, an individual's input into his or her work, the outcomes given by the organization for this work, and the individual's evaluation of these outcomes lead to the individual's affective satisfaction/dissatisfaction with work. Such affective reaction directly influences the individuals subsequent behavior. Satisfied and dissatisfied workers react differently to the same work situation. Organ and Konovsky (1989) suggest that the cognitive component of job satisfaction is responsible for premeditated, intentional, and sustained contributions to the organization. This is because the satisfied individual wishes to maintain fairness in the social exchange relationship between him or her and the organization. On the other hand, dissatisfied workers will do something to reduce their negative feelings. Because the actions that are employed to reduce these negative feelings are often detrimental to the well-being of the organization, it is important that the organiza-

tion be aware of job dissatisfaction through regularly administered measures of employees' attitudes. The actions taken by individuals in response to dissatisfaction usually include attempts to increase work outcomes, attempts to decrease work inputs, attempts to reduce work role inclusion, and formal attempts to change the job/work role.

Organizational Commitment

Steers (1977) defined commitment to an organization as the relative strength of an individual's identification with and involvement in that organization. He also discussed three factors of such identification. One is an individual's strong belief and acceptance of the organization's goals. A second is the individual's willingness to exert considerable effort on behalf of the organization. The third is the individual's desire to maintain membership in the organization.

Based on this definition, organizational commitment is frequently measured by multiple-item scales, such as the one described by Porter, Steers, Mowday, and Boulian (1974).

> This 15-item questionnaire was designed to measure...commitment....Included in this instrument are items pertaining to the subject's perceptions concerning his loyalty toward the organization, his willingness to exert a great deal of effort to achieve organizational goals, and his acceptance of the organization's values. All items represent statements to which the subject responds on 7-point Likert scales, ranging from "strongly disagree" to "strongly agree" (p. 605).

Marrow (1993) has reviewed studies of commitment and concludes that there are two perspectives in the academic community as to the nature of commitment. One perspective focuses on calculative/continuance commitment which centers on the exchange components of the employee-organization relationship. That is, this type of commitment focuses on both perceived ease of movement by the individual from the present organization to another one, and also on those variables which might be lost if a person were to leave an organization (e.g., seniority, relationships, retirement benefits, and insurance). According to this perspective, the individual is bound to the organization because of these variables. Calculative\continuance commitment has been correlated with a number of antecedents and outcomes. For example, age of respondent, organizational tenure, and level within the organization are all significantly, if weakly, correlated to this type of commitment. Other research indicates that employees may form calculative/continuance commitment with an organization based on both their feelings about the organization and the costs of leaving. That is, provision of an excellent health care insurance policy might engender feelings of loyalty based on a positive emotional attachment to the organization (i.e., the organization really cares about me) or a more "sunk costs" orientation (i.e., few firms could offer me an equivalent policy). Reliabilities of measures of this construct were usually reported at .80+.

Affective commitment is the second form of commitment and, conceptually, is identical to the construct defined by Porter and colleagues (1974). Their measure of commitment is by far the most extensively used instrument to measure this construct. Its reported internal consistency reliability averaged .88 over 90 samples (Mathieu & Zajac, 1990). Factor analysis of the instrument yields a single-factor solution that is independent of work ethic, career commitment, professional commitment, and job involvement (Marrow, 1993). Affective commitment has been the subject of extensive empirical research which, in turn, has served to more precisely define the construct. For example, affective commitment is positively related to age, organizational tenure, group cohesion, communication, more complex job design, and job satisfaction. Negative relationships have been found with education, stress, absenteeism, intention to leave an organization, and actual turnover. Because of this extensive research, affective commitment is considered to be an important work attitude which is critical in understanding employees' behaviors and performance.

ASSESSMENT OF WORK PERFORMANCE

Assessments of individual and group work performance are important for organizations because they directly relate to the organization's profitability, competitive advantage, and survival. Therefore, one use of such performance data is in formulating organizational strategy and setting goals. A second use is to validate recruitment, selection, and training programs. Third, performance data are also prominent in the distribution of compensation among employees. Although there are several different measures of work per-

formance, we can classify these measures into three main groups.

Production Data

Production data consist of the results of work. The data comprise things that can be counted, seen, and compared directly from one worker to another. Related terms that have been used to describe these data are output, objective, and nonjudgmental performance measures. These measures are usually based upon the specific job tasks and quite different measures can be used for the same job title. Usually the various measures reflect some aspect of either quantity or quality of completed product/service, for example, number of units produced per week, dollar volume of sales, profit of unit, number of errors per hour, weight of scrap, and number of complaints from customers.

These data are often considered to be the most desirable measure of performance for several reasons. First, such data are often easy to gather because they are routinely collected for business operations such as production planning and budgeting. Also, the importance of such measures is thought to be obvious. Finally, these data are understood and accepted by workers as indicators of their performance.

However, there are measurement limitations in the collection of production data which should be considered before they are used. For one, these data can be deficient because the measure usually focuses on only one aspect of performance (e.g., volume of sales), and does not include other important aspects (e.g., service after sale). Also, production data are often collected for work groups or administrative units rather than for individuals. While group-level performance data are certainly appropriate for some HR decisions within specific managerial frameworks (e.g., Total Quality Management), they are inappropriate for others (e.g., individual achievement and bonus plans). Finally, there are differences in work situations among individuals that lead to nonstandardized assessment conditions among employees. For example, sales personnel are usually assigned to territories that differ greatly in terms of sales opportunities, economic purchasing power of residents, and the number of competitors. As a result, the sales volume of a salesperson may more strongly reflect characteristics of the territory rather than characteristics of the individual. Organizations frequently attempt to adjust specific sales data to control these differences but such adjustments are usually based upon opinions rather than statistical data and are subject to measurement limitations themselves.

Employee Data

Absenteeism, turnover, grievances, accidents, and promotions are the most common variables in the second type of performance data, employee data. These variables have characteristics which are similar to those of production data. For one, they are often routinely collected; however, unlike production data, this collection is usually at the individual level. Also, data are countable and appear to be free of bias. Finally, they are generally accepted as being important parts of work behavior.

However, similar to production data, employee data also have important measurement limitations. First, there is often very little variance in these measures because the large majority of workers have very few absences, accidents, grievances, or promotions. This is especially true for data that are collected for short time periods. Second, each of the variables in this type of data have several different operationalizations which are not interchangeable. Therefore, the measurement specialist must be clear about the construct that is to be measured. For example, absenteeism has been measured as: the number of separate instances, the total number of days taken, the number of short (one-or two-day) absences, and the number of Monday absences (Landy & Pharr, 1983). Similarly, turnover can be either voluntary (the individual initiates the separation) or involuntary (the organization initiates the separation). The two are quite different in terms of the construct being measured. Finally, employee data can also be deficient as a measure of overall work performance because they do not directly address either quantity or quality of production/service.

Judgmental Data

In judgmental data an individual familiar with the work of another is required to judge this work. In most cases this individual is the immediate supervisor. However, 360-Feedback, which is becoming increasingly popular, uses data from subordinates, peers, and supervisors to gather

Table 20.4. Dimensions Used in Judgmental Performance Review

TRAIT DIMENSIONS SCALES

DIMENSIONS	UNSATISFACTORY	MARGINAL	SATISFACTORY	GOOD	SUPERIOR
Attitude	1	2	3	4	5
Enthusiasm	1	2	3	4	5
Cooperation	1	2	3	4	5
Motivation	1	2	3	4	5

BEHAVIORALLY ANCHORED RATING SCALE

PERFORMANCE DIMENSION: INTERACTING WITH BANK CUSTOMERS

SCALE POINT	BEHAVIOR
7	Employee smiles, greets customer by name, asks to be of service
6	Employee greets customer by name, asks to be of service
5	Employee smiles, asks to be of service
4	Employee asks to be of service
3	Employee greets customer
2	Employee smiles at customer
1	Employee remains silent until customer speaks

BEHAVIORAL OBSERVATION SCALE

PERFORMANCE DIMENSION: REVIEWS SUBORDINATES WORK PERFORMANCE

1.	Communicates mistakes made in job activities by subordinates.						
	Almost Never	1	2	3	4	5	Almost Always
2.	Praises subordinates for good work behavior.						
	Almost Never	1	2	3	4	5	Almost Always
3.	Discusses hindrances in completing projects.						
	Almost Never	1	2	3	4	5	Almost Always
					♦		
					♦		
					♦		
11.	Inspects quality of output materials.						
	Almost Never	1	2	3	4	5	Almost Always
12.	Reviews inventory of necessary parts and equipment.						
	Almost Never	1	2	3	4	5	Almost Always

Total Score _____

complete data about an individual's work performance. Because judgmental data are opinions of individuals, a number of measurement limitations must be addressed before such data can be used. For example, Murphy and Cleveland (1991) have identified a number of errors that are frequently made by raters (halo, leniency, central tendency, severity) that reduce both reliability and validity. They also discuss a number of effective steps which can be taken to limit these rating errors and increase the psychometric properties of the instrument used. One step is to develop the performance appraisal system in such a way that the raters are asked to evaluate job behaviors rather than personal traits of individuals. Table 20.4 presents examples of both behaviors and traits that have been used in judgmental performance review. Research has determined that the use of traits is unacceptable to all parties involved in the performance review. That is, neither the rater nor the ratee are receptive to the rating of personal characteristics. Second, the theoretical basis for the presumed relationship between the traits and performance has not been well specified or empirically verified. Because of these deficiencies in the evaluation of traits, judgmental data have focused on the individual's performance of specified job behaviors.

The two most frequently used types of job behavior scales are Behaviorally Anchored Rating Scales (BARS) and Behavioral Observation Scales (BOS). Examples of both of these types of scales are presented in Table 20.4. Development of BARS includes gathering descriptions of specific work behaviors, sorting these incidents into dimensions of work performance, and determining the scale points to be assigned to each incident within a dimension (Bernardin & Beatty, 1984). When using this type of scale, the rater is asked to circle the behavior of each dimension which reflects the actual work behavior of the individual being evaluated. The individual being evaluated, therefore, is assigned the points associated with the behavior which is circled.

The BOS technique is similar to BARS in that descriptions of specific work behaviors are gathered and sorted into separate dimensions. However, multiple scales are produced for each work dimension. That is, the specific work behaviors which have been sorted into each dimension are used to form a number of scales for that dimension. As is indicated in Table 20.4, each of these behavior scales uses a measure of frequency as its response format. Therefore, when the rater uses the BOS technique, he or she reviews each of these scales and rates the frequency of occurrence of each behavior in the BOS instrument in the work performance of the individual being evaluated (Latham & Wexley, 1981).

Because judgmental performance data are extensively used in many organizational programs (e.g., compensation, promotion, termination, training), they have been part of several court disputes. These court cases have produced rulings which have comments on specific characteristics of judgmental performance systems. According to McEvoy and Beck-Dudley (1991), characteristics that courts have regarded as being necessary for defense against illegal discrimination are:

1. presence of written instructions given to raters which describe the performance review system.
2. use of behaviorally based dimensions for judgment rather than trait dimensions.
3. basing the behaviorally based dimensions on formal job analysis.
4. having employees review the appraisal results in a formal session in which feedback is given by the supervisor.
5. presence of an appeal channel for use by employees who wish to disagree with the ratings of a supervisor.
6. having multiple raters, such as the employee's immediate supervisor and a manager one level above the immediate supervisor, participate in the evaluation.

SUMMARY

Assessment of employee characteristics is necessary for development and implementation of various programs that seek to improve the work performance and psychological reactions to work of employees. The five major classes of employee characteristics that have been assessed for such programs are: knowledge, skills, and abilities, which are especially important for the performance of the technical portions of tasks; personality, which is necessary for work group interaction; physiological attributes, which are essential for acquiring information and completing the physical components of tasks; attitudes, which reflect an individual's emotional reactions to or cognitive evaluations of features of the job or the organization; work performance, which is a primary indicator of the employee's contribution to the success of the organization and has been linked to the organization's ability to function in the competitive market.

As this chapter describes, in addition to development of organizational programs, scientists have used data from assessment of these five characteristics of individuals to study the relationships of these five with numerous other variables. The results of these studies have produced an understanding of the antecedents and consequences of individual work attitudes and work performance. Such understanding, in turn, has contributed to our understanding of human behavior.

REFERENCES

Allport, G. (1961). *Pattern and growth in personality.* New York: Holt, Rinehart & Winston.

Anastasi, A. (1982). *Psychological testing* (5th ed.). New York: Macmillan.

Ash, R., Johnson, J., Levine, E., & McDaniel, M. (1989). Job applicant training and work experience evaluation in personnel selection. In K. Rowland & G. Ferris, (Eds.), *Research in personnel*

and human resource management. Greenwich, CT: JAI Press.

Barrick, M., & Mount, M. (1991). The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology 44*, 1–26.

Barrick, M., Mount, M., & Strauss J. (1993). Conscientiousness and performance of sales representatives: Test of the mediating effects of goal setting. *Journal of Applied Psychology, 78*, 715–722.

Bernadin, H., & Beatty, R. (1984). *Performance appraisal: Assessing human behavior at work.* Boston: West Publishing Co.

Borman, W. C. (1991). Job behavior, performance, and effectiveness. In M. Dunnette & L. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd ed., Vol. 2). Palo Alto, CA: Consulting Psychologists Press.

Bray, D. W., Campbell, R. J., & Grant, D. L. (1979). *Formative years in business.* Huntington, NY: Robert E. Krieger Publishing.

Brown, S. (1978). Long-term validity of a personal history item scoring procedure. *Journal of Applied Psychology, 63*, 673–676.

Campbell, J. (1990). An overview of the army selection and classification project (Project A). *Personnel Psychology 43*, 243–257.

Cascio, W., & Phillips, N. (1979). Performance testing: A rose among thorns? *Personnel Psychology, 32*, 751–766.

Cook, J. D., Hepworth, S. J., Wall, T. D., & Warr, P. B. (1981). *The experience of work.* New York: Academic Press.

Cornelius, E. T. (1983). The use of projective techniques in personnel selection. In K. Rowland & B. Ferris (Eds.), *Research in personnel and human resources management.* Greenwich, CT: JAI Press.

Day, D., & Silverman, S. (1989). Personality and job performance: Evidence of incremental validity. *Personnel Psychology, 42*, 25–36.

Dipboye, R. (1992). *Selection interviews: Process perspectives.* Cincinnati, OH: SouthWestern.

Dreher, G., & Sackett, P. (1983). Commentary. In G. Dreher, & P. Sackett (Eds.), *Perspectives on employee staffing and selection.* Homewood, IL: Richard D. Irwin.

Edward, K. (1990). The interplay of affect and cognition in attitude formation and change, *Journal of Personality and Social Psychology, 59*, 202–216.

England, G. (1971). *Development and use of weighted application blanks.* Minneapolis, MN: Industrial Relations Center, University of Minnesota.

Fleischman, E., & Mumford, M. (1988). Ability requirements scales. In S. Gael (Ed.), *Job analysis handbook for business, industry, and government, Vol. 2.* New York: Wiley.

Gatewood, R., & Feild, H. (1998). *Human resource selection, 4th ed.* Fort Worth, TX: Dryden Press.

Gaugler, B., Rosenthal, D., Thornton, G., & Bentson, C. (1987). Meta-analysis of assessment center validity. *Journal of Applied Psychology, 72*, 493–511.

Ghiselli, E. E. (1973). The validity of aptitude tests in personnel section. *Personnel Psychology, 26*, 461–477.

Gordon, M. F., & Kleiman, L. S. (1976). The prediction of trainability using a work sample test and an aptitude test: A direct comparison. *Personnel Psychology, 29*, 243–253.

Grimsley, G., & Jarrett, H. (1975). The relation of past managerial achievements to test measures obtained in the employment situation: Methodology and results—II. *Personnel Psychology, 28*, 215–231.

Guion, R. M., & Gottier, R. F. (1965). Validity of personality measures in personnel selection. *Personnel Psychology, 18*, 135–164.

Haetner, J. (1977). Race, age, sex, and competence as factors in employer selection of the disadvantaged. *Journal of Applied Psychology, 62*, 199–202.

Harris, M. (1989). Reconsidering the employment interview: A review of recent literature and suggestions for future research. *Personnel Psychology, 42*, 691–726.

Hogan, J. (1991). Physical abilities. In M. Dunnette & L. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd ed., Vol. 2). Palo Alto, CA: Consulting Psychologists Press.

Hogan, R. (1992). Personality and personality measurement, in M. Dunnette & L. Hough (Eds). *The handbook of industrial and organizational psychology (2nd ed., Vol. 2).* (pp. 873–919). Palo Alto, CA: Consulting Psychologists Press.

Hulin, C. (1991). Adaptation, persistence, and commitment in organizations. In M. Dunnette & L. Hough (Eds.), *Handbook of organizational and industrial psychology* (2nd ed., Vol. 2). Palo Alto, CA: Consulting Psychologists Press.

Hunter, J. (1986). Cognitive ability, cognitive attitudes, job knowledge, and job performance. *Journal of Vocational Behavior, 29*, 340–362.

Hunter, J., & Hunter, R. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96*, 72–88.

Hunter, J. E., Schmidt, F. L., & Hunter, R. (1979). Differential validity of employment test by race: A comprehensive review and analysis. *Psychological Bulletin, 85*, 721–735.

Janz, T. (1982). Initial comparisons of patterned behavior description interviews versus unstructured interviews. *Journal of Applied Psychology, 67*, 577–580.

Judge, T. (1993). Does affective disposition moderate the relationship between job satisfaction and voluntary turnover? *Journal of Applied Psychology, 78*, 395–401.

Keenan, A. (1977). Some relationships between interviewers' personal feeling about candidates and their general evaluation of them. *Journal of Occupational Psychology, 50*, 275–283.

Kinslinger, H. J. (1966). Application of projective techniques in personnel psychology since 1949. *Psychological Bulletin, 66*, 134–149.

Landy, F., & Pharr, J. (1983). *The measurement of work performance methods, theory, and applications*. New York: Academic Press.

Latham, G., Saari, L., Pursell, E., & Campion, M. (1980). The situational interview. *Journal of Applied Psychology, 65*, 422–427.

Latham, G., & Wexley, K. (1981). *Increasing productivity through performance appraisal*. Reading, MA: Addison-Wesley Publishing Co.

Ledvinka, J., & Scarpello, V. (1991). *Federal regulation of personnel and human resource management, second edition*. Boston: PWS-Kent Publishing Company.

Locke, E. A. (1976). The nature and causes of job satisfaction. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology*. Chicago: Rand McNally College Publishing Co.

Marrow, P. (1993). *The theory and measurement of work commitment*. Greenwich, CT: JAI Press Inc.

Mathieu, J. E., & Zajac, D. M. (1990). A review and Meta-analysis of the antecedents, correlates, and consequences of organizational commitment. *Psychological Bulletin, 108*, 171–194.

McCormick, E. J., & Jeanneret, P. R. (1989). Position Analysis Questionnaire (PAQ). In S. Gael (Ed.), *The job analysis handbook for business, industry, and government*. New York: Wiley.

McDaniel, M., Schmidt, F., & Hunter, J. (1988). A meta-analysis of the validity of methods for rating training and experience in personnel selection. *Personnel Psychology, 41*, 283–314.

McEvoy, G., & Beck-Dudley, C. (1991). Legally defensible performance appraisals: A review of federal appeals court cases. *Sixth Annual Conference of the Society for Industrial and Organizational Psychologists*.

Millar, M., & Millar, K. (1990). Attitude change as a function of attitude type and argument type. *Journal of Personality and Social Psychology, 59*, 217–228.

Miner, J. B. (1977). *Motivation to manage: A ten-year update on the "studies in management education" research*. Atlanta, GA: Organizational Measurement Systems Press.

Mount, M., Barrick, M., & Strauss, J. (1994). Validity of observer ratings of the big five personality factors. *Journal of Applied Psychology, 79*, 272–280.

Mumford, M., & Owens, W. (1987). Methodology review: Principles, procedures, and findings in the application of background data measures. *Applied Psychological Measurement, 11*, 2–6.

Murphy, K., & Cleveland, J. (1991). *Performance appraisal: An organizational perspective*. Boston: Allyn & Bacon.

Murray, H. (1943). *Thematic apperception test*. Cambridge: Harvard University Press.

Olson, J., & Zanna, M. (1993). Attitudes and attitude change. *Annual Review of Psychology, 44*, 117–154.

Organ, D. W., & Konovsky, M. (1989) Cognitive versus affective determinants of organizational citizenship behavior. *Journal of Applied Psychology, 74*, 157–164.

Organ, D. W., & Near, J. P. (1985). Cognition vs. Affect in measures of job satisfaction. *International Journal of Psychology, 20*, 241–253.

Pearlman, K., Schmidt, F., & Hunter, F. (1980). "Validity generalization results for tests used to predict job proficiency and training success in clerical occupations." *Journal of Applied Psychology, 65*, 373–406.

Porter, L., Steers, R. T., Mowday, R. T., & Boulian, P. V. (1974). Organizational commitment, job satisfaction and turnover among psychiatric technicians. *Journal of Applied Psychology, 59*, 603–609.

Reilly, R., & Chao, G. (1982). Validity and fairness of some alternative employee selection procedures. *Personnel Psychology, 35*, 1–62.

Reilly, R., & Warech, M. (1991). The validity and fairness of alternative to cognitive tests. In L. Wing (Ed.), *Employment testing and public policy*. Boston: Kluwer.

Robinson, D. (1981). Content-oriented personnel selection in a small business setting. *Personnel Psychology, 34*, 77–78.

Rowe, P. (1963). Individual differences in selection decisions. *Journal of Applied Psychology, 47*, 305–307.

Roznowski, M., & Hulin, C. (1991). The scientific merit of valid measures of general constructs with special reference to job satisfaction and job withdrawal. In C. J. Cranny, P. C. Smith, & E. F. Stone (Eds.), *Job satisfaction*. New York: Lexington Books.

Scarpello, V. G., & Vandenberg, R. (1992). Some issues to consider when surveying employees' opinions. In J. Jones, B. Steffy, & D. Bray (Eds.), *Applying psychology in business: The manager's handbook*. Lexington, MA: Lexington Books.

Schmidt, F., Greenthol, A., Hunter, J., Berner, J., & Seaton, F. (1977). Job sample vs paper-and-pencil trade and technical tests: Adverse impact and examiner attitudes. *Personnel Psychology, 30*, 187–197.

Schmidt, F., Gast-Rosenbery, I., & Hunter, J. (1980). "Validity generalization for computer programmers." *Journal of Applied Psychology, 65*, 643–661.

Schmidt, F., Hunter, F., Pearlman, K., & Shane, T. (1979). "Further tests of the Schmidt-Hunter Bayesian validity generalization procedure." *Personnel Psychology, 32*, 257–281.

Schmidt, F., Hunter, J., & Caplan, J. (1981). "Validity generalization results for two job groups in the petroleum industry." *Journal of Applied Psychology, 66*, 261–273.

Schmitt, N. (1976). Social and situational determinants of interview decisions: Implications for the employment interview. *Personnel Psychology, 29*, 79–101.

Schmitt, N., Gooding, R., Noe, R., & Kirsch, M. (1984). Metaanalysis of validity studies published between 1964 and 1982 and the investigation of study characteristics. *Personnel Psychology, 37*, 407–422.

Steers, R. M. (1977). *Organizational effectiveness: A behavioral view*. Santa Monica, CA: Goodyear.

Task Force on Assessment Center Guidelines (1989). Guidelines and Ethical Considerations for Assessment Center Operations. *Public Personnel Management, 18*, 457–470.

Ulrich, L., & Trumbo, D. (1965). The selection interview since 1949. *Psychological Bulletin, 63*, 100–116.

Valenzi, E., & Andrews, I. R. (1973). Individual differences in the decision process of employment interviewers. *Journal of Applied Psychology, 58*, 49–53.

Wright, P., Kacmar, M., McMahan, G., & Deleeuw, K. (1992). P=f(M x A): Cognitive ability as a moderator of the relationship between personality and job performance. *Academy of Management Best Paper Proceedings*. Madison, WI: Omnipress.

This Page Intentionally Left Blank

CHAPTER 21

PSYCHOLOGICAL ASSESSMENT OF ETHNIC MINORITIES

Antonio E. Puente
Miguel Perez Garcia

INTRODUCTION

Whereas the study of abnormal behavior through the use of scientifically based psychological instruments has a relatively lengthy and interesting history (see introductory chapter), the assessment of individuals who fall toward the edges of the "bell curve" poses unique social, political, and scientific challenges (Olmedo, 1981; Scarr, 1988). Traditionally, approaches to the study of individuals who are considered outside the mainstream of whatever society they belong to have been politically and socially based. The understanding of the study of these individuals was grounded on the assumptions that it is morally correct to understand these individuals (Fowlers & Richardson, 1996). Their "abnormal" functioning may be more saliently expressed by understanding their affiliation to a culture that is not appreciative or reflective of the majority group which rules or guides the social context in which they live. However, we propose that while such a motivation would appear reasonable and politically correct, it is still insufficient scientifically.

First, such an approach presupposes that the role of psychology is partially if not largely oriented toward righting the wrongs of a society's ancestors and, hence, primarily a social enterprise. While a reasonable goal, that would appear to us as insufficient. Second, one might assume that understanding others who are, by design, difficult to understand, is again a reasonable goal. While we believe that for individual cases and in clinical situations this is not only desirable but also ethically appropriate, again this paradigm is insufficient. A third goal, rarely addressed by workers in this field, is that we believe that the study of culture and psychopathology combined (especially from a cognitive or neuro-cognitive perspective) provides a much larger pool of data about the human condition than previously used paradigms.

An example of this approach is found in a study from the World Health Organization (1973). They reported that with regards to schizophrenia, in Nigeria 58 percent and in India 51 percent of hospitalized individuals experienced complete remissions after two years after treatment. In contrast, in Denmark only 6 percent remission had been reported. The question then becomes what aspects of Nigerian and Indian culture is present (that are not in Denmark) which allow for such a high rate of recovery. The Basic Behavioral Science Task Force of the National Advisory Mental Health Council (1996) reported that in Los Angeles, Mexican Americans indicated that schizophrenia was a transitory condition associated with nervousness whereas Anglo-Saxon counterparts believed that schizophrenia was a permanent and total deterio-

ration of mental functions. The question becomes how culture is defined and how it helps focus the more global issue of human function and dysfunction. Thus, we believe that basic principles about the human condition are the best understood by studying the interrelations of culture and psychopathology through the use of scientifically and cognitively (possibly neuro-cognitively) based psychological instruments. Such an approach could potentially yield unique insights into individual differences and general theories of psychological function and dysfunction.

However, individual differences and diversity are often viewed as impediments to the development of general principles of behavior. Of significance is the lack of understanding and sensitivity for larger group differences. Few would question the basic ability of specific psychological tests (e.g., Wechsler Memory Scale) to assist in the discrimination or classification of specific diagnostic groups (i.e., those with, from those without memory dysfunction). In contrast, few would disagree that affiliation with specific demographic groups (non-diagnostic) would be of great value in diagnostic classification. Presumably, this assumption is based on the concept that psychopathology (or for that matter, skills, abilities, or any other behavioral variable) is relatively free from the contamination of these potential confounds (Westermeyer, 1987a). Thus, this perspective suggests that other (non-diagnostic) group membership, while possibly important in some capacity, would have little or no effect on nomenclature issues. Such a belief is deeply rooted in nonempirical foundations, and its beginnings lie in a number of historical trends, none well documented or acknowledged. Thus, speculation rather than definitive analysis is the source for the following observations.

Traditionally, few attempts have been made to understand the behavior of individuals in minority groups. Brislin (1988) and others have cogently argued that psychologists for too long have categorically shown poor understanding of behavior traits and patterns of individuals who do not belong to groups associated with mainstream society. This limited perspective of the nature of behavior was first addressed by Frank Beach (1950) in his now classic article, "The Snark was a Boojum." In more contemporary terms, Robert Guthrie's (1976) book, *Even the Rat was White*, cites clear evidence not only of restricted sampling but of limited understanding of many other species (in the case of Beach) or other racial and ethnic groups (in the case of Guthrie).

One direct outcome of this situation, shown in recent statistics, suggests that few individuals appear interested in studying how understanding racial and ethnic group membership may contribute to understanding behavior. The article by the American Psychological Association's (APA) Committee for Human Resources, "The Changing Face of American Psychology" (Howard, Pion, Gottfredson, Flattau, Oskame, Pfafflin, Bray, & Burstein, 1986) underscores the paucity of minorities pursuing study and being associated with all areas of psychology. Of special concern is the limited number of minorities in graduate schools and in faculty positions. Hall (1997) has suggested that "cultural malpractice" exists across all aspects of psychological pedagogy, research, and clinical activities. Bernal and Castro (1994) indicated that only 12 percent of all clinical programs require courses involving cultural issues but 89 percent of the programs indicated that they "integrated" such issues into their program. Interestingly, they reported that approximately half of all clinical programs did not have an ethnic minority on the staff. To add insult to injury, less than 10 percent of clinical students are of color. These disturbing trends persist a decade after the historical report in 1978 by the President's Commission on Mental Health.

The lack of understanding combined with the lack of resources to solve the problem will clearly leads to further complications of an already complex issue. Nevertheless, the common denominator is limited understanding. This limited understanding of minority populations has resulted in overrepresentation of minority groups in several distinct psychopathology groups. Maheady, Towne, Algozzine, Mercer, and Ysseldyke (1983) and others have observed that members of minority or underrepresented groups tend to be overrepresented in special education programs, especially programs for the mildly handicapped. However, it is unclear that less "biased" tests will produce less overrepresentation. Thus, we believe that basic principles about the human condition is best understood by studying

The overrepresentation of minority groups in handicapped conditions has, in turn, resulted in negative stereotypes. In 1991 the Department of Education reported that African Americans com-

prise 21 percent of total enrollment but an astonishing 42 percent of individuals labeled educable mentally retardate, 38 percent of those in educable mentally retardate, and 22 percent of those considered learning-disabled. Hispanics comprise 13 percent of the total enrollment, but 10 percent of educable mentally retardated, 22 percent trainable mentally retardate, and 12 percent learning disabled. In contrast, for Asian students total enrollment is reflective of enrollment of special programs. Such stereotypes in the short term encourage the assignment of individuals to incorrect diagnostic groups (e.g., learning disabled). In the long term this stereotypical and grossly incorrect database may eventually serve as a foundation for potentially incorrect theories and research programs on racial and ethnic differences (e.g., Jensen, 1980). While all valid programs of inquiry should exist (Kuhn, 1970), constraints on the scientific process fueled by emotional and unempirical variables have little value for the discipline, for the science, for society, and most of all, for members of ethnic-minority groups. But as late as the end of the 20th century, we still are surprised and disappointed to read that "distinguished" historians of American culture continue to misunderstand the very essence of the issues at heart.

The purpose of this chapter will be to avoid such an orientation by focusing as much as possible on the data that are available. Initially, this contribution will focus on providing both historical and clinical background of testing of ethnic-minority group members. Standard clinical and psychometric practices involving individuals of minority groups will be presented and critiqued. Suggestions for theoretical shifts as well as practical clinical and psychometric approaches will be outlined, with cognizance of the potential pitfalls, perceived or real, that presently exist.

This chapter is primarily intended for North American audiences. Numerous limitations in the available data set, whether clinical or otherwise, would make a more geographically ambitious approach impossible. In fact, it could be argued that the most fertile research database is found in the states. Nevertheless, the approach (though not necessarily the data) should be considered a model for workers in other cultures, groups, or locations (e.g., Native Indians in mainstream Brazilian culture) in order to address the issues of psychological assessment of ethnic-minority group members.

An initial step in understanding members of minority groups is to define such groups. According to accepted practice, individuals are different from larger groups if they are not members of that group. Group composition can be determined by social, legal, biological, statistical, and behavioral variables. Possibly the easiest and most socially acceptable variable is biological, such as color of skin. However, other variables may also play a role. Statistical methods define group memberships by numerical scores obtained, while social and legal approaches may use societal tradition to define membership. Behavioral and psychological variables represent the most robust method as they should be free of bias due to the use of empirical methods and the criterion in question, the function of the person. After all, the color of an individual's skin is much less critical than their thinking patterns when it comes to understanding such issues as capacity to learn.

Standard practices have used overt and obvious variables to classify members into minority groups. For example, if an individual is not white (Caucasian) in North America he or she must belong to a minority group. One need look no further than the disciplines of animal behavior and neuropsychology to realize that gross morphological signs are often not well correlated with clear behavioral patterns. For this chapter, Brislin's (1988) classification system for human diversity is adopted. Contrary to popular belief, only three races exist. These include Caucasian (e.g., white), black, and Indian. The Indian race can be subdivided into Native American (e.g., Cherokee, Incas, etc.) and Asian (e.g., Japanese, Chinese, etc.). Ethnicity is another variable that can be used to differentiate mainstream from minority groups. Here, ethnicity is defined as a collective identity (e.g., Jew, Italian, etc.). Next, group composition can be determined by culture (e.g., southern, urban, etc.). This variable implies that groups can be defined according to social and personal identification. While less understood and accepted, other variables could also assist in determining group membership. These include, but should not be limited to, gender, sex, physical status (e.g., disability), social class, and religion. In 1990 the United States Bureau of the Census has more or less compressed these distinctions avoiding the differences between race, culture, and ethnicity. In a bold step, they proposed five different groups; Spanish/Hispanic/

Table 21.1. Ethnicity and Race According to the 1990 U.S. Census Data: Origin, Total Number, and Subgroups

Spanish/Hispanic/Latin Background or Origin
Origin = Latin America or Spain; Total = 22,354,059
Cuba
Mexican/Mexican-American/Chicano
Puerto Rican
Hispanic Latin America (e.g., Panamanian, Peruvian, Venezuelan, Ecuadorian, Guatemalan, etc.)
Spaniard
African American/Black/Negro
Origin = African or Caribbean; Total = 29,986,060
Asian or Pacific Islander
Origin = Far East, Southeast Asia, Indian Subcontinent, or Pacific Islands; Total = 7,273,662
Asian Indian
Chinese
Japanese
Korean
Vietnamese
Filipino
Hawaiian
Indian (American) or Alaska Native
Origin = North America; Total = 1,959,234
Aleut
American Indian
Eskimo
White
Origin = Europe, North America, Middle East; Total = 199,686,070

Latin Background or Origin, African American/Black/Negro, Asian (American) or Alaska Native, and White. The Spanish group contains five different groups whereas the Asian has seven separate subcategories and the Indians have three.

In the area of psychological assessment, race has been the most widely studied of the previous variables. Sex and, to a lesser degree, ethnicity have been considered as potential though not highly salient variables. However, culture, physical status, social class, and religion have rarely been considered important in understanding human behavior. Whether this neglect is due to collective wisdom or ignorance is not known (nor is it the focus of this chapter).

Regardless of the variable used, ethnic-minority group membership will be defined as indicated previously by groups who are both politically powerless and sparsely represented in scientific inquiry. However, what may be a minority group in terms of ethnicity in 1990 may not be by the year 2000. Census figures suggest, for example, that people of color (including African Americans, Asian Americans, Hispanics, and Native Americans) who now constitute less than 20 percent of the U.S. populations will soon constitute approximately 50 percent of the American population (Basic Behavioral Science Task Force of the National Advisory Mental Health Council, 1996).

A necessary outcome of appropriately defining group membership is the implication that a minority member will engage in behavior that is different from the mainstream norm but not necessarily abnormal. Thus, clearer understanding of human behavior is the goal. Such an understanding is not only academically useful but also contains treatment implications. The importance of ethnic-minority group membership for psychological treatment has been outlined by Sue and Zane (1987), while Lawson (1987) has reported its implications for psychopharmacological intervention. Caution should be inserted here. Careful between-group comparison often implies limited concern for within-group analysis. Using the Hispanic population in the United States as an example, the behavioral patterns of Cubans,

Mexicans, and Puerto Ricans may actually differ more from each other than the entire group of Hispanics differs from Caucasians. In an ongoing translation and standardization of the Wechsler Intelligence Scale for children, Hispanics have been further subdivided into Central Americans, Cubans, Mexicans, Puerto Ricans, and South Americans. Thus, within-minority group analysis will eventually become as important as minority versus majority group comparisons.

HISTORICAL PRECEDENTS

Galton's "Inquiries into Human Faculty and Its Development" written in 1883 is most often considered the beginnings of psychological assessment (Boring, 1950). In order to evaluate human disabilities (and not sins, as had commonly been the case prior to Galton), this British pioneer developed the "mental test." While the test intended to measure such variables as color discrimination and auditory reaction time, the purpose of establishing the Anthropometric Laboratory at the International Health Exhibition in London was to determine the range of human abilities. Together with the founding of the journal *Biometrika* and the Eugenics Laboratory, Galton attempted to develop the concept of racial improvement (Schultz & Schultz, 1996).

The discrimination of acceptable and nonacceptable human characteristics has, unfortunately, found its way into present-day mental testing, possibly by way of James McKeen Cattell. After obtaining his Ph.D. from Wundt in Leipzig, Germany, Cattell came into contact with Galton (Boring, 1950), who in turn had enormous influence both directly (e.g., with numerous students) and indirectly (e.g., as editor of *Science*) on the study of mental ability in the United States. However, it was not until the appearance of Henry H. Goddard at Vineland Training School in New Jersey, and later Lewis Terman at Stanford University that a research program of psychological abilities became part of mainstream psychology.

Using "the evidence of mental tests," Terman (1916) indicated that "the average intelligence of women and girls is as high as that of men and boys" (p. 68). Nevertheless, he concluded in his book, *The Measurement of Intelligence*, that the "dullness" seen in "Indians, Mexicans, and Negroes raises the question of racial differences in mental tasks." Terman suggested, "Children of the group should be segregated in special classes and given instruction which is concrete and practical. They cannot master abstraction, but they can often be made efficient workers, able to look out for themselves" (p. 92). He continued, "There is no possibility at present of convincing society that they should not be allowed to reproduce, although from a eugenics point of view they constitute a grave problem because of their unusually prolific breeding" (p. 92).

Such an orientation is observed in Goddard's work and later in Robert Yerkes's groundbreaking work with the Army Alpha and Beta tests during World War I. These tests were meant to classify A (intelligent) and D and E (feebleminded) individuals with a mean mental age of 13.08. (This score may have prompted Goddard to term any adult with less than 13 years of mental age as "moron."). However, both immigrants and nonwhites tended to score lower, prompting Yerkes (1923) to write in *Atlantic Monthly* about noninherited racial differences. This conclusion readily supported the racist opinion of Madison Grant who considered Nordics superior to other races. Based on these observations, Yerkes and others encouraged strict immigration laws especially for "the negro." To curtail the reproduction of those already in the United States, several American followers of Galton (namely John H. Noyer and Victoria Woodhull) established a center for American eugenics in Cold Spring Harbor with financial support from the Carnegie Institution (Leahey, 1997). One of the greatest proponents of eugenics, Henry Goddard, published his famous book *The Kallikak Family: A Study in the Heredity of Feeblemindedness* (1912). This book, probably more than any published work of the time, was used for the control of reproduction by ethnic minorities.

Reflecting the influence of this and similar works, sterilization and vasectomy became common phenomena. According to Leahy, one of the greatest landmark decisions on the issue was that of a mental patient, Carrie Buck. After giving birth to a retarded child out of wedlock, the "feebleminded" Buck was involuntarily sterilized. She, in turn, sued the state of Virginia but lost in a split decision at the Supreme Court level. It seems as though unempirical (and presently considered unethical) approaches to the measurement of abilities are never easily resolved scientifically. Earlier in this century this issue

was far from being resolved academically. Approximately 50 years later Stephen Jay Gould (1981) continues to argue that such strong conclusions have indeed been based on weak data.

Unfortunately, this approach to the understanding of minority behavior, at best weak and spurious, was the foundation for the Jewish genocide by the Nazis. However, not until 1954 did the judicial branch in this country make strides to erase this previously accepted and now embarrassing "scientific" orientation. The Brown case in 1954 allowed for desegregating of races in the school system. However, cases specifically referring to minorities and testing did not surface until the 1970s (Reschly, 1984). Generally, the plaintiff in these cases represented the three major minority groups of the time—African American, Hispanic, and Native American—who had been poorly and unethically classified as retarded.

While most of the cases were won or favorably settled out of court, it was not until the legislative aspect of the litigation-legislation cycle occurred (Bersoff, 1981) that reform began to be developed and later implemented. According to Reschly (1984), the federal Education for All Handicapped Children Act of 1975 "was the most important and most widely applicable legislative act." This act opened the road for later litigation meant to define more succinctly the spirit of this law.

Perhaps linked to these legal efforts, psychologists have become increasingly aware of the need to document human abilities more carefully. In his introduction to the special issue, "Cultural Factors in Understanding and Assessing Psychopathology" (*Journal of Consulting and Clinical Psychology*), James Butcher (1987) stated that "the application of psychological procedure and methods with patients from different cultural backgrounds raises numerous methodological issues. "Issues such as psychological equivalence, test reliability and validity, and test utility were some of the factors that Butcher considered critical. Five years later, Bethancourt and Lopez (1993) still believe that the study of ethnic minorities still hold "at best second place" in mainstream psychology. This has occurred despite critical reports suggesting that such an approach would be detrimental not only to ethnic-minorities but to psychology at large (McGovern, Furomoto, Halpern, Kimble, & McKeachie, 1991).

In this decade undoubtedly the most significant and most detrimental work on this issue comes from Herrnstein and Murray (1994) in their highly controversial book, *The Bell Curve.* Herrnstein and Murray aggressively pursue the traditional concepts that ethnic minorities do not score well on standardized tests, including tests of achievement and intelligence, because of genetic and biological limitations. In many respects these authors provide a modern-day version of the ideas of Terman, Goddard, and others (e.g., Graham) linked to immigration laws, reproductive limitations, and intellectual and social segregation.

Sternberg (1997) has cogently addressed the importance of ultimate criterion, possibly lifelong learning capacity, rather than test intelligence. Test intelligence may be a significant though, by design, incomplete condition for understanding learning. Such intelligence appears, according to Stenrberg, to be predictive of scholastic achievement. However, this type of achievement is only partially correlated with lifelong success. Two other factors are suggested by him that are obviously not understood by Herrnstein and Murray; behavioral intelligence and intelligence. It is the concept of behavioral intelligence that we find particularly interesting. One could clearly argue that a migrant worker whose native language is Spanish would do particularly poorly on the SAT or, for that matter, the GRE. Thus, there would be little question that such a person would not gain admission to most selective colleges in the United States. Further, one could argue that such an individual would undeniably do poorly in a traditional university curriculum. But to argue that such a person is biological or genetically inferior seems downright stupid. For example, such a migrant worker has found a method to travel from rural Central America to, say the eastern United States, with little money, inadequate transportation, and limited understanding of the culture. They are able to find work, complete the task, live frugally, send money back to their homeland, and locate alternative employment within days of completion of the job at hand, often in another state. One could argue that a suburban-raised Anglo-Saxon who has played on the high school sports team and has dedicated his or her life to spectator sports and socially driven concerns could not under any circumstances go to Central America and replicate what their counterparts have

accomplished in the United States. The question becomes, are we measuring "true" intelligence or some understanding of culture.

Of course, to the typical reader of this chapter, such an example seems rather extreme. Hence, we have chosen to provide another example, which should be closer to the experiences of most psychologists. In another and more recent article Sternberg and Williams (1997) suggest that the GREs, still the most widely used standardized measure of achievement for acceptance into graduate school, predicts little in terms of graduate school performance and maybe less than that in terms of career success. Additional and also sophisticated arguments against the limited arguments of *The Bell Curve* are also found in Gould (1996). Thus, the criterion for intelligence and achievement and, for that matter pathology, according to Sternberg (and suggested by Gould) as well as accepted here, is not test scoring (especially alone and out of context) but life-long ability to adapt to the demands of life. This chapter attempts to build on this newfound scientific interest in an effort to determine the needs, limitations, and directions associated with the psychological assessment of ethnic-minority populations in North America.

THEORETICAL ISSUES

In a chapter of this type it would be essentially impossible to address all pertinent theoretical issues that apply to the psychological assessment of ethnic minorities. We have chosen to focus on three main issues, bias, acculturation, and culture believing that they are the three most critical issues involved in this area of study.

Bias

Kenneth Eells pioneered the concept of bias in mental measurement, specifically the mental test. While his work focused only on whites, it did address the importance of difference—in this case, social class—in the assessment of mental function (Eells, 1951). Although the reasons for doing so are not entirely clear, some workers in psychometrics generalized his findings to other populations, namely African Americans. This incorrect generalization launched a wave of poorly developed and executed studies on bias in testing.

One of the most controversial figures in mental bias research is Jensen, of the University of California at Berkeley; his most controversial book is *Bias in Mental Testing* (1980). According to Jensen, mental testing has been criticized because of one or more of the following reasons:

1. cultural bias
2. specific test items
3. inability to define or measure intelligence
4. tests that measure too narrow a range of abilities
5. failure to measure innate capacity
6. IQ tests that measure only learned skills
7. IQs that are inconsistent
8. test scores that are contaminated by extraneous factors
9. misuses, abuses, and undesirable consequences of testing

According to Jensen (1980), these criticisms are largely unfounded and confused with other factors. As he wrote, "Anxiety about one's own status, or the importance of the traits measured by tests, or sympathy for the less fortunate, may prompt the acceptance of criticisms of tests without evidence" (p. 23). Unfortunately, such critiques tend to focus on IQ tests and are emotionally interpreted. They complicate the question and prevent adequate understanding of the valid issues.

Reynolds and Brown (1984) presented a set of reasons, which are applicable to bias for a wider range of tests. These include:

1. inappropriate content
2. inappropriate standardization samples
3. examiner and language bias
4. inequitable social consequences
5. measurement of different constructs
6. differential predictive validity

Regardless of the source of bias, the definition of bias must also be considered. Unfortunately, numerous definitions are available in the literature—some more heuristic and plausible than others. The following are two samples of the many available.

> [Bias results from] differences in the extent to which the child being tested has had the opportunity to know and become familiar with the specific subject matter or specific process required by the test item. (Eells, 1951, p. 54)

> Psychometric bias is a set of statistical attributes conjointly of a given test and two or more specified subpopulations (Jensen, 1980, p. 375)

Flaugher (1978) has suggested that test bias can mean more than simple knowledge or psychometric deficiencies. Indeed, bias could be represented in a wide variety of concerns including, but not limited to, both psychometric issues (mean differences, differential validity, item content, internal validity) and test usage (over-interpretation, selection model, and atmosphere). He concluded that in 1978 the research was promising, but the results were still disappointing. Twenty years later, the research and the results are both disappointing.

Among the more current research findings, an excellent example is Drasgow's (1972) article, "Biased Test Items and Differential Validity." In this review the author addresses differences between majority and minority groups in validity coefficients. The results of his study provide support for earlier findings suggesting that validity coefficients may not prove useful in examining test bias. He concludes: "Test scores *can* be used to predict criterion performance for minority group members. Nevertheless, it *may be inappropriate to compare test scores for minority group members with test scores for majority group members*" (p. 529, italics added). In a similar vein, Cole (1981) concluded in her article, "Bias in Testing," that "there is no large-scale, consistent bias against minority groups." Nevertheless, both "subtle aspects of the testing situation" and presumably more refined understanding still evade workers in the field. In contrast, Humphries (1986) has argued that even if items differ between groups, these items should not be labeled as biased if adequate measurement properties are taken into account.

Recently, the American Psychological Association released the results of a task-force study on these issues (Neisser, Boodoo, Bouchard, Boykin, Bordy, Ceci, Halpern, Loehlin, Perloff, Sternberg, & Urbina, 1996). Presumably the focus of this task force was to address authoritatively the issues brought out by *The Bell Curve*. Whereas the article addressed numerous critical issues, it fell far short when addressing ethnic-minority issues. The task force assumes, without question, that if tests are to be used as predictors of future performance, these tests do not seem to be biased against African Americans" (p. 93). However, less clear evidence of bias is presented against other ethnic-minority groups. Unfortunately, one is left with the sense that outside Asian Americans, other ethnic minorities score below their Anglo-Saxon counterparts. Further, they suggest that numerous factors, but not necessarily bias, ranging from economics to genetics, many be playing a role in these differences. Despite these opinions, may questions have yet to be formulated and, of course, answered. Until then, as Reynolds and Brown have concluded, the verdict on test bias is still not in (scientifically).

However, Susuki and Valencia (1997) have suggested that a significant drop in bias research exists. They believe in this precipitous drop is due to the fact that belief that bias does not exist. This erroneous belief is due to the following reasons:

1. Some minority groups have not been studied extensively and, in some cases, not at all.
2. Test bias in school-based tests has been done only with some but not a large variety of tests.
3. In actuality, there are mixed results. They report several studies by Valencia and colleagues which suggest that the K-ABC contain bias in predictive validity and content but not in construct validity or reliability.
4. Bias research has traditionally been done with nonpatient populations.

Malgady (1996) proposes an interesting twist to the theoretical foundations of bias research. He proposes that what is necessary is to reverse the null hypothesis. That is, we must assume that bias exists. If one commits an error in measurement, then it is better to assume that bias exists so that precautions are taken to protect the individual.

Acculturation

If a minority group does poorly on a test, relative to a majority group, two interpretations may be used to account for the discrepancy. A rather emotional one is provided by Jensen (1980)—

that the difference is accounted for by biological factors such as genetics. A less popular interpretation used by researchers studying integration of an immigrant group into a majority or mainstream culture is that of acculturation.

Assimilation into a larger, more mainstream culture allows an individual to understand and adjust to the cultural, social, and psychological requirements of that culture. Conversely, those who do not adapt are considered to exhibit greater degrees of psychopathology. An illustration of the lack of adaptation was reported by Hoffmann, Dana, and Bolton (1985) who found that Sioux Native Americans with strong ties to tribal values and language were more likely to exhibit psychopathology as measured by the MMPI. These findings have also been replicated with other minority groups, including Hispanics (e.g., Montgomery & Oroz, 1984). Focusing on cognitive style and intelligence, Gonzales and Roll (1985) reported differences between Mexican Americans and whites on several test measures. However, no group difference was observed between Anglo-Americans and a subgroup of the original sample of Mexican Americans who had been shown to be acculturated to Anglo-American culture.

One method to determine whether acculturation has been achieved, and thus controlled, is to administer an acculturation scale. Marin, Sabogal, Marin, and Otero-Sabogal (1987) have developed a 12-item scale, which measures acculturation in Hispanic populations. The validation criteria included generation, length of residence in the United States, age at arrival, ethnic self-identification, and an acculturation index. These findings have been extended to children (e.g., Franco, 1983) as well as to other cultural groups such as Asian Americans (Suinn, Rickard-Figueroa, Lew, & Vigil, 1987). Preliminary findings suggest that age (younger), sex (male), and length of exposure to the predominant culture (Bumam, Telles, Kamo, & Hough, 1987) as well as cultural awareness and ethnic loyalty (Padilla, 1985) are critical factors in the acculturation process. Another scale used for acculturation is the Acculturation Rating Scale for Mexican Americans (ARSMA and ARSMA-II) (Cuellar, Arnold & Maldonado, 1995; Cuellar, Harris & Jasso, 1980). While this is a promising scale, more research is necessary to generalize use to other Hispanic and ethnic-minority populations. Fradd and Hallman (1983) concluded that until an individual has been taught strategies to build bridges from a previous to a current domain of knowledge, the validity of test measures is questionable.

The process of acculturation must be understood, however, as a dynamic rather than static process. Acculturation does not imply reaching an imaginary threshold at which time one becomes clearly acculturated. Knight and Kagan (1977) reported that it took about three generations for Mexican-American children to develop modal responses on Anglo-Saxon children with regard to social motives.

Four stages have been postulated (Basic Behavioral Science Task Force of the National Advisory Mental Health Council, 1996). They are; assimilation (becoming part of the majority culture), acculturation (adapting to the majority culture), alternation or biculturalism (adequately engaging two cultures), and multiculturalism (holding on to a personal and non-majority identity while participating in a goal-directed activity of the majority culture).

Further, acculturation is not dichotomous, instead it is multifaceted (Phinney, 1996; Magana, de La Rocha, Amsel, Magana, Fernandez, & Rulnick, 1996). Triandis (1982) has suggested that culture could be physical (e.g., buildings, tools, etc.) or subjective (e.g., social norms, roles, beliefs, and values). The subjective could include family dynamics, religious beliefs, language limitations, individualism, and so forth. Thus, one could conceivably be adapted to a culture physically, live and appear to be American (i.e., live in North Carolina, dress in Brooks Brothers clothing, etc.) but have specific behavioral patterns that would clearly identify the person as non-North Carolinian (i.e., native language would be Spanish, have extended family, practice Catholicism, and so forth).

Berry (1990) has proposed an interesting theory of acculturation. The process involves three levels:

1. Antecedents—internal, external, and traditional.
2. Processes—cultural change, acculturation, psychological acculturation,
3. Consequents—changed cultural and social system, changed psychological status of persons.

Such a comprehensive approach appreciates the multidimensional and phasic nature of acculturation. Bethancourt and Lopez (1993) have also suggested that lack of acculturation implies perceived stress, presumably because of the individual's inability to "fit" with the majority culture. The problem is in determining why that "stress" exists by defining the specific cultural values (e.g., lack of understanding of social norms) that produce the lack of acculturation and eventually the perceived stress. And even though not suggested by the authors, but certainly implied, is that perceived stress must be controlled in order to make certain that what is measured is acculturation and not stress secondary to limited acculturation.

While it is clear that acculturation is seen as the ability of immigrants (e.g., Hispanics and Asians) to adapt to a majority group culture (e.g., the United States), the possibility is considered that an analogous concept could be applied to nonimmigrant ethnic minorities already living in North America. If, as it is argued in the next section, intelligence is largely, if not completely, a cultural phenomenon, then acculturating to the majority culture is a prerequisite for the development of successful learning strategies and eventual intelligence. Thus, it is argued that African Americans living in the United States but not participating fully in the American culture may not be acculturated. Hence, the same issues would apply to this ethnic-minority group as it would to Hispanics and Asians.

Culture and Ethnicity

According to recent position papers, neither bias nor acculturation may be the most salient variables that need to be addressed in understanding ethnic minorities. Bethancourt and Lopez (1993) suggested that most studies to date on the issues at hand have been at best descriptive of the differences between cultures. Whereas such approaches appear on the surface useful, they suggest that they are at best an initial step in the more important question of culture. To date, they argue, the field does still not understand the role of culture in behavior and cognition. For example, they believe that using race as a variable, either dependent or independent, is inappropriate. Zuckerman (1990) has reported that within-race differences appear larger than between-race differences with regard to biological variables. The hypothesis is made that similar assumptions can be generated with regard to behavioral and cognitive variables as well. Indeed, social class might be a more salient variable in grouping individuals than race. In other words, individuals from a high social class would more likely be different than persons from a low social class than a white and black from similar social classes. First, they suggest the bottom up approach which starts with the data generated from cross-cultural studies in order to then alter existing theories about human behavior. They suggest the work of Triandis and colleagues as a benchmark for this approach. Second, and a more novel approach, is what they term the top down approach. The essence of this approach is to determine how culture helps define the larger concept of human behavior and cognition. In other words, culture is seen, much like psychopathology in the earlier ideas of Neal Miller, as a unique way of being able to understand "normal" human function. Instead of examining how cultures are different, the focus would shift to determining what are the most salient variables in helping distinguish individuals. Using Hispanics as an example, one would ask not how Hispanics are different than say Anglo-Saxons and what variables contribute to those differences. In this ethnic-minority group the following variables might be pertinent; language, social orientation, family dynamics, and religion. Then, the question would become how, for example, does language affect a person's behavior to the point of excluding them from a majority group essentially making them pathological. A third-tier question then would be what should be pathological. Essentially, the final question would be what is the criterion for labeling pathology.

Phinney (1996) has furthered this concept by proposing that ethnicity, a subset of culture, could be studied at three levels. First, she suggests cultural values, attitudes, and behaviors that help define a group. Second, group identification is partially based on a subjective sense of what it means to belong to the group. Finally, she suggests that group identification is based partially on the specific experiences associated with that group identification.

Finally, an alternative to doing cross-cultural investigations is to begin by understanding culture. According to Greenfield (1997), "ability assessments don't cross cultures". Specifically,

she suggests that values and meanings, knowledge, models of knowing, and conventions of communication are not easily translatable across cultures and could be culture-specific. That is, the criterion of a particular meaning must be understood before it is "translated". She concludes that "tests are not universal instruments".

Social Policy

Whether tests are biased or culturally free, whether an individual belongs to a minority or a majority group, whether different groups are biologically equal or unequal, group differences still exist. To deny the obvious would be inappropriate. Certain minority groups perform differently, more often than not worse, than majority groups on specific items, tasks, or tests. These differences drive social policy. Academic psychology would undoubtedly prefer to research these problems and discrepancies more thoroughly before allowing findings to affect the judicial and legislative process, because the data for any of these questions are at best inconclusive as well as emotional and at worst confusing.

However, policy must be and will be made in the absence of adequate data and in the presence of emotion (see Bersoff, 1981). This reality could explain why Cole (1981) concluded that test bias research is likely to have only a small impact on complex social policy issues. Regardless, there are issues that relate to the available data. In the first edition of this handbook, Reschly (1984) addresses the concept of fairness. According to him, two approaches have been adopted. Equal treatment implies no bias or documentation in selection procedures and that all candidates, regardless of demographic affiliation, are treated equal. An alternative to this approach is equal outcomes, which implies that selection should match population demographics. Regardless of the approach and the data, North American society has adopted in principle the concept of fairness. The question remaining is which method described by Reschly will be chosen and what, if any, implications will the current paucity of data and lack of scientific agreement have on social policy formation and implementation. Another and politically limited approach would be to assume that representation of the American population (or for that matter whatever criterion population was chosen) is a criterion of choice. Next, one could use the currently used measures described in this chapter within the context of subsamples. Specifically, if selection to a college is the goal, then a college would first choose to accept representation from all groups as desirable. Next, they would apply the traditional standards (e.g., standardized test scores) within the accepted or chosen subsamples (e.g., Caucasian, Hispanics, African Americans, etc.). Considering the predictive validity of these tests, it is hypothesized that within-group (or subsample) variance would be greater than between-group (i.e., high-scoring blacks and high-scoring whites) variance.

To assist policymakers, researchers need to place greater importance on studying issues of race, culture, ethnicity, and related variables. The findings must then be applied to broaden our limited understanding of differences in psychological test performance of minority group members. Hall (1997) suggests the following steps in attempting to reach these objectives; (1) Ensure that the psychology curriculum is culturally diverse, (2) recruit and retain diverse faculty, (3) recruit and retain diverse students, (4) monitor, for the sake of accountability, efficacy of initiatives, (4) encourage culturally diverse research and publications, (5) increase the number of editors and reviewers of diverse background, (6) ensure minimum cultural competency for psychology students, (7) understand state-of-the-art research on topics of diversity, and (8) increase diversity within membership and leadership of the American Psychological Association.

Of course, there is the issue of who is to pursue these questions, both in academic and research settings. In the seminal article, "The Changing Face of American Psychology" (Howard et al., 1986), the future for ethnic-minority group representation is presented as quite dismal. While women have made significant strides, African Americans, Hispanics, Asian-Americans, and Native Americans continue to lose ground, in terms of representation in graduate school ranks (Hall, 1997). Similar trends exist in academic ranks, and presumably in clinical settings as well (Bernal & Castro, 1994). Programs within the American Psychological Association, including the Minority Fellowship Program and the Minority Neuroscience Fellowship Program, may aid talented minorities to pursue graduate training. Unfortunately, undergraduate majors in psychology mirror the same trend (Puente, 1993). Indeed, by the time minori-

Table 21.2. Several Structured Interviews Applicable to Ethnic-Minority Populations

INTERVIEW	REFERENCES
Brief Psychiatric Rating Scale	Overall & Gorham (1962)
Inpatient Multidimensional Rating	Lorr & Lett (1969)
Mental Status Schedule	Spitzer, Endicott, & Flenn (1967)
Present State Examination	Wing (1970)
Structural Clinical Interview (DSM III)	Spitzer & William (1983)
International Classification of Disease	World Health
Organization (199?)	
Interview (for ICD-10)	
Mini-Method Station Examination	Folstein, Folstein & McHugh (1975)

ties have chosen a college, they most likely have committed to a course of study. Simply put, despite the urgency of the questions raised, the future for a better understanding of psychological assessment of ethnic-minority group members looks bleaker than its past especially when issues such as "pipeline" of prospective students is considered. Whereas one would hope that the natural forces or evolution of psychology would "take care of the problem", social policy initiatives may have to jump-start what psychology has verbalized yet never realized.

ASSESSMENT METHODS

This section of the chapter will focus on specific assessment methods, including interviews, standard measures, culturally sensitive methods, and behavioral assessment methods. As feasible, each section will cover a variety but, not an exhaustive set, of tests or assessment strategies including application (and/or translation), norms, limitations and cautions, and suggestions for use.

Interview

The interview, whether structured or unstructured, remains the initial step of any psychological assessment and also the most commonly used method for obtaining information. The interview is a frequently used method for obtaining data in cross-cultural contexts. As Zubin (1965) and others have pointed out, however, the unstructured interview poses problems since it may yield unreliable data resulting from a host of uncontrolled factors.

Structured interviews may help in avoiding these pitfalls. Numerous interview methods, including several presented in this volume, seem generally well suited for use with minority populations, especially since they are often based on objective diagnostic criteria (e.g., Research Diagnostic Criteria). Several of these methods are found in Table 21.2.

Although most of these structured interviews have been well studied and validated, validity studies often use the judgment of the clinician as the criterion variable (Spitzer & Williams, 1980). Further, it is well accepted that cultural and ethnic variables—such as behavior patterns, nonverbal cues, translation equivalence, concept equivalence, gender differences, and general cultural beliefs—are often misunderstood by even the most sensitive clinician (Hall, 1997; Spitzer, Endicott, & Fleiss, 1967; Westermeyer, 1987a). Recent research has also revealed that expression of psychological symptoms is differentially affected by culture. Interviews and diagnostic conclusions are based on signs and symptoms which could be considered normal (versus abnormal) in specific cultures (e.g., belief in the devil, describing somatic abnormalities using metaphors, etc.). The end result is confounding symptoms with culture (Basic Behavioral Science Task Force of National Advisory Mental Health Council, 1996).

One way to avoid this complication is to use interview methods that either have been formally validated or are in current use with these populations. For example, the Present State Examination was an interview used for the international pilot study of psychopathology (World Health Organization, 1973). Another method is that of using a translator or someone knowledgeable about ethnic-minority groups. However, even this approach has limitations. It is not

unusual for the translator to be a lay person with limited understanding of psychological principles as well as an individual with personal interest in the patient. Further, translators may not approximate a balanced bilingual, or worse yet, not understand the culture in question. Distortion or misconception further impairs data gathering, especially with severely disorganized patients or individuals whose culture is very different from that of the diagnostician. Velazquez, Gonzales, Butcher, Castillo-Canez, Apocada, and Chavira (1997) suggested that an important and often over-looked first step in an evaluation is to allow the patient to choose the language to be used in the evaluation.

Several steps might be taken to attempt to control interview distortion. First, in order to bridge the language and cultural gap between patient and psychologist, rapport should be established prior to the interview. Greenfield (1997) has reported that ease in speaking to strangers, even though they are professionals, varies across cultures. For collective cultures (e.g., Asian and Hispanic), it is typical to limit discussions to only known individuals (Kim & Choi, 1994). In other words, you only relate one's problems with intimate or close friends or family friends. In contrast, in North America providing personal information to a stranger, but presumably a professional, increases the perception of objectivity and effectiveness.

Westermeyer (1987b) suggested that diagnostic interviews may take up to twice the usual time of a standard interview. Also, the clinician should make sure that ambiguous (whether real or imagined) questions or answers are clarified. Confrontation, the hallmark of some structured interview methods, should be avoided if possible since it may adversely affect client-clinician rapport.

By far the most important aspect of any diagnostic interview is to place the client in his or her own bio-psychosocial context and not the psychologist's context. Otherwise, a patient's behavior could be incorrectly interpreted as maladaptive (Adebimpe, 1981). To avoid erroneous conclusions, the psychologist must put special emphasis on understanding the patient's culture, race, ethnicity, class, or social context that grants him or her membership in a minority group. Not only must that context be understood but it should be understood as it relates to the patient's relationship to majority culture (e.g., Mexican migrant worker employed as a field hand in Colorado). Finally, and possibly most important, the clinician must understand his or her own limitations in other sociocultural situations. To enhance his or her understanding of others, the psychologist must become aware of, and possibly experience, other cultures and ethnic behavior patterns and cognitions. Hall (1997) has suggested that all clinicians must be well versed in these issues, initially in completing a multicultural graduate course and subsequently in clinical training.

Intellectual

Tests which attempt to measure the construct of intelligence are not only the most commonly used psychological tests (see related chapters in this volume) but also the most criticized (Neisser at al. 1996, Sternberg, 1997). The literature is replete with controversies about the efficacy of the construct of intelligence and its measurability (Gould, 1996; Helms, 1992), and strong and often emotional arguments have been levied against tests of intelligence by members of ethnic-minority groups (Herrnstein & Murray, 1994). Before these arguments are considered, the most commonly used tests of intelligence will be reviewed relative to their applicability to minority populations.

The application of intelligence tests to children of minority populations has yielded the most empirical data as well as the most controversy. Of the tests applicable to children, the Wechsler Intellectual Scale for Children (WISC) is one of the most popular psychometric tests of intelligence (Puente & Salazar, 1998). Despite that the WISC-III was published in 1991 (Wechsler, 1991), most data on this topic exists with the first two versions of this test. Excluding Asian children, the results, in general suggest up to one standard deviation difference between ethinic-minority groups and the criterion sample, Anglo-Saxon children. Using Hispanic children as an example, it appears that these differences are erased if the child is a third-generation American. Thus, the issue might be more that the WISC might be measuring some type of acculturation process.

Nevertheless, conflicting and nonconclusive evidence is often found. For example, in one thorough review of the literature, the race of the

examiner did not seem to affect the validity of intelligence scores in African-American children (Graziano, Varca, & Levy, 1982). Using the criteria outlined by Jensen (1980) for determining bias in testing, Sandoval (1979) concluded that the "WISC-R appears to be non-biased for minority group children." Other factors are presented by Sandoval to explain observed minority versus majority group scores. These findings are supported by Ross-Reynolds and Reschly (1983) in a study involving Anglo, African American, Hispanic, and Native-American Papago. While no bias in the WISC-R was found against African Americans and Hispanics, ceiling effects influenced the response pattern of the Papagos.

Language, however, may be confounded in bilingual children and thus needs to be clarified prior to the administration of the WISC-R. Sandoval (1979) examined the evidence of cultural bias for Anglo, Hispanic, and African-American children. Further, the Spanish version of the WISC-R does not have acceptable norms for each cultural or ethnic group and should be used with extreme caution. Concern is also cited by Dana (1984) who indicated that the WISC-R is biased for traditional Native-American children. He indicated that a pattern of spatial > sequential > conceptual > acquired knowledge exists across both ages and tribes. The difficulties associated with using the WISC in other ethnic-minority cultures is discussed in Puente and Salazar (1998).

Lampley and Rust (1986) examined the validity of the Kaufman Assessment Battery for Children and found that African Americans scored significantly lower on this test. These findings are supported by others (e.g., Sandoval & Miille, 1980). Nevertheless, these conclusions are in direct contrast to those of Hickman and Reynolds (1986-1987) who reported that "blacks did not perform significantly better in the test form developed solely on their own item statistic."

It seems that regardless of the data, contrasting interpretations abound. An interesting and eloquent attack on these issues was leveled by George Jackson, chair, Association of Black Psychologists in 1975. A more balanced perspective on this issue is presented by Cole (1981) as well as Reynolds and Brown (1984) and Helms (1992). Additional commentaries and rebuttals are found in the 1985 article by Jensen in *Behavioral and Brain Sciences.*

Little information is found for adult intelligence testing with the Wechsler Adult Intelligence Scale Revised (WAIS-R) and, due to its recent publication date, the WAIS-III (though the items appear to be much less culturally biased and the norms are more reflective of the American population). For example, in the first edition of this handbook, Lindenmann and Matarazzo (1984) indicated that the Army Alpha was developed for literates and the Army Beta for the non-English speaking. The implicit assumption is that non-English-speaking individuals were illiterate. Of course, if the dominant language becomes that of the client, then it is the psychologist who is illiterate.

Using both the WAIS and the WAIS-R, Whitworth and Gibbons (1986) reported that differences were found using both tests and that the most significant differences appeared to be the conversion of race to scale scores. Reynolds, Chastain, Kaufman, and McLean (1987) re-analyzed the data for the 1981 standardized sample of the WAIS-R and reported a 141/2 point difference between whites and African Americans on the Full Scale IQ. In attempting to resolve these discrepancies, Grubb (1987) examined the IQ differences in profoundly and severely mentally retarded individuals using Weschler's test. He reported no differences between whites and African Americans in this sample of subjects and concluded that lower IQ scores of African Americans were not biologically determined and, instead, were attributable to other factors.

Unfortunately, few data other than the results of the Weschler tests exist on measures of intellectual abilities. While one might expect that such tests as the Raven Progressive Matrices and the Beta would be less ethnically biased, the data provide little support for this (or contradictory) views. For example, using minority group offenders, Hiltonsmith (1984) reported that these subjects actually scored lower on the Beta than on the WAIS-R.

Obviously, one of two things must be occurring. There is either incorrect measurement of intellectual function or some difference (not deficiency) is present. Before accepting the possibility of difference, measurement error must be eliminated or reduced to the lowest possible level. One possible way to address this is to use greater care and ingenuity in the construction of intellectual tests. Using the WISC as an example. Care must be taken in the development of

items that are not culturally biased, both across cultures and within subcultures (e.g., Cubans versus Mexicans). Second, greater care must be taken in the standardization process. A typical protocol might include two phases, a try-out phase that helps develop further an item pool and a standardization phase that would closely mimic the U.S. population.

Another possibility is to consider intelligence from a totally different perspective. For example, Sternberg (1996) has suggested that intelligence is really nothing more than success in life. Hence, tests such as the Learning Potential Assessment Device (Feuerstein et al., 1979), Cognitive Assessment System (Naglieri & Das, 1996), as well as standard tests from the neuropsychological literature provide a more unique way to address the possible underlying variable, problem solving, in intelligence. Indeed, it is expected that future tests of intelligence will have strong foundations in neuropsychological performance.

Neuropsychological

It is often assumed that brain functions are not affected by non-neurological variables. A review of the table of contents of major neuropsychological texts of the 1980s and 1990s suggests that issues of culture, ethnicity, and race have not been addressed to date. Even more revealing are the reference sections of the books, which indicate that very few articles on these issues exist. A review of the existing journal literature also exposes the paucity of references surrounding neuropsychological assessment and the effects of culture, ethnicity, and race. *In Reliability and Validity in Neuropsychological Assessment,* Franzen (in press) presents an excellent overview of issues concerning most measures of neuropsychological ability. While different forms of validity are considered, no mention is made of the application of the tests to minority group members.

Most of the sparse data that do exist on this topic are found in the non-neuropsychological literature. For example, Lopez and Romero (1988) assessed intellectual functions in Spanish-speaking adults using both the WAIS and the Puerto Rican version of the WAIS. While the authors report that differences did exist, test equivalence is generally elusive and its application for these tests to a neuropsychological sample would be at best haphazard. On a more theoretical orientation, Drasgow (1972) addressed test-item bias and differential validity by using a "profoundly" biased test. However, in this case (as with all others), no direct or indirect mention is made of neuropsychological tests.

Anecdotal and clinical evidence indicate that these variables may have little, if any, effect on specific sensory and possibly motor measures. Some support for this contention exists. For example, Roberts and Hamsher (1984) administered both the Facial Recognition and Visual Naming Tests of the Multilingual Aphasia Examination to African Americans in a consultation setting. They reported negligible racial bias. In contrast, Adams, Boake, and Crain (1982) found that bias did exist with regard to several variables, including ethnicity, in neuropsychological performance. In both brain-damaged and normal samples, African Americans and Mexican Americans exhibited more errors than did Caucasian participants. One may extrapolate from early (though questionable) motor-learning studies on race that motor measures may be affected by race. However, as implied, the data are questionable because of numerous methodological and theoretical issues. Other individual variables are definitely affected. Language, for example, is a difficult variable to measure across groups because it contains syntactical, grammatical, and cultural content that precludes a direct translation/interpretation of a specific concept. For example, the location in a sentence of nouns and verbs differs across certain languages. Another example involves the Spanish alphabet, which contains two additional letters, ñ and Π. Cognitive styles may similarly be affected because of variables, which directly affect cognitive manipulations, such as specific style or analysis of information. Additionally, indirect variables may play a role. Asians or Hispanics not acculturated to North American norms may find it difficult to permit a professional to examine "their minds." In certain subcultures this probing is allowed only by medicine men, witch doctors, or "curanderos." Thus, it may be impossible to obtain valid data because of the client's fear of testing.

While few individual neuropsychological tests have been adapted or translated, the two most widely used batteries, the Halstead-Reitan and Luria-Nebraska Neuropsychological Batteries, have been used with diverse populations. Both of

these batteries have been translated into Spanish (HRNB by Melendez; LuriaNebraska by Puente and colleagues) and are presently being used in other cultures (e.g., Chinese). Of the two, the Halstead-Reitan may prove, at least initially, to be more adaptable since the focus is less on language function than is the Luria-Nebraska. In both cases, however, the lack of data from diverse populations is presently hindering their application.

The data that do exist, though extremely sketchy, may indicate the direction for future research. For example, complications are introduced in a report on sex, age, developmental variables, and cognitive functioning by Denno, Meijs, Nachsshon, and Anrand (1982). Differences were noted on a variety of cognitive tests (e.g., Stanford-Binet) but only for four and eight year olds. Specifically, "white males scored the highest on all tests, followed by white females, black females and black males." Thus, variables such as sex and age may interact with race (and other variables). If these studies are found to be valid examples of neuropsychological measures, then a clear and easy identification of variables contributing to diversity of neuropsychological performance may not be feasible.

More recently, Perez-Arce and Puente (1996) reviewed the literature with a focus on understanding ecological validity of neuropsychological tests for Hispanics living in North America. If neuropsychological assessment focuses on problem solving, they suggest that different problem-solving strategies are employed by Hispanics. For example, many neuropsychological tests use time in assessing brain dysfunction. Whereas, in a competitive culture, like the American, time is a critical variable not to be wasted, and so on, the opposite is often true for other cultures. Hence, slowed performance, which could actually be interpreted by the Hispanic as prolonging a task of interest, would be incorrectly interpreted as "brain-damage". Unfortunately, in their review the authors conclude that outside the work of a few researchers such as Ardila (e.g., Ardila, Rosselli, & Puente, 1994), little is found in the literature.

Personality and Pathology

Tests of personality could be generally categorized as one of two types—projective or objective. Projective or cognitive-perceptual tests (e.g., Rorschach) are quite commonly used with minority members because of their inherent ease of administration and superficial adaptability and interpretation. According to preliminary analyses by Exner and Sciara (personal communication, July 7, 1989), the Rorschach, an internationally accepted measure of cognitive-perceptual status, does not appear to be biased against Asian Americans, African Americans, or Mexican Americans. However, it is important to state that no data exist to support this contention.

In contrast, limitations of test adaptability are more readily investigated for issues such as bias with objective measures. Clearly, the most often used test of personality and psychopathology is the Minnesota Multiphasic Personality Inventory (MMPI). The homogeneity of the original MMPI sample limits its ready application to minority groups (see Butcher, Braswell, & Raney 1982). According to Dahlstrom, Welsh, and Dahlstrom (1972), the norms used in the original MMPI sample were Caucasian, married, rural, blue-collar workers, with an eighth-grade education. However, Dahlstrom, Diehl, and Lachar (1986); Dahlstom, Lachar, and Dahlstrom (1986), and Lachar, Dahlstrom, and Moreland (1986) have suggested that even when important demographic variables are taken into account (e.g., race and socioeconomic status), approximately 12 percent to 13 percent of the total variance of the basic scales is accounted for. Still, the popularity of the test has resulted in translation into approximately 100 languages (Butcher, 1984; Williams, 1987) and a wealth of cross-cultural, ethnic, and racial studies based on research using this instrument have been published.

In an early review of ethnicity and the MMPI, Greene (1987) did an exhaustive examination of studies. Over 100 studies were analyzed according to type of scale and item level across groups including African American white, Hispanic white, Asian American, and Native American. Greene concluded that too many variables and too few adequately completed studies prevent conclusions of bias. The variables in question include subject parameters, ethnic group membership profile validity, moderator variable, and scores analyzed. Additional methodological considerations include appropriate statistical analyses, adequate sample size, and validity of statistical (versus clinical) significance. Based on

his review, Greene provided the following four conclusions:

1. At this stage of our understanding, it is too premature to develop norms for specific ethnic and racial groups.
2. Subjects have to be identified with an ethnic group using subjective self (not clinician or experimenter) identification.
3. Empirical and not clinical differences should be emphasized.
4. Finally, more research needs to be focused on the special scales of the MMPI.

As exhaustive as the review is and as heuristic as Greene's conclusions may be, others advocate different orientations. For example, Gynther (1981), Gynther and Green (1980), and others argue that specific norms—and, in some cases, items—be developed, using an empirical methodology rather than a review of the literature. Better understanding of ethnic, cultural, and race differences and their application to interpretation of *T* scores, specific scale scores, or patterns preclude widespread use of the MMPI with minorities. For example, it seems foolish to group all Hispanics together as Greene and others have done. As Sue and Zane (1987) have indicated, being culturally sensitive is being aware of within-group heterogeneity. Further, little understanding appears evident in the MMPI research with regard to differences among culture, ethnicity, and race. Until such issues, as well as those outlined by Greene (1987), are resolved, not only will the MMPI data as it now stands be premature; it will be incorrect. According to a recent announcement from the Restandardization Committee of the University of Minnesota Press (1989) concerning the MMPI-2, published in 1989, the revised version will have "national norms that are much more representative of the present population of the U.S." (p. 4).

Dana (1995) criticizes the sampling methods of the MMPI. In the MMPI, for example, Hispanics were underrepresented by 2.8 percent. In the second MMPI-2 similar limitations are still noted. Further, differences are noted in scale performance between Hispanics and their Anglo-Saxon counterparts. Main differences are noted in the following scales; L, K, 3, and 4. He suggests that clinical interpretation must hinge on understanding the acculturation of the individual.

Using the recommendations of Velazquez, and colleagues (1997) for the use of the MMPI with Hispanics, the following recommendations are suggested for all ethnic minorities:

1. When options are available, use the most recent version of the test.
2. Administer the entire, and not a short, version of the test.
3. Appreciate prior test-taking experience of the test-taker.
4. Test in the language selected by the test-taker.
5. Evaluate results within a bio-psychosocial context.
6. Appreciate the effects of acculturation on test results.
7. Interpret results based on research literature and not on cultural stereotypes.
8. Always use a variety of test sources to arrive at conclusions.

Achievement, Aptitude, and Interest

Achievement tests are still widely used in a variety of settings. A starting point involving achievement assessment is that of achievement motivation (Basic Behavioral Science Task Force of the National Advisory Mental Health Council, 1996). Indeed, it assumed that motivation plays a relatively small role in achievement testing. Nevertheless, access to models, ethnic-minority status, and related variables produce an initial "handicap" in such testing. Unfortunately, motivation appears highly correlated with scores on achievement tests as well as academic performance. Hence, what one might be measuring with ethnic minorities is not achievement as much as motivation.

What does exist, as with many other psychometric instruments, is a paucity of data. In chapter 7 of this handbook, there is a comprehensive review of achievement tests. Of the tests discussed in that chapter, the California Achievement Test (in education) and the Wide Range Achievement Test (in education and clinical application) are two of the most frequently used tests, which have been applied to nonmajority samples of the U.S. population. Initial findings regarding test bias in these measures reflect the conclusions outlined by Fox and Zirkin (1984) in the first edition of this handbook. Specifically,

they suggest that while attention should be paid to the possibility of such bias, and while it may be intuitive that such bias would exist (at least on specific items), these tests should not be considered biased. This conclusion is in direct contrast to others, however. For example, Weiss (1987), considered the Scholastic Aptitude Test especially biased in the verbal section. While Golden Rule procedures have been applied to reduce such biases, the reliability and validity of these tests may be in jeopardy (Linn & Drasgow, 1987). Thus, conflicts exist in terms of having a useful but unbiased test of achievement.

For tests of interest, even fewer data exist. While separate scales for sex are the rule and not the exception for measures of occupation it is generally assumed that other variables are of little importance. The same applies for interest surveys. For example, the Kuder Occupational Interest Survey (Form DD) (Kuder, 1966) as well as the Holland Interest Inventories (1979) consider academic major, occupational status, and even personality type, but not cultural, race, or ethnic factors. The Strong-Campbell is available in Spanish but the norms presumably are from non-Spanish-speaking samples. In a recent study, Drasgow and Hulin (1987) attempted to answer the question of whether scores on the Job Description Index (a vocational measure) varied across different Hispanic populations. Specifically, they compared bilingual Mexicans in Mexico City to other Hispanics residing elsewhere. While few differences were noted between the New York and Miami samples, large differences were noted between the U. S. and Mexican samples. Drasgow and Hulin concluded that both linguistic and cultural measurement equivalence must be addressed in measures of vocational interest.

What appears to exist is that differences between ethnic minorities and the criterion counterparts exist as early as the first grade (Jackson, 1975; Task Force of the National Advisory Mental Health Council, 1996). Asian students appear to perform better than either Anglo-Saxons or other ethnic minorities. Such differences are even greater by the fifth grade. Asian families tend to explain this in motivational differences. In contrast, American mothers consider good performance of their children to be related to "natural" ability. Unfortunately, similar data are not available on Hispanic and African-American children. However, it is often thought that, even when intelligence is controlled, achievement differences may be due to biological variables in these ethnic-minority groups. According to the task force report, "Americans regard the need to try harder as evidence of low innate ability and are less likely to value or encourage such effort" (p. 725).

While it is assumed that ethnicity, race, and related variables have been explored by the Educational Testing Service and related psychological test corporations, again few scientific data exist in the public domain regarding tests of aptitude. Terman (1916) helped develop the now widely used Stanford Achievement Test for precollege screening with no reference to minority groups. At the college level, the College Advanced Placement Examination is also widely used and accepted. However, data on minority populations is still lacking for both of these instruments since relatively few ethnic minorities enroll in such programs.

Culturally Sensitive Measures

Traditionally, one method of avoiding test bias with regard to culturally different populations is to use instruments that are sensitive to and factor out cultural variables. Of the attempts to diminish test bias, the most significant effort has been by Cattell (1963). His Culture-Fair Intelligence Test measures intellectual abilities that allegedly factor out culture.

Cattell's basic aim was to factor out both cultural and educational variables from intellectual factors. Items were developed on common rather than culturally specific knowledge. Based on initial speculation, Cattell suggested that fluid intelligence was a function of biological factors including genetic and constitutional ones. In contrast, crystallized intelligence was a result of the development of fluid intelligence through environmental and cultural opportunities. While the Culture-Fair Test has been regularly used in the United States, its popularity has extended to non-North American populations. To date the instrument has been used with Nigerian (Nenty, 1986), Bulgarian (Paspalanova & Shtetinski, 1985), Italian (Stepanile, 1982), Spanish (Ortega-Esteban, Ledesma-Sanz, Lopez-Sanchez, & Prieto-Adanez, 1983), Israeli (Zeidner, 1987), and Indian (Ravishankar, 1982) groups. Unfortunately, the test has been shown to exhibit bias in

some (e.g., Nigerian) though not all populations. In addition, these studies were completed with individuals residing in their own culture. It would be interesting to explore the efficacy of this test with minority cultures residing in the United States. While this test is promising in theory, additional research both in the United States and abroad will have to occur prior to its wider clinical acceptance. It is interesting to note that Cattell who died very recently has been accused as being biased. Indeed, this accusation was dramatically brought to the attention of psychologists by the 1997 American Psychological Association Keynote Speaker, Ellie Weisel. Unfortunately, the blue-ribbon panel that was convened to explore these allegations was disbanded before any discussions ensued (partially due to the fact that Cattell removed his name from further consideration for the award for which he was being considered).

Of all standardized tests, the WAIS has received most attention with regard to cultural standardization. Two excellent examples are the Canadian and Puerto Rican versions of the test (Wechsler, 1960). Violato (1984) administered the standard or a revised version of the WAIS to 101 Canadians. The revised version contained eight items that were changed to increase face validity for Canadians. While bias effects were limited, the author did suggest that changes for Canadian administration of the WAIS were necessary. The WAIS has also been translated and standardized with Puerto Rican populations (1980). It was assumed that all translations would be appropriate; this assumption, however, is incorrect. Puerto Rican, Chicano, Mexican, Latin American, South American, and Castilian Spanish not only have their own dialects and idiosyncrasies but in many cases, their own language. Thus, the Puerto Rican translation of the WAIS has limited usefulness with non-Puerto Rican subjects. Further, though yet to be researched, the issue of norms needs to be addressed. For example, Puerto Rican norms may differ from Argentinean norms. Also, there is the question of when an individual, from one culture but residing in another, becomes acculturated enough to be administered the "new" culture's tests. These and related questions remain to be answered.

Other tests of intellectual ability which are purported to be culture-reduced or fair include Raven's Progressive Matrices—both Colored and Standard versions—as well as the Peabody Picture Vocabulary Test, the Quick Test, and the Army Beta. However, little evidence exists on the ability of these tests to be culture free. With the Picture Vocabulary Test serving as an example, several of the pictures on this test are useful for North American but not British populations. Another interesting example is that of the LuriaNebraska Neuropsychological Battery. Certain sections and stimuli are deemed culture free or culture reduced; but several of the visual stimuli come from Denmark and not Nebraska, making clear identification of specific items (e.g., nutcracker) an often difficult if not impossible task.

Less and less bias research is being done, as indicated earlier, since the belief that bias does not exist is so prevalent (Suzuki & Valencia, 1997). Hence, one might assume that tests such as the ones developed by Cattell will continue to lose favor as psychologists continue to, possibly incorrectly, assume that culture is not a critical factor in psychological assessment.

Behavioral Assessment

In another section of this handbook, chapters on behavioral assessment are found. One major focus of this type of assessment is the assumption that behavioral, versus psychometric, approaches to assessment reduce the risk of focusing on psychic and nonobservable attributions. Psychometric focus may increase the potential for incorrect understanding of the behavior in question and, of course, is more likely to introduce bias in the assessment process.

Behavioral assessment focuses on empirically based methods of understanding behavior and, thus, the application to minority populations seems obvious. If psychometric tests are riddled with questions of culture, race, and ethnicity, then an assessment procedure, which focuses on behavior, and places the individuals in question in their environmental context, would seem an excellent alternative. Hence, it is surprising to note that this application has not been considered and researched adequately.

What scientific literature does exist is limited and, at best, preliminary. For example, Slate (1983) attempted to compare three nonbiased "behavioral" measures in retarded and nonretarded children across race and social class.

Unfortunately, the results are so convoluted that they preclude an adequate understanding of any of these measures. Further, the possibility exists that behavioral measures may themselves be biased, partially due to the rater as well as the rating instruments. Lethermun, Williamson, Moody, and Wozniaz (1986) examined the effects of the race of the rater on the rating of social skills of African-American and white children. The results support earlier findings that the race of the child affects the ratings received. In addition, the researchers reported that racial bias effects were noted with both African-American and Caucasian raters.

While intuitive support exists for the use of behavioral assessment with non-mainstream populations, complications are evident in the literature. First, little data and even less clinical application of this approach are available. In addition, initial studies suggest that bias may still be present both in terms of the rated and the rater.

SUMMARY

Understanding human behavior requires an understanding of human diversity. Unfortunately, historical foundations have dictated an incorrect understanding of how culture, race, ethnicity, and related demographic variables affect human behavior. This situation is evident in the traditional and current use of psychological tests to measure such variables as intelligence, achievement, abilities, aptitude, personality, and neuropsychological function.

Two factors appear to have guided this incorrect measurement of human diversity. First, pioneers such as Terman not only suggested that minorities were inferior but that their "proliferation should be controlled." Legislation and adjudication addressing minority bias continues to this day, especially in California and Texas, even at the level of the Supreme Court. Indeed, recent rulings on affirmative action call these issues into question. Second, few researchers, academicians, or clinicians have devoted time and effort to answering pertinent questions on human diversity, and even fewer have studied psychological assessment of diversity. Recently published statistics indicate that fewer ethnic minorities than in earlier years are pursuing graduate training in psychology or the study of human diversity. The lack of interested personnel is mirrored in faculty and clinical positions throughout North America.

Of course, the possibility exists that what is actually occurring is what has been previously described as "the false uniqueness effect" which is the tendency to overestimate one's personal positive attributes and underestimate other's abilities (Basic Behavioral Science task force of the National Advisory Mental Health Council, 1996). This Task Force reports that American children tend to think better of themselves when compared to others, presumably ethnic minorities. As adults, Americans tend to think of themselves as more attractive and intelligent than average. Further, 60 percent of students believed that they were in the top 10 percent of ability to get along with others—clearly an impossibility. One might assume that what has transpired is that American psychology has failed to understand the possibility of "false uniqueness effect" and has confused such variables as test intelligence, which are highly influenced by economic and related factors, with biological and genetic superiority (see Jensen, 1980 for further information). It appears that psychologists have decided that when an ethnic-minority group does better than the majority group (as in the case for Asian Americans), the attributing variable is motivation and cultural variables (e.g., Suzuki & Valencia, 1997). In contrast, when an ethnic minority does poorer than a majority group it is often ascribed to biological variables. Differential attribution of ethnic-minority differences—if better, it must be due to effort; if worse, it must be due to genetics—represents intellectual imperialism.

The obvious outcome is a field lacking in adequate data and much emotionality. The data that are available are clouded not only by a host of methodological problems but by researchers' gross misunderstanding of ethnic-minority group members and membership (including but not limited to within-group heterogeneity), especially in the context of majority-group behavior patterns. Regardless of the absence of data, social policy continues forward—often guided by political but not scientific correctness. Thus, much effort needs to be directed to the areas of research, teaching, and services to minority group members. Until additional adequate information is available, extreme caution should be used in the application of present knowledge of the psychological assessment of ethnic-minority group

members and in the acceptance of previously considered "universal" theories of human function.

REFERENCES

Adams, R. L., Boake, C., & Crain, C. (1982). Bias in a neuropsychological test classification related to education, age, and ethnicity. *Journal of Consulting and Clinical Psychology, 50,* 143–145.

Adebimpe, V. R. (1981). Overview: White norms and psychiatric diagnoses of black patients. *American Journal of Psychiatry, 138,* 279–285.

Ardila, A., Rosselli, M., & Puente, A. E. (1994). *Neuropsychological evaluation of the Spanish speaker.* New York: Plenum.

Basic Behavioral Science Task Force of the National Advisory Mental Health Council (1996). Basic behavioral science research for mental health. Sociocultural and environmental processes. *American Psychologist, 51,* 722–731.

Beach, F. (1950). The snark was a boojum. *American Psychologist, 5,* 115–124.

Bernal, M., & Castro, F. (1994). Are clinical psychologists prepared for service and research with ethnic minorities? *American Psychologist, 49,* 797–805.

Berry, J. W. (1990). Psychology of acculturation. In J. J. Berman (Ed.), *Nebraska symposium of motivation, 1989: Cross-cultural perspectives, 37,* 201–234.

Bersoff, D. N. (1981). Testing and the law. *American Psychologist, 36,* 1047–1056.

Betancourt, H., & Lopez, S. R. (1993). The study of culture, ethnicity, and race in American psychology. *Americant Psychologist, 48,* 629–637.

Boring, E. G. (1950). *A history of experimental psychology* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Brislin, R. W. (1988). Increasing awareness of class, ethnicity, culture, and race by expanding students' own experience. In E. S. Cohan (Ed.), *The G. Stanley Hall lecture series* (Vol. 8, pp. 137-180). Washington, DC: American Psychological Association.

Bumam, M. A., Telles, C. A., Kamo, M., & Hough, R. L. (1987). Measurement of acculturation in a community population of Mexican Americans. *Hispanic Journal of Behavior Science, 9,* 105.

Butcher, J. N. (1984). Current developments in MMPI use: An international perspective. In J. N. Butcher & C. D. Spielberger (Eds.), *Advances in personality assessment.* Hillsdale, NJ: Erlbaum.

Butcher, J. N. (1987). Introduction to the special series: Cultural factors in understanding and assessing psychopathology. *Journal of Consulting and Clinical Psychology, 55,* 459–460.

Butchner, J. N., Braswell, L., & Raney, D. (1982). A cross-cultural comparison of American Indian, black & white inpatients on the MMPI and persisting symptoms. *Journal of Consulting and Clinical Psychology, 51,* 587–594.

Cattell, R. R. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology, 54,* 1–22.

Cole, N. S. (1981). Bias in testing. *American Psychologist, 36,* 1067–1077.

Cuellar, I., Arnold, B., & Maldonado, R. (1995). Acculturation Rating Scale for Mexican American II: A revision of the original ASRMA scale. *Hispanic Journal of Behavioral Science, 17,* 275–304.

Cuellar, I., Harris, L. C., & Jasso, R. (1980). An acculturation scale for Mexican American normal and clinical population. *Hispanic Journal of Behavioral Science, 2,* 197–217.

Dahlstrom, W. G., Diehl, L. A., & Lachar, D. (1986). MMPI correlates of the demographic characteristics of black and white normal adults. In W. G. Dahlstrom, D. Lachar, & L. E. Kahistrom (Eds.), MMPI patterns of American minorities (pp. 104-138). Minneapolis: University of Minnesota Press.

Dahlstrom, W. G., Lachar, D., & Dahlstrom, L. E. (Eds.) (1986). *MMPI patterns of American minorities.* Minneapolis: University of Minnesota Press.

Dahlstrom, W. G., Welsh, G. S., & Dahlstrom, L. E. (1972). *An MMPI handbook. Volume I: Clinical interpretation.* Minneapolis: University of Minnesota Press.

Dana, R. H. (1984). Intelligence testing of American Indian children: Sidesteps in quests of ethical practice. *White Cloud Journal, 3,* 35–43.

Dana, R. H. (1995). Culturally competent MMPI assessment of hispanic populations. *Hispanic Journal of Behavioral Sciences, 17,* 305–319.

Denno, D., Meijs, B., Nachshon, I., & Aurand, S. (1982). Early cognitive functioning: Sex and race differences. *International Journal of Neuroscience, 16,* 159–172.

Drasgow, F. (1972). Biased test items and differential validity. *Psychological Bulletin, 92,* 526–531.

Drasgow, F., & Hulin, C. L. (1987). Cross-cultural measurement. *Revista Interamericana de Psicologia, 21,* 1–24.

Eells, K. (1951). *Intelligence and cultural differences.* Chicago: University of Chicago Press.

Feuerstein, R., Rand, Y., & Hoffman, M. (1979). *The dynamic assessment of retarded performes: The Learning Potential Assessment Device, theory, instruments, and techniques.* Baltimore: University Park Press.

Flaugher, R. L. (1978). The many definitions of test bias. *American Psychologist, 33,* 671–679.

Fowlers, B. J., & Richardson, F. C. (1996). Why is multicuralism good? *American Psychologist, 51,* 609–621.

Fox, C. H., & Zirkin, B. (1984). Achievement tests. In G. Goldstein & M. Hersen (Eds.), *Handbook of psychological assessment.* New York: Pergamon Press.

Fradd, S., & Hallman, C. L. (1983). Implications of psychological and educational research for assessment and instruction of culturally and linguistically different students. *Learning Disability Quarterly, 6,* 468–478.

Franco, J. N. (1983). An acculturation scale for Mexican-American children. *Journal of General Psychology, 108,* 175–181.

Franzen, M. D. (in press). *Reliability and validity in neuropsychological assessment* (2nd ed.). New York: Plenum Press.

Goddard, H. H. (1912). *The Kallikakfamily: A study in the heredity offeeble-mindedness.* New York: Macmillan.

Gonzales, R. R., & Roll, S. (1985). Relationship between acculturation, cognitive style, and intelligence: A cross-sectional study. *Journal of Cross-Cultural Psychology, 16,* 190–205.

Gould, S. J. (1981). *The mismeasurement of man.* New York: W. W. Norton & Company.

Gould, S. J. (1996). *The mismeasurement of man.* New York: W. W. Norton & Company.

Graziano, W., Varca, P., & Levy, J. (1982). Race of examiner effects and the validity of intelligence tests. *Review of Educational Research, 52,* 469–497.

Greene, R. L. (1987). Ethnicity and MMPI performance: A review. *Journal of Consulting and Clinical Psychology, 55,* 497–512.

Greenfield, P. M. (1997). You can't take it with you. Why ability assessment don't cross cultures. *American Psychologist, 52,* 1115–1124.

Grubb, N. J. (1987). Intelligence at the low end of the curve: Where are the racial differences? *Journal of Black Psychology, 14,* 25–34.

Guthrie, R. (1976). *Even the rat was white.* New York: Harper & Row.

Gynther, M. D. (1981). In the MMPI an appropriate device for blacks. *Journal of Biack Psychology, 7,* 67–75.

Gynther, M. D., & Green, S. B. (1980). Accuracy may make a difference but does a difference make for accuracy? A response to Pritchard and Rosenblatt. *Journal of Consulting and Clinical Psychology, 48,* 268–272.

Hall, Ch.C. (1997). Cultural malpractice. The growing obsolescence of Psychology. *Americant Psychologist, 52,* 642–651.

Helms, J. E. (1992). Why is there no study of cultural equivalence in standardized cognitive ability testing? *American Psychologist, 47,* 1083–1101.

Herrnstein, R. J., & Murray, C. (1994). *The bell curve.* New York: The Free Press.

Hickman, J. A., & Reynolds, C. R. (1986–1987). Are race differences in mental test scores an artifact of psychometric methods? A test of Harnngton's experimental model. *Journal of Special Education, 20,* 409–430.

Hiltonsmith, R. W. (1984). Predicting WAIS-R scores from the Revised Beta for low functioning minority group offenders. *Journal oyclinical Psychology, 40,* 1063–1066.

Hoffmann, T., Dana, R., & Bolton, B. (1985). Measured acculturation and MMPI-168 performance of Native American adults. *Journal of Cross-Cultured Psychology, 16,* 243–256.

Holland, J. L. (I 979). *The self-directed search professional manual.* Palo Alto, CA: Consulting Psychologists Press.

Howard, A., Pion, G. M., Gottfredson, G. D., Flattau, P. E., Oskame, S., Pfafflin, S. M., Bray, D. W., & Burstein, A. G. (1986). The changing face of American psychology: A report from the Committee on Employment and Human Resources. American *Psychologist, 41,* 1311–1327.

Humphries, L. G. (1986). An analysis and evaluation of test and item bias in the prediction context. *Journal ofapplied Psychology, 71,* 327–333.

Jackson, G. D. (1975). On the report of the Ad Hoc Committee on educational tests with disadvantaged students: Another psychological view from the Association of Black Psychologists. *American Psychologist, 30,* 88–93.

Jensen, A. (1980). *Bias in mental testing.* New York: Free Press.

Jensen, A. R. (1986). Construct-validity and test bias. *Phi Delta Kappan, 58,* 340–346.

Kim, U., & Choi, S-H. (1994). Individualism, collectivism, and child development: A korean perspecitve. In P. M. Greenfield & R. R. Cocking (Eds.), *Cross-cultural roots of minority child development* (pp. 227–257). Hillsdale, NJ: Erlbaum.

Knight, G. P. & Kagan, S. (1977). Acculturation of prosocial and competitive behaviors among second- and third- generation Mexican-American children. *Journal of Cross-Cultural Psychology, 8,* 273-284.

Kuder, G. F. (1966). *General manual: Occupational interest survey form DD.* Chicago:Science Research Association.

Kuhn, T. S. (1970). *The structure of scientific revolutions* (2nd ed.). Chicago: University of Chicago Press.

Lachar, D., Dahlstrom, W. G., & Moreland, K. L. (1986). Relationship of ethnic background and other demographic characteristics to MMPI patterns in psychiatric samples. In W. G. Kahlstrom, D. Lackar, & L. E. Dahlstrom (Eds.), *MMPI patterns of American minorities* (pp. 139–178). Minneapolis: University of Minnesota Press.

Lampley, D. A., & Rust, J. V. (1986). Validation of the Kaufman Assessment Battery for Children with a sample of preschool children. *Psychology in the Schools, 23,* 131–137.

Lawson, W. B. (1987). Racial and ethnic factors in psychiatric research. *Hospital and Community Psychiatry, 37,* 50–54.

Leahey, T. H. (1997). *A history of psychology* (4th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Lethermun, V. R., Williamson, D. A., Moody, S. C., & Wozniak, P. (1986). Racial bias in behavioral assessment of children's social skills. *Journal of Psychopathology and Behavioral Assessment, 8,* 329–332.

Lindenmann, J. E., & Matarazzo, J. D. (1984). Intellectual assessment of adults. In G. Goldstein & M. Hersen (Eds.), *Handbook of psychological assessment* (pp. 77–79). New York: Pergamon Press.

Linn, R. L., & Drasgow, F. (1987). Implications of the Golden Rule settlement for test construction. *Educational Measurement Issues and Practice, 6,* 13–17.

Lopez, S., & Romero, A. (1988). Assessing the intellectual functioning of Spanish-speaking adults: Comparison of the EIWA and the WAIS. *Professional Psychology: Research and Practice, 19,* 263–270.

McGovern, T. V., Furumoto, L., Halpern, D. F., Kimble, G. A., & McKeachie, W. J. (1991). Liberal education, study in depth, and the arts and sciences major-Psychology. *American Psychologist, 46,* 598–605.

Magana, J. R., de la Rocha, O., Amsel, J., Magana, H. A., Fernandez, M. I., & Rulnick, S. (1996). Revisting the dimensions of acculturation: Cultural theory and psychometric practice. *Hispanic Journal of Behavioral Sciences, 18,* 444–468.

Maheady, L., Towne, R., Algozzine, B., Mercer, J., & Ysseldyke, J. (1983). Minority overrepresentation: A case for alternative practices prior to referral. *Learning Disability Quarterly, 6,* 448–456.

Malgady, R. G. (1996). The question of cultural bias in assessment and diagnosis of ethnic minority clients: Let's reject the null hypothesis. *Professional Psychology: Research and Practice, 27,* 73–77.

Marin, G., Sabogal, F., Marin, B., & Otero-Sabogal, R. (1984). Development of a short acculturation scale for Hispanics. *Hispanic Journal of Behavioral Sciences, 9,* 183–205.

Montgomery, G. T., & Orozl, S. (1984). Validation of a measure of acculturation for Mexican Americans. *Hispanic Journal of Behavioral Sciences, 6,* 53–63.

Naglieri, J. A., & Das, J. P. (1996). *Das Naglieri Cognitive Assessment System.* Chicago: Riverside.

Neisser, U., Boodoo, G., Bouchard, T. J., Boykin, A. W., Brody, N., Ceci, S. J., Halpern, D. F., Loehlin, J. C., Perloff, R., Sternberg, R. J., & Urbina, S. (1996). Intelligence: Knowns and unknows. *American Psychologist, 51,* 77–101.

Nenty, H. J. (1986). Cross-culture bias analysis of Cattell Culture-Fair Intelligence Test. *Perspectives in Psychological Researches, 9,* 1–16.

Olmedo, E. L. (1981). Testing linguistic minorities. *American Psychologist, 36,* 1078–1085.

Ortega-Esteban, J., Ledesma-Sanz, A., Lopez-Sanchez, F., & Prieto-Adanez, G. (1983). Profit of the academically successful student in the Spanish universities. *Sciencia Pegagogica Experimentalis, 20,* 62–82.

Padilla, A. M. (1985). *Acculturation and stress among immigrants and later generation individuals.* Spanish-Speaking Mental Health Research Center: Occasional Paper, 20, 11–60.

Paspalanova, E., & Shtetinski, D. (1985). Standardization of the CF 2A Intelligence Test of Cattel for Bulgarian Population. *Psikhologiia Bulgaria,* 1985, 12-–22 (translation).

Perez-Arce, P., & Puente, A. E. (1996). Neuropsychological assessment of ethnic minorities: The case of assessing Hispanics living in North America. In R. Sbordone & C. Long (Eds.) *Ecological validity in neuropsychological testing.* Delray Beach, FL: St. Lucie Press.

Phinney, J. S. (1996). When we talk about american ethnic groups, what do we mean? *American Psychologist, 51,* 918–927.

Prasse, D. (1979). Federal legislation and school psychology: Impact and implication. *Professional Psychology, 9,* 592–601.

Puente, A. E. (1993). Towards a psychology of variance. In T. McGovern (Ed.), Enhancing the undergrauduate psychology curriculum. Washington, DC: American Psychological Association.

Puente, A. E., & Salazar, G. (1998). Assessment of minority and culturally diverse children. In A. Prifitera & D. Saklofske (Eds.), *WISC-III. Clinical use and interpretatio.* New York: Academic Press.

President's Commission on Mental Health (1978). *A report to the President from the President's Commission on Mental Health.* Washington, DC: U.S. Government Printing Office.

Ravishankar, V. (1982). A correlational study of Cattell's personality factor B. and 1. Q. as measured by his culture free test. *Indian Psychological Review, 22,* 9–1 1.

Reschly, D. J. (1984). Aptitude tests. In G. Goldstein & M. Hersen (Eds.), *Handbook of psychological assessment.* New York: Pergamon Press.

Restandardization Committee of the University of Minnesota Press. (1989). MMPI-2 Minnesota Multiphasic Personality Inventory. *Critical Items, 4* (2), 1–4.

Reynolds, C. R., & Brown, R. T. (1984). Bias in mental testing. In C. R. Reynolds & R. T. Brown (Eds.), *Perspectives on bias in mental testing.* New York: Plenum Press.

Reynolds, C. R., Chastain, R. L. Kaufman, A. S., & McLean, J. E. (1987). Demographic characteristic of IQ among adults: Analysis of the WAIS-R standardization sample as a function of the stratification variables. *Journal of School Psychology, 25,* 323–342.

Roberts, R. J., & Hamsher, K. D. (1984). Effects of minority status on facial recognition and naming performance. *Journal of Clinical Psychology, 40,* 539–545.

Ross-Reynolds, J., & Reschly, D. J. (1983). An investigation of the item bias on the WISC-R with four sociocultural groups. *Journal of Consulting and Clinical Psychology, 51,* 144–146.

Sandoval, J. (1979). The WISC-R and internal evidence of test bias with minority group. *Journal of Consulting and Clinical Psychology, 47,* 919–927.

Sandoval, J., & Miille, M. P. W. (1980). Accuracy of judgments of WISC-R item difficulty for minority groups. *Journal of Consulting and Clinical Psychology, 48,* 249–253.

Scarr, S. (1988). Race and gender as psychological variables: Social and ethical issues. *American Psychologist, 43,* 56–60.

Schultz, D. P., & Schultz, S. E. (1996). *A history of modern psychology* (6th ed.). New York: Academic Press.

Slate, N. (1983). Nonbiased assessment of adaptive behavior: Comparison of three instruments. *Exceptional Children, 50,* 67–70.

Spitzer, R. L., Endicott, J., & Fleiss, W. (1967). Instruments and recording forms for evaluating psychiatric status and history: Rationale, method of development and description. *Comprehensive Psychiatty, 8,* 321–343.

Spitzer, R. L., & Williams, J. B. W. (1980). Classification of mental disorders and DSM-111. In H. I. Kaplan, A. M. Freedman, & B. J. Sadock (Eds.), *Comprehensive textbook of psychiatry-III.* Baltimore: William and Wilkins.

Sternberg, R. J. (1996). *Successful intelligence: How practical and creative intelligence determine success in life.* New York: Simon & Schuster.

Sternberg, R. J. (1997). The concept of intelligence and its role in lifelong learning and success. *American Psychologist, 52,* 1030–1037.

Sternberg, R. J., & Williams, W. M. (1997). Does the Graduate Record Examination predict meaningful success in the graduate training of psychologists? *American Psychologist, 52,* 630–641.

Stepanile, C. (1982). Contributo per una taratura intaliana de test culture fair di Cattell. *Bollettino & Applicata* (161–164), 81–86 (translation).

Sue, S., & Zane, N. (1987). The role of culture and cultural techniques in psychotherapy. *American Psychologist, 42,* 37–45.

Suinn, R. M., Richard-Figueroa, K., Len, S., & Vigil, P. (1987). The Suinn-Law Asian Self Identity Acculturation Scale: An initial report. *Educational and Psychological Measurement, 6,* 103–112.

Suzuki, L. A., & Valencia, R. R. (1997). Race-ethnicity and measured intelligence. *Americant Psychologist, 52,* 1103–1114.

Terman, L. W. (1916). *The measurement of intelligence.* Boston: Houghton-Mifflin Company.

Triandis, H. C. (1982). Acculturation and biculturation indices among relatively accultured Hispanic youths. *Revista Interamericana de Psicologia, 16,* 140–149.

U.S. Bureau of the Census (1992). *Survey of Business Owners and Self-Employed Pearsons: Form MB-1.* Washington, D.C.: Department of Commerce.

Velazquez, R. J., Gonzales, M., Butcher, J. N., Castillo-Canez, I., Apocada, J. X., & Chavira, D. (1997). Use of the MMPI-2 with Chicanos: Strategies for counselors. *Journal of Multicultural Counseling and Development, 25,* 107–120.

Violato, C. (1984). The effects of Canadianization of American-biased items on the WAIS and WAIS-R information subtests. *Canadian Journal of Behavioral Science, 16,* 36–41.

Wechsler, D. (1991). *Wechsler Intelligence Scale for Children-Third Edition: Manual.* San Antonio: The Psychological Corporation.

Weiss, J. (1987). The Golden Rule bias reduction principle: A practical reform. *Educational Measurement Issues and Practice, 6,* 23–25.

Weschler, D. A. (1960). *Escala de Inteligencia de Weschler-Adultos.* New York: Psychological Corporation.

Westermeyer, J. (1987a). Cultural factors in clinical assessment. *Journal of Consulting and Clinical Psychology, 55,* 471–478.

Westermeyer, J. (1987b). Clinical considerations in cross-cultural diagnosis. *Hospital and Community Psychiatry, 38,* 160–165.

Whitworth, R. H., & Gibbons, R. T. (1986). Cross-racial comparison of the WAIS and WAIS-R. *Educational and Psychological Measurement, 46,* 1041–1049.

Williams, C. L. (1987). Issues surrounding psychological testing of minority patients. *Hospital and Community Psychiatry, 38,* 184–189.

World Health Organization (1973). *International pilot project on schizophrenia.* Geneva, Switzerpland: Author.

Yerkes, R. N. (1923, March). Testing the human mind. *Atlantic Monthly,* pp. 358–370.

Zeidner, M. (1987). Test of the cultural bias hypothesis: Some Israeli findings. *Journal of Applied Psychology, 72,* 38–48.

Zubin, J. (1965). Cross-national study of diagnosis of the mental disorder: Methodology and planning. *American Journal of Psychiatry, 125,* 12–20.

Zuckerman, M. (1990). Some dubious premises in research and theory on racial differences: Scientific, social, and ethical issues. *American Psychologist, 45,* 1297–1303.

This Page Intentionally Left Blank

CHAPTER 22

PSYCHOLOGICAL ASSESSMENT OF THE ELDERLY

Karen L. Dahlman
Teresa A. Ashman
Richard C. Mohs

INTRODUCTION

As the proportion of the population over age 65 increases, the field of geropsychology has expanded exponentially. Incorporated into this field are issues related to the psychological assessment of the elderly, including the appropriate application and limitations of psychometric tools. The intent of this chapter is to present a context in which to understand the cognitive impact that occurs as a process of aging. The first part will define what constitutes normal aging. The second part outlines the importance of gathering a complete clinical picture of premorbid and present functioning exclusive of cognitive abilities (e.g., medical conditions, family history, social adaptation, psychiatric history). Part three focuses on cognitive functioning among older patients, including methods of determining dementia. The fourth section highlights the principles of the neuropsychological assessment process for older adults, along with the purposes utilized by the assessment. The fifth part highlights typical differential diagnosis questions, such as Alzheimer's disease versus depression. The last section presents case studies to illustrate the utility of a well-executed neuropsychological evaluation in answering frequent diagnostic questions.

When assessing older adults, the concept of normal aging versus degenerative decline must be considered. The aged are at least as varied a population as teens, college students, or middle-aged individuals. Some will change very little as they age, others a great deal, and still to others the change will be in only a few areas. Therefore, it is useful to know what cognitive functions normally decline with age as well as what impairments are common for age-related conditions like Alzheimer's disease.

Reports about the prevalence of psychiatric symptoms in the geriatric community estimate that 15 percent of adults over 65 suffer from depressive symptoms (Gatz & Hurwitcz, 1990; Hertzog et al., 1990; Snowdon, 1990). However, major depression or clinical depression occurs among the elderly with the same prevalence (1 to 4%) as in the general population (Gatz & Hurwitcz, 1990). Dementia has been reported to occur in 15 percent of the population aged 65 years or older (Green & Davis, 1993). Given the prevalence of both depression and dementia in the geriatric population of the United States, and the tendency to either misdiagnose or ignore these conditions (Bowler et al., 1994; Lamberty & Bieliauskas, 1993; Katzman, Lasker, & Berstein, 1988), the need for accurate assessment of both conditions is vital. Early diagnosis of dementia is more important than ever, with the introduction of new cognition-enhancing pharmacological agents (Samuels & Davis, 1998).

Current approaches to the psychological assessment of the elderly, like that of any other patient group, has developed over time. Some approaches (Adams, 1986; Adams & Heaton, 1985; Reitan & Davidson, 1974; Reitan & Wolfson, 1993; Russell, Neuringer, & Goldstein, 1970; Swiecinsky, 1978) have reflected trends in the field of psychology to become more medical-model, focusing on such constructs as deficits, and with diagnoses derived by use of actuarial formulas based on testing and/or symptom checklist. In the extreme adaptation of this view, the clinician need not even actually see the patient, but may review test scores that may have been gathered by a psychometric technician. With databases in a purely numerical form, diagnostic possibilities may even be generated by a computer. A distinct problem with such an objective approach is that the conclusions yield limited possibilities without accounting for any idiosyncratic responses or symptoms. Mistaking an assessment score for the behavior that it represents fails to classify the behaviors that do not fall neatly into specific definitions and diagnoses.

Other approaches (Christensen, 1979; Luria, 1966, 1973) have drawn on case-study literature and have emphasized careful clinical observation of the patient. This model is often associated with Luria and is advocated more recently by Lezak (1995), who argues that assessment should integrate qualitative behavioral descriptions, examinations of patients' writings and drawings, and attempts to elicit behaviors that reflect brain function as well as quantitative instruments. This approach also involves testing of hypotheses that guide clinical exploration, diagnosis, and formulation of treatment recommendations (Kaplan, 1988).

A thorough evaluation of the older patient must be function-based; that is, it must include assessments of level of functioning. Older patients with medical, cognitive and/or psychological problems also have functional and support issues that strongly affect their quality of life. Areas of concern to the assessor must include basic and instrumental activities of daily living, cognition, mood, psychiatric and medical diagnoses, balance and mobility, sensory intactness, continence, nutritional status, and living arrangements.

QUESTION OF NORMAL AGING

A key issue in psychological assessment of elderly patients is the need to discriminate between normal age-related intellectual changes and those changes that are clinically significant. Although many cognitive functions decline as a part of the normal aging process (Wechsler, 1997a, 1997b), the extent and pattern of the decline varies according to both the individual and the type of function being examined. Aspects of cognitive functioning that deal with well-rehearsed, overlearned activities change very little across the lifespan. Other cognitive functions, like speeded tasks, processing unfamiliar information, complex problem solving, delayed recall, mental flexibility, or perceptual manipulation tasks, do tend to decline as individuals age (Harvey & Dahlman, 1998).

The considerable individual differences in cognitive changes with aging indicate not only the difference between normal and impaired changes over time, but also differences between normal and successful changes as individuals age. Using the example of normative standards on the Logical Memory subtest from the Wechsler Memory Scale (Wechsler, 1997b), it becomes clear that those individuals who performed at high levels (99th percentile) in their youth on a variety of cognitive domains, tend to decline very little throughout their lifespan. Individuals who performed at lower levels (15th percentile) in their youth exhibit not only a decline, but a sharper decline than individuals in the upper percentile scores. The individuals at the top of the distribution consistently outperform those at the lower levels by a progressively greater extent as they become older.

The idea that normal adults who perform at higher baseline levels of intellectual function will exhibit little cognitive decline with age is supported by Rowe and Kahn's (1987, 1997) reports on successful aging. They define successful aging as including three main components: low probability of disease and disease-related disability, high cognitive and physical functioning, and active engagement with life. Continuing engagement with life has two major elements: maintenance of interpersonal relations and productive activities. Membership in a social network is an important determinant of longevity (House, Landis, & Umberson, 1988). Network membership research (Cassel, 1976; Kahn & Byosiere, 1992; Glass, Seeman, Hertzog, Kahn, & Berkman, 1995) has demonstrated that two types of supportive transactions may be prophylactic in aging: socio-emotional and instrumental. Socio-emotional transactions include expressions of respect and affection, while instrumental transactions are comprised of direct giving of services or money.

It is critical in the assessment of elderly individuals to take into account the relative nature of observed deficits; relative, that is, to the patient's own previous levels of functioning. Current functioning, in terms of engagement in life as well as presence/absence of disease and cognitive normalcy, must be viewed against the individual's overall level of previous functioning. Even a clinical interview of the patient combined with neuropsychological testing may be not be enough to fully assess what the patient may have been like prior to the onset of symptoms (Harwood, Hope, & Jacoby, 1997; Williams, 1997). For this reason, there is a trend to include caregiver ratings of patients as part of the assessment process. Examples are the Alzheimer's Disease Assessment Scale (Rosen, Mohs, & Davis, 1984) and the Informant Questionnaire on Cognitive Decline in the Elderly (IQCODE, 1989), both of which rely on the caregiver's knowledge of the patient's premorbid level of functioning to inform judgment regarding that patient's relative current decline.

OVERALL ASSESSMENT WITH REASONS FOR REFERRAL AND HISTORY-TAKING

A history is necessary to establish premorbid levels of functioning in all areas of the individual's life. It should include relevant medical, family, social, occupational, educational, cultural, and medication history, as well as substance abuse, if any, and a detailed description of the changes in functioning that precipitated the contact. It is important to establish the nature of the onset of these changes (whether abrupt or insidious), the progression of these changes (stepwise or steady, worsening versus fluctuating versus improving), and the duration of the changes.

Medical Conditions

Medical history should include a review of any diseases, psychiatric or medical, and known neurological disorders, noting any history of head trauma. Alcohol or other substance abuse as well as exposure to toxins should be reviewed. Because these and other contributing factors (e.g., HIV, diabetes, urinary tract infection) may affect cognitive functioning, a careful interview documents any illness or infection, past or present.

Medication history is an important part of the initial evaluation because drug-induced cognitive changes are among the most easily reversible. All medications, including over-the-counter formulas, could have an effect on cognition, especially in combination (Greenblatt et al., 191; Greenblatt et al., 1989). Caregivers can be helpful in providing an exhaustive list of all medications being taken by the patient, complete with dosages.

A physical examination should be performed as part of the initial assessment of the geriatric patient. This part of the screening includes a brief neurological evaluation, designed to identify lesions, vascular illness, and infection. Illnesses, such as urinary tract infection or medication toxicity, are assessed in order to rule out or address delirium. The physical examination needs to incorporate a check for signs of contusions that may indicate either accidental injuries or abuse/neglect of the patient. Suspected abuse must be taken seriously, as reports have indicated that severe abuse occurs in as many as 20 percent of families with a demented individual (Paveza et al., 1992).

Family History

A family history of dementia and other conditions (such as Huntington's disease and schizophrenia) should be established since the genetic component of these illnesses is significant (Bachman et al., 1993; Bierer et al., 1992; Goldberg et al., 1990; Mayeux et al., 1993; Neale & Oltmanns, 1980; Schellenberg et al., 1992). It is important, for example, when evaluating patients who present with psychotic symptoms (i.e., delusions and hallucinations), to weigh family and personal history of schizophrenia in making a decision about the primary disorder in the clinical picture.

Social Adaptation

The history-taking should include a review of educational level, career, and hobbies, along with socioeconomic, ethnic, and cultural background. All interviews should be conducted with both the patient and a caregiver, with the motives of the caregiver being assessed as well. The evaluation should take into account the possibility of either minimization or exaggeration of symptoms, depending on the family situation. Information on major life events and social supports, and especially recent changes are neces-

sary due to their possible contribution to the individual's performance on tests of cognitive functioning. The frequency of changes in living situations, support systems, and resources among elderly patients is common. Once an individual retires, income levels usually drop which can require a change in living situation. The onset of medical conditions could require that one moves to a residence that provides more hands-on care, easier accessibility into the building, or smaller living quarters that are easier to maintain. In some cases, such a move may be the first one the individual has had for many years. A decrease in social supports can also occur after retirement age. One loses the stimulation of interacting with coworkers, as one may be faced with losses of both friends and family members (including spouses) that results in feelings of loneliness and sadness (Parkes, 1986; Parkes & Weiss, 1983; Zisook, DeVaul, & Glick, 1982; Zisook & Schucter, 1986).

Psychiatric Conditions

Besides evaluating family history of psychiatric conditions, the psychologist must outline the patient's own psychiatric history. If the patient does have a psychiatric history, the evaluator should ascertain whether the previous episodes were reactive or not, and what types of situations have precipitated onset of symptoms in the past.

Depression

The assessment of depression in patients presenting with cognitive impairment involves some level of sophistication in order to parse out the relative contributions of affective, neurological, and other medical illness. This is critical because of treatment selection issues; if cognitive impairment is attributed to dementia, a treatable affective disorder may be overlooked. If the patient's cognitive dysfunction can be attributed with some degree of certainty to depression, the clinician has strong reasons to pursue vigorous antidepressant treatment. The failure to treat a primary depression is potentially disastrous for a patient, especially given the fairly good response of elderly depressed patients to various treatments (Benedict & Nacoste, 1990; Koenig & Blazer, 1992; Salzman & Nevis-Olsen, 1992). However, if the patient's cognitive impairment is primarily the cause of a primary dementing illness, then aggressive treatment of depressive symptoms may not substantially improve the quality of the patient's life. Overall, the issue is one of careful assessment of the clinical picture (Paquette, Ska, & Joanette, 1995).

The differential diagnosis of behavioral and cognitive disorders in older patients is made more complicated by depression, which can produce symptoms that mimic those of dementia. This is understandable given evidence from neuroimaging studies showing that patients with late-onset depression have enlarged ventricles and decreased brain density (Alexopoulos et al., 1989). Estimates of the incidence of depression in the elderly indicate that it may be slightly higher among persons 65 and older than in the younger population (Blazer, 1982; Marcopulos, 1989), and it may be the most common emotional problem among elderly patients (Hassinger et al., 1989; Thompson et al., 1987). Depressive symptoms are often precipitated by a traumatic loss, either of a family member, or by an event such as retirement or poor health. In cases such as these, the depression is reactive, and fits better with the diagnosis of Adjustment Disorder with Depressed Mood than with Major Depressive Disorder. While a chronic physical illness greatly increases the likelihood of depression in an older patient, making the diagnosis of depression in a physically sick patient is often complicated by the iatrogenic factors. Depressive symptoms may arise either from an illness itself or from medication used to treat it (Jenike, 1988; Gleenblatt et al., 1991, 1989).

Assessment of depression in the geriatric patient usually begins with a clinical interview of the patient, and ideally this is supplemented by corroborative information from a family member. The assessment must focus on objective symptoms of depression, including mood, behavior, anxiety, and vegetative symptoms such as sleep disturbance, anhedonia, anergia, and loss of appetite, as well as the subjective experiences outlined by the individual.

An instrument developed especially for use with elderly patients is the Geriatric Depression Scale (GDS: Brink et al., 1982; Yesavage et al., 1983). The GDS (Table 22.1) is a 30-item screening tool for depressive symptoms, but is not sufficient for a DSM-IV diagnosis of depression. Although it omits items tapping guilt, sexuality, and suicidality, items dealing with perceived locus of control are included which are particularly suitable for handicappped or hospitalized patients. Factor analysis of the GDS has established a major factor of dysphoria (unhappiness, disatisfaction with life, emptiness, downheartedness, worthlessness, helplessness) and minor

Table 22.1. Geriatric Depression Scale

1. Are you basically satisfied with your life?	Yes/**No**
2. Have you dropped many of your activities and interests?	**Yes**/No
3. Do you feel that your life is empty?	**Yes**/No
4. Do you often get bored?	**Yes**/No
5. Are you hopeful about the future?	Yes/**No**
6. Are you bothered by thoughts that you can't get out of your head?	**Yes**/No
7. Are you in good spirits most of the time?	Yes/**No**
8. Are you afraid that something bad is going to happen to you?	**Yes**/No
9. Do you feel happy most of the time?	Yes/**No**
10. Do you often feel helpless?	**Yes**/No
11. Do you often get restless and fidgety?	**Yes**/No
12. Do you prefer to stay at home rather than go out and doing new things?	**Yes**/No
13. Do you frequently worry about the future?	**Yes**/No
14. Do you feel that you have more problems with memory than most?	**Yes**/No
15. Do you think that it is wonderful to be alive now?	Yes/**No**
16. Do you often feel downhearted and blue?	**Yes**/No
17. Do you feel pretty worthless the way you are now?	**Yes**/No
18. Do you worry a lot about the past?	**Yes**/No
19. Do you find life very exciting?	Yes/**No**
20. Is it hard for you to get started on new projects?	**Yes**/No
21. Do you feel full of energy?	Yes/**No**
22. Do you feel that your situation is hopeless?	**Yes**/No
23. Do you think that most people are better off than you are?	**Yes**/No
24. Do you frequently get upset about little things?	**Yes**/No
25. Do you frequently feel like crying?	**Yes**/No
26. Do you have trouble concentrating?	**Yes**/No
27. Do you enjoy getting up in the morning?	Yes/**No**
28. Do you prefer to avoid social gatherings?	**Yes**/No
29. Is it easy for you to make decisions?	Yes/**No**
30. Is your mind as clear as it used to be?	Yes/**No**

Note: Brink et al. (1982); Yesavage et al. (1983).

factors of worry/dread/obsessive thought, and of apathy/withdrawal (Parmalee et al., 1989) Recommended cutoff points for the GDS are: normal, 0-9; mild depressives, 10-19; and severe depressives, 20-30. Research focused on the GDS has shown it to be helpful in discriminating between mildly demented depressed and non-depressed subjects (Snowdon & Donnelly 1986; Yesavage 1987), though the authors of the test have conceded that is less than ideally valid with more severely demented patients (Brink, 1984).

The common complaint of memory problems in an older adult may be associated with depression or other psychiatric disorders. Kiloh (1961) originally used the term *pseudodementia* to describe cases in which significant cognitive impairment seemed to resolve dramatically following treatment of a psychiatric condition. Because the cognitive impairment seen in depressed patients can be severe, some writers have proposed alternative terms such as *dementia syndrome of depression* (Folstein & McHugh, 1978) and *depression-related cognitive dysfunction* (Stoudemire, Hill, Gulley, & Morris, 1989).

Another factor in the differential diagnosis of dementia and depression is that they are often comorbid (Greenwald et al., 1989), with reports that depressive symptoms in Alzheimer's patients range from 0-86 percent, with most studies reporting rates in the 17-29 percent range (Teri & Wagner, 1992). Because patients' depressive symptoms may be unrecogized once AD has been diagnosed, the patient may suffer from unnecessary discomfort that would benefit from treatment.

Schizophrenia

Older patients with cognitive impairment may exhibit psychotic symptoms. Late-life onset of psychotic symptoms may occur separately or as a secondary feature of a primary dementing condition. A differential diagnosis may become neces-

Table 22.2. Diagnostic Criteria for Dementia of Alzhiemer's Type

A. The development of multiple cognitive deficits manifested by both

1) memory impairment (impaired ability to learn new information or to recall previously learned information)
2) one (or more) of the following cognitive disturbances:
   a) aphasia (language disturbance)
   b) apraxia (impaired ability to carry out motor activities despite intact motor function)
   c) agnosia (failure to recognize or identify objects despite intact sensory function)
   d) disturbance in executive functioning (i.e., planning, organizing, sequencing, abstracting)

B. The cognitive deficits in Criteria A1 and A2 each cause significant impairment in social or occupational functioning and represent a significant decline from a previous level of functioning.

C. The course is characterized by gradual onset and continuing cognitive decline.

D. The cognitive deficits in Criteria A1 and A2 are not due to any of the following:

1) other central nervous system conditions that cause progressive defits in memory and cognition (e.g., cerebrovascular disease, Parkinson's disease, Huntington's disease, subdural hematoma, normal-pressure hydrocephalus, brain tumor)
2) systemic conditions that are known to cause dementia (e.g., hypothyroidism, vitamin B12 or folic acid deficiency, niacin deficiency, niacin deficiency, hypercalcemia, neurosyphilis, HIV infection)
3) substance-induced conditions

E. The deficits do not occur exclusively during the course of a delirium.

F. The disturbance is not better accounted for by another AXIS I disorder (e.g., Major Depressive Disorder, Schizophrenia)

Note: Reprinted from DSM- IV, 1994.

sary to determine whether the psychotic symptoms are a component of a dementing condition like dementia of the frontal type, Huntington's disease, or Alzheimer's disease or whether the patient is experiencing a late-onset primary psychiatric condition without dementia. While new onset psychiatric conditions in individuals with no history of psychotic disturbance is unusual, some older patients do manifest such problems later in life (Harris & Jeste, 1988; Jeste, 1993). A distinction between dementias with a psychotic component and psychotic disorders without a dementing component is that late-life psychoses are typically not accompanied by profound cognitive impairments (Rosen & Zubenko, 1991).

Another differential diagnosis occurs between progressive cognitive impairments that occur as a matter of course in elderly chronic schizophrenics and cognitive impairments that indicate a dementing condition comorbid with a diagnosis of chronic schizophrenia. There is evidence that as schizophrenic patients age, their already impaired cognitive abilities worsen (Harding et al., 1987; Harvey et al., 1997). Some have argued that geriatric patients with chronic schizophrenia and severe cognitive impairment meet criteria for dementia (Arnold et al., 1994; Davidson et al., 1995).

This has led to a relatively new area of inquiry focused on cognitive impairment in chronic schizophrenic patients (Gold & Harvey, 1993). These patients have clear cognitive deficits that worsen with age and do meet criteria for dementia. Recent work has attempted to determine if there is a separate and distinct "dementia of schizophrenia," or whether Alzheimer's dementia is comorbid with schizophrenia among chronic patients and to blame for much of their late-life decline. While some brain studies have found an increased prevalence of AD-like pathology in schizophrenic brains on autopsy (Prohovnik et al., 1993), other research has concluded that neither vascular pathology nor AD can be the sole cause of the gross impairments in chronic schizophrenic patients (Arnold et al., 1993, 1994). An inspection of the types of cognitive impairments found in geriatric schizophrenics and individuals with AD reveals that the two groups have different deficits (Heaton et al.,1994; Davidson et al., 1996). Comparisons between AD patients and young versus old schizophrenic patients reveal that the performance deficit in delayed recall is much more profound in AD patients (Heaton et al., 1994) while the schizophrenics were

more impaired in constructional praxis and naming performance (Davidson et al., 1996).

COGNITIVE FUNCTIONING

Definition of Dementia

Dementia is defined in several different diagnostic systems (e.g., American Psychiatric Association, 1994; World Health Organization, 1992) as a condition marked by the loss of memory functions, deterioration in adaptive functioning from a higher level of functioning, and the presence of at least one additional sign of major cognitive deficit. The changes characteristic of dementia may be delineated into three general categories: cognitive, functional, and behavioral (Juva et al., 1994). DSM-IV criteria for dementia of the Alzheimer's type, arguably a prototypical dementia, may be found in Table 22.2. They include cognitive deficits both in memory and in one or more other areas of cognitive functioning, such as aphasia (language disturbance), apraxia (impaired ability to follow directions despite intact motor function), agnosia (failure to name objects despite intact sensory function), and disturbance in executive functions such as planning and organization. Though a progressive course is not necessarily a feature of dementia, many dementing conditions do entail gradual deterioration. Dementia is also distinguished from other conditions involving losses in one isolated area of cognitive function, such as amnesia. Dementia should be coded according to etiology when it can be identified. Patients who present for either medical or psychiatric evaluation may show evidence, either on examination or through complaints by either the patient himself or by a concerned relative, of the following symptoms:

Difficulty learning new information: Patient is repetitive, has trouble remembering recent conversations, events, and appointments; frequently misplaces objects.
Difficulty handling complex tasks: Patient has more trouble than expected following a complex train of thought, performing tasks that require many steps such as paying bills or preparing a meal.
Impaired Reasoning: Patient is unable to problem solve as effectively as in the past, shows surprising disregard for rules of social contact.
Impaired spatial ability and disorientation: Patient has trouble navigating in a car or with public transportation, organizing possessions in the household, or becomes confused trying to find his or her way around familiar settings.
Language impairment: Patient has difficulty finding words to express what he or she wants to say, and has trouble following conversations.
Behavioral abnormalities: Patient is less responsive and more passive, may be irritable or suspicious, may misinterpret behavior of others.

Establishment of Premorbid Functioning

As indicated by the discussion of normal versus impaired aging, the evaluator must establish the patient's premorbid level of cognitive functioning. Educational and occupational history will give some indication (Williams, 1997); however, we cannot dismiss the idea that some individuals can learn to cover any deficits in either of these realms. Objective measures have been shown to accurately represent premorbid functioning, including AMNART (Grober & Sliwinski, 1991; Smith et al., 1997), and certain subtests of the Wechsler Adult Intelligence Scale (Wechsler, 1987, 1997a): specifically, vocabulary, information, and block design (Albert & Moss, 1988). The most accurate estimates are accomplished through the combination of objective measures such as AMNART and vocabulary, combined with educational and occupational level (Williams, 1997).

Overview of Brief Dementia Assessment Instruments

Along with establishing the patient's premorbid level of cognitive functioning it is often useful to administer a brief dementia rating scale. The Mini-Mental State Examination (MMSE, Folstein et al., 1975) is a widely used instrument intended for use as a basic preliminary screening of cognition in geriatric patients. It contains 11 cognitive tasks and can be administered in five to 10 minutes. The exam covers orientation, memory, and attention, as well as confrontation naming, praxis, and the ability to both write a sentence spontaneously and to copy overlapping pentagons. Summing the points earned for each successfully completed task produces a score of 0 to 30, with 30 as a perfect score. Usually the score of 23 is viewed as a threshold below which

cognitive impairment is indicated (Cockrell & Folstein, 1988). However, MMSE does not measure mood, perception, nor thought content.

The Alzheimer's Disease Assessment Scale (ADAS: Mohs et al., 1983; Rosen et al, 1984) is a 21-item scale designed to assess the severity of cognitive, emotional, and behavioral symptoms in patients with Alzheimer's dementia. The cognitive portion of the scale includes both short neuropsychological tests and items rated by the examiner based on both observations of the patient's behavior and an interview with the patient's caregiver. The cognitive part of the scale assesses memory, language, and praxis, while the non-cognitive portion of the scale targets mood, vegetative functions, agitation, delusions, and hallucinations. The scale is designed to assess all core abnormalities, both cognitive and behavioral, that are typical of AD patients. Total scores on the cognitive subscale range from 0 to 70 and on the non-cognitive subscale from 0 to 50, with increasing scores indicating greater impairment.

Clinical Dementia Rating (CDR: Hughes et al., 1982) was developed in order to enable the clinician to arrive at a "global" rating of dementia based on clinical testing of cognition as well as a rating of cognitive behavior in everyday activities. The original CDR has been revised several times (Berg, 1988). It rates cognitive performance in six major categories (memory, orientation, judgment and problem solving, community affairs, home and hobbies, and personal care). Scores on these six ratings are synthesized into a single score ranging from none (0) to questionable (0.5), mild (1), moderate (2), or severe (3).The CDR does not rate apraxia, aphasia, mood, or personality change, though aphasia is measured indirectly by the assessment of both language and nonlanguage function in each cogntive category (Berg, 1984, 1988).

The Dementia Rating Scale (DRS, Mattis, 1976) is somewhat more comprehensive than some of the briefer scales such as the CDR and the MMSE. However the DRS requires a longer administration time, as does the ADAS (both take 30-45 minutes). The DRS tests orientation, registration, retention, cognitive processing, memory, and abstract reasoning. It also screens for initiation and perseveration in though & processes motor activity and for visuospatial impairments. A score on the DRS ranges from 0 to 144. Normal scores are 140 and above, and a DRS score under 100 indicates severe impairment.

The usefulness of brief dementia rating scales is clear when a gross measure of functioning is needed, either for screening or research. Brief dementia rating scales that have multiple forms, such as the ADAS are particularly useful in outcome studies, when the information sought is objective change. However, more extensive neuropsychological batteries are often indicated in both research and clinical practice. This is because of the additional information that can be gleaned regarding an individual's specific cognitive strengths and weaknesses, in order to identify predictors of the course of a particular illness and to help defferentiate between different subsets of certain psychiatric disorders. In addition, neuropsychological data is useful in the development of treatment strategies tailored to the pattern of individual strengths and weaknesses demonstrated on testing (Keefe, 1995).

GENERAL PRINCIPLES IN NEUROPSYCHOLOGICAL ASSESSMENT OF THE OLDER ADULT

Aging itself is associated with changes in virtually every function that becomes impaired in dementia. Most of the cognitive deficits seen in dementing illnesses are essentially exaggerations of normal age-related changes. Only when these deficits exceed expected levels for the patient's age and educational group, when those deficits affect adaptation, or when psychiatric symptoms (delusions, hallucinations, depression) occur, do cognitive deficits fall into the realm of a dementia. Those functions that are most resistant to decline in normal aging, such as word knowledge, are also preserved longest in most dementias.

The initial assessment of a geriatric patient is complicated by several factors including the patient's age, premorbid intelligence and previous level of functioning, educational attainment, cultural background, comorbid psychiatric illness, sensory deficits, and medical status. Thus, these factors must be considered when working with patients. Once symptoms of a possible dementia have been recognized, a thorough assessment should be initiated. This assessment consists of a detailed history, physical examination, and neuropsychological assessment of functional ability and mental status. Neuroimaging is indicated very often.

The neuropsychological assessment is an actuarial approach to the quantification of impairments reported by the patient, such as those mentioned above. The subjective complaints of the patient translate into cognitive domains targeting for eval-

uation, including perception, attention, learning and memory, verbal skills, motor skills, and executive function.

A neuropsychological test battery for the geriatric patient should begin with a thorough history, as discussed above. The primary purpose of the history is to establish a strong foundation on which to base estimates of a patient's premorbid level of functioning. There would be very different assumptions drawn about the premorbid level of functioning of individuals testing in the average range on the WAIS-III, depending on the history. For example, if that person was known to have only a tenth-grade formal education, and worked as a janitor off and on throughout his life, it would be reasonable to generate the hypothesis that that individual may have had some psychiatric problems that interfered in his ability to function at a level comporting with his intellectual capacity. On the other hand, if the individual being tested had achieved a doctoral degree and had functioned until her recent retirement as chairman of an academic department at a university, an average performance on the WAIS-III would suggest a recent intellectual deterioration.

In order to establish a baseline, or premorbid level of functioning, it is useful to estimate from performance on tests of old learning, because of the minimal effect of aging on such tests. Two often-used tests for this purpose are the vocabulary subtest of the WAIS-R and the reading subtest from either the Wide Range Achievement Test (WRAT: Jastak & Wilkinson, 1984) or the Wechsler Individual Achievement Test (WIAT: 1992). All cognitive impairments identified in geriatric patients must be referenced to age-corrected norms as well as to the patient's previous level of cognitive functioning, as noted above.

Learning and Memory: Memory impairment is a necessary but not sufficient criterion for the diagnosis of dementia. The memory domain includes the ability to retain information for a very brief period (primary or working memory), encode information for transfer to long-term storage, acquire information with repeated trial exposure (serial learning), retrieve information from memory after a delay, either with or without cues (delayed recognition versus recall), resistance to interference during the retention interval, and the ability to retrieve information that was learned long ago and bring it into current usage (long-term memory). Many of these processes are examined in the typical neuropsychological examination, with slightly different patterns of impairment depending on the etiology of the cognitive impairments in question.

Attention: Attention is a construct about which there is considerable controversy. In general, this construct refers to the ability of individuals to identify (register), focus on, process, and sustain contact with information to the extent that other operations can be performed on it. There is, however, substantial overlap between attentional processes and others that are labeled perception and memory. For example, an object must be recognized at the same time it is being perceived. Working memory, the ability to sustain information in memory while it is being processed, interacts with sustained attention. Sustained attention necessitates, in turn, that the object is maintained in working memory. Regardless of these interactions, attentional skills usually deteriorate in a broad sense in various dementing conditions. Many studies that have focused on attention target both verbal and spatial stimuli.

Perception: Perception is the ability to identify objects and other information on the basis of interpretation of sensory input. Each of the five senses is involved in this process and each may potentially be impaired in certain dementing conditions. Structured tests are used to examine each, although the majority of the attention has been focused on visual and tactile functions.

Verbal Skills: This area of functioning refers to the ability to use language adaptively, both expressively and receptively. Generally, demented patients have difficulty with generating coherent speech, with reduced complexity and content. In the assessment of verbal skills in dementia, several aspects of functioning have received considerable attention. Fluency, the ability to consistently produce words in response to a situational or task demand, has been closely examined, as has the ability to verbally identify objects (confrontation naming). In addition, vocabulary, reading ability, and other well-learned verbal skills, including the ability to use appropriate grammar and syntax, are also often affected by certain types of dementing disorders. Since deficits in receptive language ability can result in a difficulty expressing oneself, identification of specific verbal impairments are important to accomplish during the course of a neuropsychological evaluation.

Motor Skills: Motor skills can be simple, such as opening a door using a doorknob, or much more complex, for example, reproducing a complex drawing or performing a sonata on the piano. Some motor-skills tasks require an external stimulus, such

Table 22.3. Standard Neuropsychological Battery

DOMAINS	RECOMMENDED COGNITIVE MEASURES		
Attention/Concentration	Digit Span (WMS-III) Letter/Number Sequencing (WMS-III)	Spatial Span (WMS-III)	Mental Control (WMS-III) Trails A
Serial Learning	Rey Auditory Verbal Learning Test	California Verbal Learning Test	Beiber Figure Learning Test
Delayed Recall	Logical Memory II	Visual Reproduction II	
Cued Delayed Recall	Rey Auditory Verbal Learning Test	California Verbal Learning Test	
Verbal Functions Reading	WRAT-R (Reading subtest)	WIAT (Reading subtest)	NART (North Am. Reading Test)
Confrontation Naming	Boston Naming		
Overlearned Information	Vocabulary (WAIS-III)	Information (WAIS-III)	
Fluency	Animal Naming Test	Controlled Oral Word Association Test	
Executive Functions	Wisconsin Card Sorting	Trailmaking B	Ramparts/MNMNMN
Motor Speed	Finger Tapping Test		
Visuospatial Functioning/ Constructional Ability	Rey-Osterrieth Complex Figure-Copy	Hooper Visual Organization Test	Block Design (WAIS-III)
Tactile Perception	Rey Skin Writing Test		

as a model that is copied by the subject. These tasks are viewed as tapping "constructional praxis."

Executive Functioning: This domain refers to the ability to plan and organize behavior, to process more than one simultaneous stream of information, and to coordinate the application of several cognitive resources to accomodate competing demands. The clearest prototype of an executive functioning test is the Wisconsin Card Sorting Test (WCST: Heaton, 1981, Heaton et al., 1993), which requires an individual to identify basic constructs in order to solve problems, to learn through trial-and-error, to utilize working memory skills, and to make appropriate motor responses as well as to inhibit inappropriate responses. As a result, performance on this task is based on the intactness of every other aspect of cognitive functioning. Deficits in executive functioning tasks can be the result of deficits in any one of the other cognitive domains, or in their integration.

Praxis: Praxis refers to deliberate control of the motor skills employed in the execution of complex learned movements. It is usually tested by giving the patient a series of commands to follow, from simple (*pretend to comb your hair*) through facial (*whistle*) to more complex (*address a letter to yourself; pretend to knock at the door and open it*). Praxis is often subsumed under other categories such as constructional abilities, which include the ability to construct figures according to verbal directions (e.g., draw a clock) and is related to visuomotor integrative skills, such as the ability to copy a figure in two-dimensional space (e.g., Bender-Gestalt Test: Bender, 1938) or three-dimensional space (e.g., Block Design subtest of the WAIS-III).

Visuospatial Organization: Related to the above are visuointegrative skills, defined as the ability to put together pieces of a puzzle so that they form a whole (Block Design, Hooper Visual Organization Test, 1983). Disorders of praxis and visu-

ospatial organization tend to go together, though they can be seen in isolation. Thus, an individual may experience difficulty with all constructional tasks, or may be able to copy well but not be able to perform mental rotations of parts to create a whole percept. Others may be able to perform mental rotations but not be able to organize a complex drawing on paper (Rey Osterrieth Complex Figure: Osterrieth, 1944; Rey, 1941).

There are many texts available that describe the above-mentioned cognitive domains in greater detail (e.g., Lezak, 1995), and which provide a comprehensive list of neuropsychological tests (Spreen & Strauss, 1998). We refer the reader to these texts, but have also provided a table detailing a typical basic neuropsychological battery (Table 22.3).

DIFFERENTIAL DIAGNOSIS

Profile Analysis

In analyzing data from neuropsychological evaluations, it is clear that different profiles emerge for different patients, depending on the etiology of the complaint. We will limit our discussion of typical profiles here to those most commonly encountered in neuropsychology; namely, dementia and depression. The typical neuropsychological referral in geropsychological practice is generated when patients complain to their psychiatrist or internist about cognitive deterioration. While the competent clinician can confirm by using a brief screening measure such as the Mini-Mental State Exam (MMSE: Folstein, Folstein, & McHugh, 1975) that cognitive changes have occurred, more in-depth evaluation is indicated. Neuropsychological testing reveals differing patterns of scores that may be helpful in distinguishing the etiology of the cognitive disturbance, shedding light on strengths and weaknesses that will potentially affect the development of treatment plans for the patient.

Alzheimer's Disease (AD)

This disease, first reported in the early part of this century, is the most common cause of dementia. At least half of all patients with dementia over age 65 will be found to have Alzheimer's disease on postmortem evaluation (Arriagada et al., 1992), with the proportion of AD rising with increasing age (Rebok & Folstein, 1993). Some estimates have indicated that as many as 50 percent of the over-85 population meet criteria for AD (Evans et al., 1989). The course of Alzheimer's disease is about 10 years from the first sign of illness. Risk factors include age, family history of AD, little formal education, Down's Syndrome, and female gender (Cummings & Benson, 1983). On autopsy, the brain is found to have amyloid plaques and neurofibrillary tangles. These abnormalities are initially located in the medial temporal cortex and hippocampus, and eventually spread to the temporal lobe, parietal cortex, and the frontal lobe.

Alzheimer's disease is distinguished from other dementias by a deteriorating course. The first indication of an AD dementia is a profound deficit in serial learning and delayed recall. This corresponds with the very common presenting complaint of the patient: forgetfulness and difficulty learning new material. This deficit is profound, and is apparent on neuropsychological testing even in patients who may have virtually normal MMSE scores. These patients, though scoring in the mildly impaired or better range on the MMSE, will perform much more poorly than expected on tests of delayed recall (such as the Logical Memory II and Visual Reproduction subtests of the WMS) and serial learning (e.g., Rey Auditory-Verbal Learning Test: Crawford et al., 1989; California Verbal Learning Test: Delis et al., 1987). On tests of delayed recall, AD patients display virtually no retention, compared with over 85 percent retained by normal adults (Welsh et al., 1991). Opportunities to rehearse new material does not seem to benefit AD patients, nor does cueing, though these conditions allow normals to improve their performance on memory tests (Weingartner et al., 1993).

Impairments in memory and learning are followed by deficits in verbal skills. In particular, category fluency as measured by a test such as Animal Naming, has been shown to be an early hallmark of Alzheimer's type dementia (Bayles et al., 1989; Butters et al., 1987; Pasquier et al., 1995; Monsch et al., 1992) while phonemic fluency does not decline until later in the course. This has been shown to be related to a deficit in semantic knowledge that affects relationships among lower-level concepts, more so than the relationship between the concepts and their higher-order category of membership (Glosser et al., 1998).

Executive functioning deficits appear early in AD, with confrontation naming, praxis, and visuospatial deficits appearing later and progressing in

Table 22.4. Diagnostic Criteria for Vascular Dementia (formerly Multi-infarct Dementia)

A. The development of multiple cognitive deficits manifested by both

1) memory impairment (impaired ability to learn new information or to recall previously learned information)
2) one (or more) of the following cognitive disturbances:
   a) aphasia (language disturbance)
   b) apraxia (impaired ability to carry out motor activities despite intact motor function)
   c) agnosia (failure to recognize or identify objects despite intact sensory function)
   d) disturbance in executive functioning (i.e., planning, organizing, sequencing, abstracting)

B. The cognitive deficits in Criteria A1 and A2 each cause significant impairment in social or occupational functioning and represent a significant decline from a previous level of functioning.

C. Focal neurological signs and symptoms (e.g., exaggeration of deep tendon reflexes, extensor plantar response, pseudobulbar palsy, gait abnormalities, weakness of an extremity) or laboratory evidence indicative of cerebrovascular disease (e.g., multiple infarctions involving cortex and underlying white matter) that are judged to be etiologically related to the disturbance.

D. The deficits do not occur exclusively during the course of delirium.

Note: Reprinted from DSM- IV, 1994.

a linear fashion. Motor speed is impaired early in the illness, and gets progressively worse (Nebes & Brady, 1992). Attention and concentration is intact early, though orientation is often impaired initially. As the disease progresses, concentration declines gradually (Kasniak et al., 1986). Alzheimer's disease progresses steadily until performance on all tests reaches the floor (Zec, 1993). Many patients suffer from behavioral and mood disturbances, including delusions, hallucinations, agitation, and depression. These symptoms are viewed diagnositically as becoming secondary to the dementia once the dementia has been diagnosed (DSM IV, 1994). Case #1 illustrates a classic Alzheimer's disease profile, early in its course.

Vascular Dementia

Because stroke can affect any and all regions of the brain, there is no single profile for cognitive impairment caused by vascular disease. Patients who have vascular dementia as well as those with mixed dementia (AD and vascular) have been found to have deficits in memory, orientation, language, and concentration and attention with the only marked difference between the two groups the presence of gait disturbance and lesser impairments in naming and praxis among those with vascular dementia alone (Thal, Grundman, & Klauber, 1988).

Vascular dementia patients are also seen as displaying a pattern of "patchy" or irregular deficits, with clear deficits that do not follow any pattern across patient groups. A demented patient whose deficits are predominantly in the area of executive functioning would be likely to have suffered infarction in the frontal lobes, while a demented patient with aphasia may have strokes in the frontotemporal region. Subcortical vascular dementias are often characterized by profound slowing of movement (bradykinesia) and thought (bradyphrenia) such as that in the subcortical dementias asssociated with Parkinson's and Huntington's diseases.

In the assessment of vascular dementias (Table 20.4) it is also important to get a thorough history of the course of the impairment. A single stroke may lead to a focal pattern of impairment, in which memory is largely unscathed, to a diffuse pattern in which memory and other domains are affected. While AD is characterized by a persistent deteriorating course, vascular dementia is traditionally "stepwise" in its pattern of decline (Hachinski et al., 1974). There has been some indication of recovery of cognitive function in patients following treatment of vascular disease (Hershey et al., 1986) just as there is often continued mental decline related to additional infarction.

There have been recent contributions to the field that indicate that infarction is not the only vascular condition that may lead to cognitive changes. White matter disease has also been associated with a dementia syndrome that has particular impact on frontal lobe abilities such as executive function, attention, and overall intellectual level, with relative sparing of language, memory, and visual-spatial skills (Libon et al., 1997; Boone, Miller, &

Table 22.5. Depression versus Dementia

	DEPRESSION	EARLY ALZHEIMER'S DISEASE
Cognitive Function		
Memory - Recognition	Relatively intact	Impaired
- Immediate	Mild attentional difficulties	Moderately to severely impaired
- Delayed Recall	Near normal rate	Little to no retention
Learning - New Information	Intact	Severely impaired
- Complex Tasks	Distractible	Loses train of thought easily
- Reasoning	Intact	Impaired
Attention	Mild difficulties	Intact
Perception	Normal	Impaired
Language Skills	Normal expressive and receptive functioning; reduced verbal fluency	Decline in expressive and receptive functioning
Executive Functioning	Intact	Mild impairments evident early especially parrallel processing
Praxis	Slowed	Intact
Visuospatial	Normal	Impaired
Course of Illness		
Onset	Rapid	Insidious
Awareness of Impairment	Intact, complaints of memory problems	Impaired
Duration	Few weeks to months; reversible with treatment	Deteriorating course over approximately 10 years
Mood		
Symptoms	Stable level of depression, apathy, and withdrawal	Labile - between normal and withdrawn
Somatic		
Symptoms	Vegetative signs: insomnia, eating disturbances, minor physical complaints	Some sleep disturbances

Lesser, 1993). Other studies point to additional types of vascular disease, such as atrophy, gliosis and spongiosis (Gustafson, 1987), white and gray matter changes (Libon et al., 1997; Gydesen et al., 1987), atrophy and gliosis (Neary et al., 1990), all of which may lead to cognitive impairment.

Depression versus Dementia

The differential diagnosis between depressed and demented patients may be made using a combination of a neuropsychological evaluation and a thorough mood assessment using one of the measures mentioned above. The neuropsychological evaluation will show mild differences between depressed and normal patients, on performance in cued and uncued recall and delayed recognition memory, as well as verbal learning (King et al., 1998). Substantial differences have been shown to exist between demented and normal patients (Christensen et al., 1997). The differences between the demented and the depressed patient are primarily on delayed recall relative to immediate recall; the depressed patient will usually perform lackadaisically on encoding tasks, resulting in a relatively low score on imme-

diate recall tests. Delayed recall will likely be poor as well, but with the difference that there will not necessarily be much loss between the raw scores for encoding and retrieval. Demented patients will encode adequately, but will demonstrate significant if not complete forgetting following a delay. Recognition, or cued recall, generally can discriminate as well between the depressed and demented patient. Table 22.5 summarizes the differences in neuropsychological test data between these patients.

SPECIAL PROBLEMS IN GERIATRIC ASSESSMENTS

While it is usually interesting to do a follow-up assessment on psychiatric patients to evaluate changes in functioning since a previous evaluation, reexamination is a virtual necessity in geropsychology. There are several times when retesting is especially important. Patients with little formal education and/or a borderline or lower IQ may initially perform at such low levels on neuropsychological testing that it is extremely difficult to distinguish a dementia from baseline performance. In such cases, retesting is necessary in order to establish presence or absence of a deteriorating course. Even though patients may be performing at an extremely low percentile level when first tested, it is possible to discern changes over time in the raw scores. If the raw scores drop noticeably and consistently across the different domains tested, even while changes in percentile level are not discernable because of an extremely low baseline, it is possible to conclude that there has been a global deterioration over time. A deteriorating course is a hallmark of Alzheimer's disease; in order to establish the existence of such a course in patients with extremely low baseline performance, reexamination is necessary.

At the other end of the intelligence spectrum, and presenting another diagnostic conundrum, are those elderly patients who have extemely high levels of intellectual functioning. Notwithstanding their advanced age, individuals with IQ's in the upper reaches of the scale (Superior and Very Superior ranges) often present themselves for evaluation because of subjective complaints of memory loss. These individuals, accustomed to enjoying great mental acuity, may be particularly sensitive to any diminishment of their abilities. They will often hit the ceiling on a gross screening measure such as the Mini-Mental Status examination, achieving perfect or near-perfect scores. Neuropsychological assessment will be necessary to determine whether these patients have suffered significant cognitive losses, or whether they are performing at expected levels. While the profiles of two such patients may be similar in that their premorbid levels of functioning are in the Very Superior range, the delayed recall performance will differentiate a patient with early dementia from another who has suffered some changes in functioning but whose memory performance is still within the Very Superior range for his age. The patient who is not demented will have a relatively lower raw score than he may have achieved on previous testing, but his age-scaled score remains essentially the same. Another patient with a premorbid IQ of 150 with early AD may have delayed recall scaled scores as high as 50th percentile: this is still in the normal range but represents a significant deficit for her.

Regardless of whether the patient in question is at one or the other end of the IQ curve, the extremity of their scores dictates that retesting will play a critical role in the assessment process. The first testing is necessary to establish a baseline, and the second, usually one year later, will be useful in determining course.

CASE STUDIES

Case #1—Alzheimer's Disease

Referral Information

This first case represents a classic Alzheimer's disease neuropsychological profile (see Table 22.6). The patient, Mrs. K. was an 87-year-old, widowed, white, female who was referred as part of a diagnostic work-up for dementia. At the time of the testing she was living alone in an apartment, with no familial support living in her city. She had suffered two major losses in the past 10 years and there was some concern that her recent cognitive impairments were related to feelings of grief and depression. The first loss was the death of her granddaughter, a teenager killed in a car accident ten years prior to the testing. The second was the death of her husband five years later. Since becoming a widow, she had lost 20 pounds, gradually. She was not particularly active anymore, though she had traveled and played bridge when her husband was alive. She had worked until her late 60s

Table 22.6. Case #1—Alzheimer's Disease (Mrs. K.)

Premorbid and Current Overall Functional Level

Wechsler Adult Intelligence Scale-Third Edition (WAIS-III)

	RAW SCORE	AGE SCALED SCORE	PERCENTILE	CLASSIFICATION
Vocabulary	49	10	50	Average
Comprehension	17	10	50	Average
Information	11	6	9	Low Average

Orientation, Attention, Concentration, Distractibility

Wechsler Memory Scale—Third Edition (WMS-R)	*Raw Score*
Information and Orientation	10/14 (missed current and past president, date, and place)
Mental Control	4/6

	TIME	ERRORS	PERCENTILE
Trailmaking Test A	74	0	>50

Memory Functioning

	Raw Score	Age scaled score	Percentile
Wechsler Memory Scale—Third Edition (WMS-R)			
Verbal Paired Associates I	1	6	9
Verbal Paired Associates II	1	8	25
Logical Memory I	28	10	50
Logical Memory II	3	6	9
Wechsler Memory Scale—Revised (WMS-R)			
Visual Reproduction I	12	9	
Visual Reproduction II	3	9	

Rey Auditory Verbal Learning Test

TRIAL	I	II	III	IV	V	B	VI
Number Recalled	6	4	5	5	7	5	0
Normative Mean	4.0	6.0	7.4	7.9	9.1	3.1	6.2
Standard Dev.	1.5	1.8	2.2	2.4	2.3	1.4	2.6

	SCALED SCORE	PERCENTILE
Learning Over Trials	6	9
Short Term Percent Retention	2	<1

Verbal Functioning (also see Vocabulary, Information, and Comprension)

	RAW SCORE	NORM MEAN	NORM SD
Boston Naming Test (CERAD Abbreviation)	15/15		
Animal Naming Test	9	15.09	4.25
Controlled Word Association Test (FAS)	18	35.20	11.90

Visuospatial Functioning

Rey Osterrieth Complex Figure Drawing	mildly impaired (Copy)

Praxis

Western Apraxia Examination	Upper Limb	15	Facial	15
	Instrumental	14	Complex	13

Executive Functioning

Wisconsin Card Sorting Test (WCST)	achieved 0 categories in 56 trials, severely impaired
Trailmaking Tests B	failed

doing office and sales work. She had no history of alcohol nor substance abuse, and had been consuming less alcohol than in years past. She had given up smoking many years ago. Her medical history was significant for ulcerated colitis.

During the testing she presented as a woman who appeared slightly younger than her stated age. Her grooming and hygiene were intact. She arrived to the appointment on time and was cooperative with all test demands. Her speech was normal in rate and volume, though somewhat sparse in content. Her affect was slightly constricted and mood was neutral.

Tests Administered. Wechsler Adult Intelligence Scale, Revised (WAIS-R), selected subtests; Wechsler Memory Scale, Revised (WMS-R) and WMS-III, selected subtest; Cancellation Test; Trailmaking Test, Parts A and B; Controlled Oral Word Association Test; Boston Naming Test (CERAD abbreviation); Animal Naming Test; Rey Osterrieth Complex Figure Drawing, copy; Rey Auditory Verbal Learning Test; Wisconsin Card Sorting Test; Western Apraxia Examination.

Areas of Cognitive Functioning

Memory functioning. Mrs. K.'s performance on memory tests indicated marked impairment relative to her estimated premorbid ability as indicated by her Vocabulary and Comprehension scores. Her ability to retrieve previously learned information was mildly impaired; she was beginning to have trouble retrieving overlearned information such as the current and past presidents. She was able to benefit very little from repeated trial learning, as evidenced by a lower than expected (9th percentile) learning over trials score on a five-trial list learning task. On a test of story recall, her delayed recall abilities placed her at the 9th percentile for her age group. Her visual recall was poor as well, with 75 percent forgetting of encoded information. She performed almost as poorly on recognition tasks as on free recall tasks, another indicator that there was a primary memory impairment and not merely an encoding problem.

Attention and Orientation. Mrs. K. demonstrated impaired orientation to date and place, though she was oriented to person. However, her attentional abilities are within normal limits (WNL).

Verbal functioning. Performance on verbal functioning tests was quite deficient, following the pattern of impairment often seen in early AD. Performance on verbal fluency tasks was more than one standard deviation below normal, both in response to phonemic and to semantic prompts. Confrontation naming was within normal range.

Praxis. Mrs. K. made several errors on instrumental and complex tasks, though upper limb and facial tasks were performed normally to command. This demonstrated a subtle deterioration in higher level praxic functioning.

Executive Functioning. On the Wisconsin Card Sorting Test, her performance was significantly impaired. She made many errors, both perseverative and non-perseverative, and was not able to generate any concepts for sorting cards. She had trouble retaining a set when the examiner suggested one. This performance suggested that the patient had an impaired ability to form and retain concepts. Difficulty with parallel processing was also seen on Trailmaking B, which she failed due to trouble alternating set.

Depression Screening. Mrs. K. did not meet criteria for major depression, though she was somewhat dysthymic. Her score on the Geriatric Depression Scale is 7, which was below the cutoff of 10 for mild depression.

Summary. Consideration of the assessment results suggested that Mrs. K., a woman whose estimated premorbid IQ was approximately 100, or in the Average range, of intellectual functioning, had some abnormal cognitive impairment. There is no evidence of formal thought disorder nor did this patient meet criteria for a major depressive disorder. She reported impairment in memory and cognitive functioning, and there was indeed clear indication on the testing that some deficits do exist relative to her estimated premorbid ability. These impairments were primarily in memory, orientation, executive functioning, and verbal fluency. This pattern of results indicate the likely early presentation of an Alzheimer's disease dementing process.

Table 22.7. Case #2 Vascular Dementia (Mrs. P.)

Premorbid and Current Overall Functional Level

WECHSLER ADULT INTELLIGENCE SCALE	1996 PERCENTILE (WAIS-R)	1998 PERCENTILE (WAIS-III)
Vocabulary	37	N/A
Information	5	N/A
Comprehension	25	16
Block Design	25	5

Orientation, Attention, Concentration, Distractibility

WECHSLER MEMORY SCALE	1996 (WMS-R)	1998 (WMS-III)
Information and Orientation	10/14	8/14
Mental Control	WNL	WNL
Digit Span	76th percentile	N/A
Trailmaking Test A	20th percentile	>90th percentile
Cancellation Test	severely impaired	severely impaired

WECHSLER ADULT INTELLIGENCE SCALE	(WAIS-R)	(WAIS-III)
Picture Completion	5th percentile	2nd percentile

MEMORY FUNCTIONING	1996 (WMS-R)	1998 *(WMS-III)*
Wechsler Memory Scale		
Logical Memory I	<1(raw score 6)	1(raw score 2)
Logical Memory II	1 (raw score 1)	4 (raw score 0)
Word List Recall	50% forgetting	

Current Memory Profile (Percentiles)

Auditory Immediate	Visual Immediate	Immediate Memory	Auditory Delayed	Visual Delayed	Aud Recog. Delayed	General Memory
.3	.2	.1	1	10	9	2

VERBAL FUNCTIONING (RAW SCORES)	1996	1998
Animal Naming	9	18
COWAT	33	33
Boston Naming	8/15	9/15

Visuospatial Functioning

Rey-Osterrieth Copy	Impaired	WNL

PRAXIS	1996	1998
Western Apraxia Examination		
Up. Limb	15	13
Facial	8	15
Instrumental	8	11
Complex	9	12

Executive Functioning

(WAIS-III) Similarities	N/A	16th percentile
Trails B	failed	failed
Wisconsin Card Sorting Test (WCST)	0 categories	0 categories

Table 22.8. Case #3—Depression with Cognitive Impairment (Mr. T.)

General Functioning

Wechsler Memory Scale-Third Edition (WMS-III)

Information and Orientation	10/14
Mental Control	8/25

Wechsler Adult Intelligence Scale-Third Edition (WAIS-III)

	Age scaled score	Percentile		Age scaled score	Percentile
Information	15	95	Picture Completion	10	91
Vocabulary	15	95	Block Design	10	63
Comprehension	15	95	Matrix Reasoning	15	95
Similarities	11	63			

Attention and Concentration	*Time*	*Percentile*	*Errors*
Trailmaking A (percentile)	***75***	***25-10***	***0***

	Age scaled score	*Percentile*
WMS-III Digit Span	10	50

Memory

WMS-III	*Age scaled scores*	*Percentiles*
Logical Memory I	2	0.4
Logical Memory II	6	9
Visual Reproduction I	2	0.4
Visual Reproduction II	7	16
Visual Reproduction Recognition	7	16

Word Lists I	*RawScore*		*RawScore*	*Scores*	*Age scaled scores*
Trial I	1	Trial B	1	Total recall	5
Trial II	3	Trial V	2	Learning slope	4
Trial III	5			Contrast 1	6
Trial IV	4			Contrast 2	12

Verbal Functioning (see also *WAIS-III* Information, Vocabulary, Comprehension)

	Raw Score	*Norm Mean*	*Norm SD*
Animal Naming Test	13	17.07	4.93
Controlled Oral Word Association Test	46	39.08	14.17
Boston Naming Test	52/60	51.5	7

Visuospatial Functioning (see also *WAIS-III* Picture Completion, Block Design)
Rey Osterrieth Complex Figure Test 23/36 (Mean = 32.90, SD = 2.69)

Sensation	*Right Hand correct*	*Left Hand correct*
Rey Skin Writing Test:	3 letters, 2 numbers	3 letters, 2 numbers

Executive Functioning (see also *WAIS-III* Similarities subtest, *Rey Osterreith Complex Figure Test*)

	Time	*Percentile*	*Errors*
Trailmaking B	476s	10-25	5
Wisconsin Card Sorting Test (WCST)	discontinued		

Personality Functioning

Geriatric Depression Scale	11 (mild depression)

Case #2 Vascular Dementia

Referral Information

The second case will present a neuropsychological profile representative of a vascular dementia (see Table 22.7). The patient, Mrs. P. was a 76-year-old, Irish-American female who was referred for a neuropsychological re-evaluation in order to follow up on a previous examination done a year earlier. That evaluation had revealed serious cognitive impairments consistent with dementia. The second testing was intended to monitor changes in cognitive functioning over the past year as well as to clarify etiology.

The patient has had a history of major depressive illness, with several hospitalizations. She has also experienced TIAs in the past. Her depression has been well controlled on medication. During the testing sessions she appeared her stated age, with grooming and hygiene very intact. She arrived on time accompanied by two family members for both testing sessions. She was cooperative with all test demands. Her speech was normal in rate and volume, though somewhat sparse in content. Her affect was slightly constricted and mood was neutral.

Tests Administered. Wechsler Adult Intelligence Scale, Third Edition (WAIS-III), selected subtests; Wechsler Memory Scale, Revised Third Edition (WMS-III), selected subtests; Cancellation Test; Trailmaking Test, Parts A and B; Controlled Oral Word Association Test; Boston Naming Test (CERAD abbreviation); Animal Naming Test; Rey Osterrieth Complex Figure Drawing, copy; Rey Auditory Verbal Learning Test; Wisconsin Card Sorting Test; Western Apraxia Examination.

Test Findings. Mrs. P.'s premorbid intelligence was estimated to have been in the Average range, or 25-37th percentile, with a Verbal IQ of approximately 90-95. The second testing revealed no significant deterioration in functioning relative to levels from the year before, nor was there any notable improvement. It should be noted that while different versions of both the Wechsler Adult Intelligence and Memory Scales for each of the testing evaluations, accurate comparisons can be made based on the high correlations on all subtests for both versions (Wechsler Adult Intelligence Scale-III and Wechsler Memory Scale-III Manual, 1997). The changes in the revised versions to the third editions do not impact comparisons based on percentile ranks. Mrs. P. continued to experience major cognitive deficits in the areas of orientation, attention and concentration, memory functioning, and executive processes. These deficits were consistent with a dementia of vascular etiology that did not have a deteriorating course as of the second assessment.

Cognitive decline due to depression could be ruled out because her cognitive impairments persisted despite remission of her depression. Based on the pattern of her scores on memory tests, particularly a marked superiority of cued over free recall, and in the absence of a clear deteriorating course, typical Alzheimer's disease is clearly not the only cause of this patient's cognitive impairment.

Case #3 Depression with Cognitive Impairments

Referral Information

The final case represents a neuropsychological profile consistent with a diagnosis of depression that includes marked cognitive impairment (see Table 22.8). The patient, Mr. T. was a 91-year-old, widowed, white, male who lived alone in a residence for the elderly.

Mr. T. referred himself for evaluation, complaining of deterioration in hearing in his (right) ear and dermatitis on his lower arms. He also reported minor difficulties in adjusting to living at his new residence. He found some of the rules "silly," such as not being allowed to take newspapers into the dining room. His medical history included a cataract in his right eye, arthritis in his left hand, and two minor operations. He denied any prior psychiatric history, any use of alcohol beyond social drinking, and any history of head trauma.

Mr. T. was the youngest of three children born into an intact family. He reported completing high school, an undergraduate degree, and a Master's degree. In addition, Mr. T. described having served in the army as an adjutant during World War II. After leaving the service he taught junior high school and was married and had one child. He retired early to travel abroad. His wife had died three years prior to the testing.

A CT scan conducted just prior to the neuropsychological evaluation revealed no evidence of abnormal enhancing mass lesions. There was no evidence of significant territorial infarcts. At the

time of the testing, Mr. T. appeared younger than his stated age. He arrived for the testing session on time. Though he was able to walk unaided, he was accompanied by his daughter. He was neatly groomed and well dressed for the testing session. On interview, he was alert, well-related, cooperative, pleasant, and made good eye contact. His affect and mood were euthymic. In answering test questions, Mr. T. was able to follow even lengthy instructions, although he did have mild difficulties with hearing that required repetition of instructions. He also gave indication of mild anxiety during testing, for instance, he stiffened slightly as he sat in his chair and at times commented that his performance must be quite poor. Mr. T.'s speech was normal in rate, volume, and fluency. His verbal answers were goal-directed, logical, and coherent. On nonverbal tasks, he made drawings and manipulated objects with only mild difficulty due to his arthritis.

Tests Administered. Weschler Adult Intelligence scale, Third Edition (WAIS-III), selected subtests; Weschler Memory Scale, Third Edition (WMS-III), selected subtest; Controlled Oral Word Association Test; Animal Naming Test; Boston Naming Test; Rey Skin Writing Test; Rey Osterrieth Complex Figure Drawing; Trailmaking Test, parts A & B; Wisconsin Card Sorting Test

Areas of Cognitive Functioning

Orientation/Attention/Concentration: Mr. T. was oriented to person and place, but severely disoriented to day, month, date, and year. His brief passive auditory attention and mental control were in the Average range (at the 50th and 25th percentile levels, respectively).

Memory Functioning. Memory functioning was impaired. Mr. T.'s immediate recall for verbal contextual material (stories), noncontextual verbal material (word lists), and nonverbal visual material (geometric shapes) was in the Extremely Low range (all 2nd percentile or below). Such scores are well below his estimated level of premorbid functioning. However, delayed recall for stories and geometric shapes were both in the Low Average range (at the 9th and 16th percentile level, respectively). Although these scores are below estimated premorbid functioning, they represent an improvement over immediate recall. In addition, on word lists, although serial learning (increase in recall with repetition) was poorer than expected, there was minimal forgetting following a distractor. The global and severe nature of deficits in memory can be consistent with a diagnosis of mild dementia. However, the pattern of memory deficits also points to difficulties with encoding of information, which raises the possibility of depression impinging on memory performance.

Verbal Functioning. Verbal functioning was intact. Mr. T.'s coherent expression of word meanings, fund of information, and verbal reasoning in answering questions about social judgments were all in the Superior range (at the 95th percentile level). Meanwhile, verbal fluency, both in response to a phonemic cue (letter fluency) and semantic cue (category fluency) as well as confrontation naming of objects were all within the norm for persons of his age and level of education.

Sensation. There was no evidence for lateralization in skin sensation with equivalent performance on left and right sides.

Psychomotor Speed. Psychomotor speed was slightly slower than the norm.

Perceptual-Motor Functioning. Perceptual-motor functioning was intact. Mr. T.'s perception of salient environmental details was in the Average range (at the 50th percentile level), consistent with estimated premorbid level of cognitive functioning. Similarly, higher-level synthesis and abstraction was in the Average range.

Executive Functioning. Executive functioning was impaired. Mr. T.'s performance on a task of forming simple and superordinate concepts in response to minimal verbal feedback was well below levels to be expected from his level of education. His parallel processing in sequencing an irregular array of letters and numbers was in the Borderline range (below the 10th percentile level). However, this result may in part be explained by psychomotor slowing. In addition, Mr. T. demonstrated limited planning and ability to use the overall gestalt to structure his copying of a complex geometric shape. However, executive functioning in the verbal realm appeared intact, with verbal abstraction performance in the High Average range (at the 75th percentile level).

Depression Screening. On a standard self-report measure, Mr. T. scored in the mildly depressed range. He acknowledged feelings of sadness, feeling the urge to cry, feelings of restlessness, fear that something bad would happen, decreased

energy, decline in activities and interests, and increased difficulty with memory.

Summary. The assessment evaluation estimated Mr. T.'s premorbid level of cognitive functioning to have been in the Superior range. There was evidence on testing of impairments in memory and executive functioning relative to premorbid cognitive functioning. However, brief passive attention and concentration, verbal functioning, and perceptual-motor functioning were consistent with estimated premorbid cognitive functioning. Meanwhile, on a mood assessment scale, Mr. T. acknowledged mild feelings of depression, particularly in connection with the death of his wife and his increasing physical frailty. The overall clinical picture is most consistent with "pseudodementia," cognitive deficits secondary to depression.

REFERENCES

Adams, K. M. (1986). Concepts and methods in the design of automata for neuropsychological test interpretation. In S. B. Filshov & T. J. Boll (Eds.), *Handbook of clinical neuropsychology* (Vol. 2). New York: John Wiley & Sons.

Adams, K. M., & Heaton, R. K. (1985). Automated interpretation of neuropsychological test data. *Journal of Consulting and Clinical Psychology, 53,* 790–802.

Albert, M. S., & Moss, M. B. (1988). *Geriatric neuropsychology.* New York: Guilford.

Alexopoulos, G. S., Young, R. C., Abrams, R. C., et al. (1989). Chronicity and relapse in geriatric depression. *Biological Psychiatry, 26,* 551–564.

American Psychiatric Association (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: APA.

Arnold, S. E., Franz, B. R., & Trojanowski, J. Q. (1993). Lack of neuropathological findings in elderly patients with schizophrenia. *Neuroscience Abstracts, 19,* 349–350.

Arnold, S. E., Franz, B. R., & Trojanowski, J. Q. (1994). Elderly patients with schizophrenia exhibit infrequent neurodegenerative lesions. *Neurobiology of Aging, 15,* 299–303.

Arriagada, P. V., Marzloff, K., & Hyman, B. T. (1992). Distribution of Alzheimer-type pathologic changes in non-demented elderly individuals matches the patterns in Alzheimer's disease. *Neurology, 42,* 1681–1688.

Bachman, D. L., Wolf, P. A., Linn, R. T., Knoefel, J. E., Cobb, J. L., Belanger, A. J., White, L. R., & D'Agostino, R. B. (1993). Incidence of dementia and probable Alzheimer's disease in a general population: The Framingham Study. *Neurology, 43,* 515–519.

Bayles, K. A., Salmon, D. P., Tomoeda, C. K., Jacobs, D., Caffrey, J. T., Kaszniak, A. W., & Troster, A. I. (1989). Semantic and letter category naming in Alzheimer's patients: A predictable difference. *Developmental Neuropsychology, 5,* 335–347.

Bender, L. (1938). *Instructions for the use of the Bender Visual Motor Gestalt Test.* American Orthopsychiatric Association.

Benedict, K. B., & Nacoste, D. B. (1990). Dementia and depression: A framework for addressing difficulties in differential diagnosis. *Clinical Psychology Review, 10,* 513–537.

Berg, L. (1984). Clinical dementia rating (letter). *British Journal Psychiatry, 145,* 339.

Berg, L. (1988). Mild senile dementia of the Alzheimer's type: Diagnostic criteria and natural history. *Mount Sinai Journal of Medicine, 55,* 87–96.

Bierer, L. M., Silverman, J. M., Mohs, R. C., Haroutunian, V., Li, G., Purohit, D., Brietner, J. C. S., Perl, D. P., & Davis, K. L. (1992). Morbid risk to first degree relatives of neuropathologically confirmed cases of Alzheimer's disease. *Dementia, 3,* 134–139.

Blazer, D. (1982). The epidemiology of late life depression. *Journal of the American Geriatrics Society, 30,* 587–592.

Boone, K. B., Miller, B. L., & Lesser, I. M. (1993). Frontal lobe cognitive functions in aging: Methodological considerations. *Dementia, 4,* 232–236.

Bowler, C., Boyle, A., Branford, M. Cooper, S. A., Harper, R., Lindesay, J. (1994). Detection of psychiatric disorders in elderly medical inpatients. *Age and Ageing, 23 (4),* 307–311.

Brink, T. L. (1984). Limitations of the GDS in cases of pseudodementia. *Clinical Gerontology, 2,* 60–61.

Brink T. L., Yesavage, J. A., Lum, O., Heersema, P. H., Adey M., & Rose, T.S. (1982). *Clinical Gerontologist, 1,* 37–43.

Butters, N., Granholm, E., Salmon, D. P., Grant, I., & Wolfe, J. (1987). Episodic and semantic memory: A comparison of amnesic and demented patients. *Journal of Clinical Neuropsychology, 9,* 479–497.

Cassel, J. (1976). The contribution of social environment to host resistance: The fourth Wade Hampton Frost lecture. *American Journal of Epidemiology, 104,* 107–123.

Christensen, A. L. (1979). *Luria's neuropsychological investigation* (2nd ed.). Copenhagen: Munksgaard.

Christensen, H., Griffiths, K., Mackinnon, & Jacomb, P. (1997). A qualitative review of cognitive deficits in depression and Alzheimer-type dementia. *Jour-*

nal of the International Neuropsychological Society, 3, 631–651.

Cockrell, J. R., & Folstein, M. F. (1988). Mini-Mental State Examination (MMSE). *Psychopharmacology Bulletin, 24,* 689–692.

Crawford, J. R., Stewart, L. E., & Moore, J. W. (1989). Demonstration of savings on the AVLT and development of a parallel form. *Journal of Clinical and Experimental Neuropsychology, 11,* 975–981.

Cummings, J. L., & Benson, D. F. (1983). *Dementia: A clinical approach* (2nd ed.). Boston: Butterworth's-Heinemann.

Dahlman, K. L., Davidson, M., & Harvey, P. (1996, April). *Cognitive functioning in late-life schizophrenia: A comparison of elderly schizophrenic patients with Alzheimer's disease.* Paper presented at The Challenge of the Dementias Conference, *The Lancet,* Edinburgh, Scotland.

Davidson, M., Harvey, P. D., Powchik, P., et al. (1995). Severity of symptoms in geriatric schizophrenic patients. *American Journal of Psychiatry, 152,* 197–207.

Davidson, M., Harvey, P. D., Welsh, K., Powchik, P., Putnam, K., & Mohs, R. C. (1996). Characterization of the cognitive impairment of old-age schizophrenia: A comparison to patients with Alzheimer's disease. *American Journal of Psychiatry, 153,* 1274–1279.

Delis, D. C., Kramer, J. H., Kaplan, E., & Ober, B. A. (1987). *California Verbal Learning Test manual.* San Antonio, TX: The Psychological Corporation.

Evans, D. A., Funkenstein, H. H., Albert, M. S., Scherr, P. A., Cook, N. R., Chown, M. J., Herbert, L. E., Hennekens, C. H., & Taylor, J. O. (1989). Prevalence of Alzheimer's disease in a community population of older persons. *Journal of the American Medical Association, 262,* 2551–2556.

Folstein M. F., Folstein S. E., & McHugh P. R. (1975). "Mini-mental state." *Journal of Psychiatric Research, 12,* 189–198.

Folstein M. F. & McHugh P. R. (1978). Dementia syndrome of depression. *Aging, 7,* 87–93.

Fromm, D., Holland, A. L., Nebes, R. D., & Oakley, M. A. (1991). A longitudinal study of word-reading ability in Alzheimer's disease: Evidence from the National Adult Reading Test. *Cortex, 27,* 367–376.

Gatz, M., & Hurwitcz, M. L. (1990). Are old people more depressed? Cross-sectional data on Center for Epidemiological Studies Depression Scale factors. *Psychology of Aging, 5,* 284-290.

Gold J. M., & Harvey P. D. (1993). Cognitive deficits in schizophrenia. *Psychiatric Clinics of North America, 16*(2), 295–12.

Glass, T. A., Seeman, T. E., Hertzog, A. R., Kahn, R. L., & Berkman, L. F. (1995). Changes in productivity in late adulthood: MacArthur Studies of Successful Aging. *Journal of Gerontology: Social Sciences, 50B,* S65–S76.

Glosser, G., Friedman, R. B., Grugan, P. K., Lee, J. H., & Grossman, M. (1998). Lexical semantic and associative priming in Alzheimer's disease. *Neuropsychology, 12*(2), 218–224.

Goldberg, T. E., Ragland, J. D., Torrey, E. F. et al. (1990). Neuropsychological assessment of monozygotic twins discordant for schizophrenia. *Archives of General Psychiatry, 47,* 1066.

Green, C. R., & Davis, K. L. (1993). Clinical assessment of Alzheimer's-type dementia and related disorders. *Human Psychopharmacology, 4,* 53–71.

Greenblatt, D. J., Harmatz, J. S., Shapiro, L., Engelhardt, N., Gouthro, T. A., & Shader, R. I. (1991). Sensitivity to triazolam in the elderly. *New England Journal of Medicine, 324,* 1691–1698.

Greenblatt, D. J., Shader, R. I., & Harmatz, J. S. (1989). Implications of altered drug disposition in the elderly: Studies of benzodiazepines. *Journal of Clinical Pharmocology, 29,* 866–872.

Greenwald, B. S., Kramer-Ginzberg, E. Marin, D. B., Laitman, L. B., Herman, C. K., Mohs, R. C., & Davis, K. L. (1989). Dementia with coexisting depression. *American Journal of Psychiatry, 146,* 1472–1478.

Grober, E., & Sliwinski, M. (1991). Development and validation of a model for estimating premorbid verbal intelligence in the elderly. *Journal of Clinical and Experimental Neuropsychology, 13,* 933–949.

Gustafson, L. (1987). Frontal lobe degeneration of the non-Alzheimer type, II: Clinical picture and differential diagnosis. *Archives of Gerontology and Geriatrics, 6,* 209–24.

Gydesen, S., Hagen, S., Klinken, L., et al. (1987). Neuropsychiatric studies in a family with presenile dementia different from Alzheimer's and Pick's disease. *Acta Psychiatry Scandanavia, 76,* 276–284.

Hachinski, V. C., Lassen, N. A., & Marshall, J. (1974). Multi-infarct dementia: A cause of mental deterioration in the elderly. *Lancet, 2,* 207–209.

Harding, C. M., Brooks, G. W., Ashikaga, T., Stauss, J. S., & Breier, A. (1987). The Vermont longitudinal study of persons with severe mental illness: II. Long term outcome of subjects who retrospectively met DSM-III criteria for schizophrenia. *American Journal of Psychiatry, 144,* 727–735.

Harris, M. J., & Jeste, D. V. (1988). Late-onset schizophrenia: An overview. *Schizophrenia Bulletin, 14,* 39–55.

Harvey, P. D., & Dahlman, K. L. (1998). Neuropsychological evaluation of dementia. In A. Chalev (Ed.) *Neuropsychological assessment of neuropsychiatric disorders*. Washington, DC: American Psychiatric Press.

Harvey, P.D., Powchik, P., Mohs, R.C., & Davidson, M. (1995). Memory functions in geriatric chronic schizophrenic patients: A neuropsychological study. *Journal of Neuropsychiatry and Clinical Neurosciences, 7,* 207–212.

Harvey, P. D., Lombardi, J., Leibman, M., Parella, M. White, L., Powchik, P., Mohs, R. C., & Davidson, M. (1997). Verbal fluency deficits in geriatric and nongeriatric chronic schizophrenic patients. *Journal of Neuropsychiatry and Clinical Neurosciences, 9(4),* 584–590.

Harvey, P. D., Powchik, P., Mohs, R. C., & Davidson, M. (1995). Memory functions in geriatric chronic schizophrenic patients: a neuropsychological study. *Journal of Neuropsychiatry and Clinical Neurosciences, 7*(2) 207–212.

Harvey, P. D., White, L., Parrella, M., Putnam, K. M., Kincaid, M.M., Powchik, P., Mohs, R. C., & Davidson, M. (1995). The longitudinal stability of cognitive impairment in schizophrenia. Mini-mental state score at one- and two-year follow-ups in geriatric in-patients. *British Journal of Psychiatry, 166*(5), 630–633.

Harvey, P. D., Lombardi, J. L., Leibman, M., White, L., Parrella, M., Powchik, P., Mohs, R. C., & Davidson, M. (1996). Performance of chronic schizophrenic patients on cognitive neuropsychological measures sensitive to dementia. *International Journal of Geriatric Psychiatry, 11,* 621–627.

Harwood, D. M. L., Hope, T., & Jacoby, R. (1997a). Cognitive impairment in medical inpatients. I: Screening for dementia-Is history better than mental state? *Age and Ageing, 26,* 31–35.

Harwood, D. M. L., Hope, T., & Jacoby, R. (1997b). Cognitive impairment in medical inpatients. II: Do physicians miss cognitive impairment? *Age and Ageing, 26,* 37–39.

Hassinger, M., Smith, G., & La Rue, A. (1989). Assessing depression in older adults. In T. Hunt & C. J. Lindley (Eds.), *Testing older adults: A reference guide for geropsychological assessments.* Austin, TX: Pro-Ed.

Heaton, R. K. (1981). *Wisconsin Card Sorting Test manual.* Odessa, FL: Psychological Assessment Resources.

Heaton, R. K., Chelune, G. J., Talley, J. L., Kay, G. G., & Curtiss, G. (1993). *Wisconsin Card Sorting Test manual* (revised and expanded). Odessa, FL: Psychological Assessment Resources.

Heaton, R. K., Paulsen, J. S., McAdams, L. A., Kuck, J., Zisook, S., Braff, D., Harris, M. J., & Jeste, D. V. (1994). Neuropsychological deficits in schizophrenics: Relationship to age, chronicity, and dementia. *Archives of General Psychiatry, 51,* 469–476.

Hershey, L. A., Modic, M. T., Jaffe, D. F., & Greenough, P. G. (1986). Natural history of the vascular dementia: A prospective study of seven cases. *Canadian Journal of Neurological Sciences, 13,* 559–565.

Hertzog, C., Dixon, R. A., Hultsch, D. F. (1990). Relationships between metamemory, memory predictions, and memory task performance in adults. *Psychology of Agining, 5,* 215-227.

Hooper Visual Organization Test. (1983). Los Angeles: Western Psychological Services.

Hotchkiss A. P., & Harvey P. D. (1990). Effects of distraction on communication failures in schizophrenic patients. *American Journal of Psychiatry, 4,* 513–515.

House, J. S., Landis, K. R., & Umberson, D. (1988). Social relationships and health. *Science, 241,* 540–545.

Huff, F. J., Growdon, J. H., Corkin, S., & Rosen, T. R. (1987). Age at onset and rate of progression of Alzheimer's disease. *Journal of American Geriatric Society, 35,* 27–30.

Hughes, C., Berg, L., Danziger, W. L., Coben, L. A., & Martin, R. L. (1982). A new clinical scale for staging of dementia. *British Journal of Psychiatry, 140,* 566–572.

Ivnik, R. J., Malec, J. F., Smith, G. E., Tangalos, E. G., Petersen, R. C., Kokmen, E., & Kurland, L. T. (1992). Mayo's Older Americans Normative Studies: WAIS-R norms for ages 56-97. *Clinical Neuropsychologist, 6* (Suppl.), 1–30.

Ivnik, R. J., Smith, G. E., Malec, J. F., Petersen, R. C., & Tangalos, E. G. (1995). Long-term stability and inter-correlations of cognitive abilities in older persons. *Psychological Assessment, 7,* 155–161.

Jastak, S., & Wilkinson, G. (1984). *The Wide Range Achievement Test-revised.* Wilmington, DE: Jastak Associates.

Jenike, M. A. (1988). Depression and other psychiatric disorders. In M. S. Albert & M. Moss (Eds.), *Geriatric neuropsychology* (pp. 115–144). New York: Guilford Press.

Jeste, D. V. (1993). Late life schizophrenia: Editor's introduction. *Schizophrenia Bulletin, 19,* 687–689.

Jorm, A. F., Scott, R., & Jacomb, P. A. (1989). Assessment of cognitive decline in dementia by informant questionnaire. *International Journal of Geriatric Psychiatry, 4*(1), 35–39.

Jorm, A. F., & Jacomb, P. A. (1989). The Informant Questionnaire on Cognitive Decline in the Elderly (IQCODE): Socio-demographic correlates, reliability, validity and some norms. *Psychological Medicine, 19*(4), 1015–1022.

Juva, K., Sulkava, R., Erkinjuntti, T., Ylikoski, R., Valvanne, J., & Tilvis, R. (1994). Staging the severity of dementia: Comparison of clinical (CDR, DSM-III-R), functional (ADL, IADL) and cognitive (MMSE) scales. *Acta Neurologica Scandinavica, 90,* 293–298.

Kahn, R. L., & Byosiere, P. (1992). Stress in organizations. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology (*Vol. 3, 2nd ed, pp. 571–650). Palo Alto, CA: Consulting Psychologists Press.

Kaplan, E. (1988). A process approach to neuropsychological assessment. In T. Boll & B. K. Bryant (Eds.), *Clinical neuropsychology and brain function: Research, measurement, and practice.*Washington, DC: American Psychological Association.

Kasznaik, A. W., & Christensen, G. D. (1994). Differential diagnosis of dementia and depression. In M. Storandt & G. R.VandenBos (Eds.), *Neuropsychological assessment of dementia and depression in older adults: A clinician's guide.* Washington, DC: American Psychological Association.

Kasznaik, A. W., Poon, L. W., & Riege, W. L. (1986). Assessing memory deficits: An information-processing approach. In L. W. Poon (Ed.), *Clinical memory assessment of older adults.* Washington, DC: American Psychological Association.

Katzman, R., Brown, T., Fuld P., et al. (1983). Validation of a short orientation-memory-concentration test of cognitive impairment. *American Journal of Psychiatry, 140,* 734–739.

Katzman, R., Lasker, B., & Berstein, N. (1988). Advances in the diagnosis of dementia: Accuracy of diagnosis and consequences of misdiagnosis of disorders causing dementia. In R. D. Terry (Ed.), *Aging and the brain* (pp. 17–62). New York: Raven Press.

Keefe, R. S. E. (1995). The contribution of neuropsychology to psychiatry. *American Journal of Psychiatry, 152,* 6–15.

Kiloh, L. G. (1961). Pseudo-dementia. *Acta Psychiatrica Scandinavica, 37,* 336–351.

Kincaid, M. M., Harvey, P. D., Parrella, M., White, L., Putnam, K. M., Powchik, P., Davidson, M., & Mohs, R. C. (1995). Validity and utility of the ADAS-L for measurement of cognitive and functional impairment in geriatric schizophrenic inpatients. *Journal of Neuropsychiatry and Clinical Neurosciences, 7,* 76–81.

King, D. A., Cox, C., Lyness, J. M., Conwell, Y., & Caine, E. D. (1998). Quantitative qualitative differences in the verbal learning performance of elderly depressives and healthy controls. *Journal of International Neuropsychological Society, 4,* 115–126.

Koenig, H. G., & Blazer, D. G. (1992). Mood disorders and suicide. In J. E. Birren, R. B. Sloane, & G. D. Cohen (Eds.), *Handbook of Mental Health and Aging* (2nd ed., pp. 379–407). San Diego, CA: Academic Press.

Kumar, A., & Gottlieb, G. (1993). Frontotemporal dementias. *American Journal of Geriatric Psychiatry, 1,* 95–107.

Lafleche, G., & Albert, M. (1995). Executive functioning deficits in mild Alzheimer's disease. *Neuropsychology, 9,* 313–320.

Lamberty, G. J., & Bieliauskas, L. A. (1993). Distinguishing between depression and dementia in the elderly: A review of neuropsychological findings. *Archives of clinical Neuropsychology, 8,* 149–170.

Libon, D. J., Bogdanoff, B., Bibavuta, J. Skalina, S., Cloud, B. S., Resh, R., Cass, P., & Ball S. K. (1997). Dementia associated with periventricular and deep white matter alterations: A subtype of subcortical dementia. *Archives of Clinical Neuropsychology, 12*(3), 239–250.

Lezak, M. (1995). *Neuropsychological Assessment, third edition.* New York: Oxford University Press.

Luria, A. R. (1966). *Higher cortical functions in man.* New York: Basic Books.

Luria, A. R. (1973). *The working brain: An introduction to neuropsychology* (trans. B. Haigh). New York: Basic Books.

Marcopulos, B. A. (1989). Pseudodementia, dementia, and depression: Test differentiation. In T. Hunt & C. J. Lindley (Eds.), *Testing older adults: A reference guide for geropsychological assessments.* Austin, TX: Pro-Ed.

Mattis, S. (1976). Mental status examination for organic mental syndrome in the elderly patient. In L. Bellak & T. B. Karasu (Eds.), *Geriatric psychiatry.* New York: Grune & Stratton.

Mayeux, R., Ottman, R., Tang, M. X., NoboaBauza, L., Marder, K., Gurland, B., & Stern, Y. (1993). Genetic susceptibility and head injury as risk factors for Alzheimer's disease among community dwelling elderly persons and their first degree relatives. *Annals Neurology, 33,* 494–501.

Mohs, R. C., Rosen, W. G., Greenwald, B. S., & Davis, K. L. (1983). Neuropathologically validated scales for Alzheimer's disease. In T. Crook, S. Ferris, & R.

Bartus (Eds.), *Geriatric psychopharmacology* (pp. 37–45). New Canaan, CT: Mark Powley Associates.

Monsch, A. U., Bondi, M. W., Butters, N., Salmon, D. P., Katzman, R., & Thal, L. J. (1992). Comparisons of verbal fluency tasks in the detection of dementia of the Alzheimer type. *Archives of Neurology, 49,* 1253–1258.

Morris, J. C. (1993). The Clinical Dementia Rating (CDR): Current version and scoring rules. *Neurology, 43,* 2412–2414.

Morris, J. C., McKeel, D. W., Storandt, M., Rubin, E. H., Price, J. L., Grant, E. A., Ball, M. J., & Berg, L. (1991). Very mild Alzheimer's disease: Informant-based clinical, psychometric, and pathologic distinction from normal aging. *Neurology, 41,* 469–478.

Neale, J. M., & Oltmanns, T. F. (1980). *Schizophrenia.* New York: John Wiley.

Neary, D., Snowden, J. S., Mann, D. M. A., et al. (1990). Frontal lobe dementia and motor neuron disease. *Journal of Neurology Neurosurgery Psychiatry, 53,* 23–32.

Nebes, R. D., & Brady, C. B. (1992). Different patterns of cognitive slowing produced by Alzheimer's disease and normal aging. *Journal of Clinical and Experiemental Neuropsychology, 14,* 317–326.

Osterrieth, P. A. (1944). Le test de copie d'une figure complex: Contribution a l'etude de la perception et de la memoire. *Archives de Psychologie, 30,* 286–356.

Paquette, I., Ska, B., & Joanette, Y. (1995). Delusions, hallucinations, and depression in a population-based, epidemiological sample of demented subjects. In M. Bergener & S. I. Finkel (Eds.), *Treating Alzheimer's and other dementias.* New York: Springer Publishing Co.

Parkes, C. M. (1986). *Bereavement: Studies of grief in adult life.* Madison CT: International Universities Press.

Parkes, C. M., & Weiss, R. S. (1983). *Recovery from bereavement.* New York: Basic Books.

Parmalee, P. A., Lawton, M. P., & Katz, I. R. (1989). Psychometric properties of the Geriatric Depression Scale among the institutionalized aged. *Psychological Assessment, 1,* 331–338.

Pasquier, F., Lebert, F., Grymonprez, L., & Perit, H. (1995). Verbal fluency in dementia of the frontal lobe type and dementia of Alzheimer's disease. *Journal of Neurology Neurosurgery Psychiatry, 58,* 81–84.

Paveza, G. J., Cohen, D., Einsdorfer, C., et al. (1992). Severe family violence and Alzheimer's disease: Prevalence and risk factors. *Gerontologist, 32*(4), 493–497.

Powchik, P., Davidson, M., Nemeroff, C. B., Haroutunian, V., Purohit, D., Losonczy, M., Bissette, G. Perl, D., Ghanbar, H., Miller, B., & Davis, K. L. (1993). Alzheimer's disease related protein in geriatric, cognitively impaired schizophrenic patients. *American Journal of Psychiatry, 50,* 1726–1727.

Prohovnik, I., Dwork, A. J., Kaufman, M. A., & Wilson, N. (1993). Alzheimer's type neuropathology in elderly schizophrenia. *Schizophrenia Bulletin, 19,* 805–816.

Purohit, D. P., Davidson, M., Perl, D. P., Powchik, P., Haroutunian, V. H., Bierer, L. M., McCrystal, J., Losonczy, M., & Davis, K. L. (1993). Severe cognitive impairments in elderly schizophrenic patients: Clinicopathologic study. *Biological Psychiatry, 33,* 255–260.

Putnam, K. M., Harvey, P. D., Parrella, M. White, L. Kincaid, M., Powchik, P., & Davidson, M. (1996). Symptom stability in geriatric chronic schizophrenic inpatients: A one-year follow-up study. *Society of Biological Psychiatry, 39,* 92–99.

Rebok, G. W., & Folstein, M. F. (1993). Dementia. *Journal of Neuropsychiatry and Clinical Neurosciences, 5,* 265–276.

Reitan, R. M., & Davidson, L. A. (1974).*Clinical Neuropsychology: Current status and applications.* New York: Winston/Wiley.

Reitan, R. M., & Wolfson, D. (1993). *The Halstead-Reitan Neuropsychological Test Battery: Theory and clinical interpretation.* Tucson, AZ: Neuropsychology Press.

Rey, A. (1941). L'examen psychologique dans les cas d'encephalopathie traumatique. *Archives de Psychologie, 28,* 286–340.

Richie, K., Ledesert, B., & Touchon, J. (1993). The Eugeria study of cognitive ageing: Who are the 'normal' elderly? *International Journal of Geriatric Psychiatry, 8,* 969–977.

Rosen, J., & Zubenko, G. S. (1991). Emergence of psychosis and depression in the longitudinal evaluation of Alzheimer's disease. *Archives of Geriatric Gerontology, 6,* 225–233.

Rosen, W. G., Mohs, R. C., & Davis, K. L. (1984). A new rating scale for Alzheimer's disease. *American Journal of Psychiatry, 141,* 1356–1364.

Rowe, J. W., & Kahn, R. L. (1997). Successful aging. *Gerontologist, 37,* 433–440.

Rowe, J. W., & Kahn, R. L. (1987). Usual and successful aging. *Science, 237,* 143–148.

Russell, E. W., Neuringer, C., & Goldstein, G. (1970). *Assessment of brain damage: A neuropsychological key approach.* New York: Wiley-Interscience.

Salzman, C., & Nevis-Olesen, J. (1992). Psychopharmacologic treatment. In J. E. Birren, R. B. Sloane & G. D. Cohen (Eds.), *Handbook of mental health*

and aging (2nd ed., pp. 722–762). San Diego, CA: Academic Press.

Samuels, S. C., & Davis, K. L. (1998). Use of cognitive enhancers in dementing disorders. In J. C. Nelson (Ed.)., *Geriatric psychopharmacology* (pp. 381–403). New York: Marcel Dekker, Inc.

Schellenberg, G. D., Bird, T. D., Wijsman, E. M., Orr, H. T., Anderson, L., Nemens, E., Bonnycastle, L., Weber, J. L., Alonso, M. E., Potter, H., Heston, L. L., & Martin, G. M. (1992). Genetic linkage evidence for a familial Alzheimer's disease locus on chromosome 14. *Science, 258,* 668–671.

Silverman, J. M., Breitner, J. C. S., Mohs, R. C., & Davis, K. L. (1986). Reliability of the family history method in genetic studies of Alzheimer's disease and related dementia. *American Journal of Psychiatry, 143,* 1279–1282.

Smith, G. E., Bohac, D. L., Ivnik, R. J., & Malec, J. F. (1997). Using word recognition tests to estimate premorbid IQ in early dementia: Longitudinal data. *Journal of the International Neuropsychological Society, 3,* 528–533.

Snowdon, J. Validity of the Geriatric Depression Scale. (1990). *Journal of the American Geriatrics Society, 38,* 722-723.

Snowdon, J., & Donnelly, N. (1986). A study of depression in nursing homes. *Journal of Psychiatric Research, 20,* 327–333.

Spreen, O. & Strauss, E. (1998). *A compendium of neuropsychological tests: Administration, norms, and commentary,* (2nd ed.). New York: Oxford University Press.

Stebbins, G. T., Gilley, D. W., Wilson, R. S, et al. (1990). Effects of language disturbances on premorbid estimates of IQ in mild dementia. *The Clinical Neuropsychologist, 4,* 64–68.

Stebbins, G. T., Wilson, R. S., Gilley, D. W., et al. (1990). Use of the National Adult Reading Test to estimate premorbid IQ in dementia. *The Clinical Neuropsychologist, 4,* 18–24.

Stoudemire, A., Hill, C., Gulley, L. R., & Morris, R. (1989). Neuropsychological and biomedical assessment of depression-dementia syndromes. *Journal of Neuropsychiatry and Clinical Neurosciences, 1,* 347–361.

Swiercinsky, D. P. (1978). *Manual for the adult neuropsychological evaluation.* Springfield, IL: C.C. Thomas.

Teri, L., & Wagner, A. (1992). Alzheimer's disease and depression. *Journal of Consulting and Clinical Psychology, 60,* 379–391.

Thal, L. J., Grundman, M., & Klauber, M. R. (1988). Dementia: Characteristics of a referral population and factors associated with progression. *Neurology, 38,* 1083–1090.

Thompson, L. W., Gong, V., Haskins, E., & Gallagher, D. (1987). Assessment of depression and dementia during the late years. In K. W. Schaie (Ed.), *Annual review of gerontology and geriatrics.* New York: Springer.

Wechsler, D. (1987). *Wechsler Adult Intelligence Scale manual* (rev. ed.). San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1997a). *Wechsler Adult Intelligence Scale manual* (3rd ed.). San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1997b). *Wechsler Memory Scale manual* (3rd ed.). San Antonio,TX: The Psychological Corporation.

Weingartner, H. R., Eckart, M., Grafman, J., Molchan, S., Putnam, K., Rawlings, R., & Sunderland, T. (1993). The effects of repetition on memory performance in cognitively impaired patients. *Neuropsychology, 7,* 385–395.

Welsh, K. A., Butters, N., Hughes, J., Mohs, R. C., & Heyman, A. (1991). Detection of abnormal memory decline in mild cases of Alzheimer's disease using CERAD neuropsychological measures. *Archives of Neurology, 48,* 278–281.

Williams, J. M. (1997). The prediction of premorbid memory ability. *Archives of Clinical Neuropsychology, 12,* 745–756.

World Health Organization (1992). *International classification of diseases* (10th ed.). Geneva, Switzerland.

Yesavage, J. (1987) The use of self-rating depression scales in the elderly. In L. W. Poon (Ed.), *Handbook for clinical memory assessment of older adults.* Washington, DC: American Psychological Association.

Yesavage, J. A., Brink, T. L., Rose, T. L., Lum, O., Huang, V., Adey, M. B., & Leirer, V. O. (1983). Development and validation of a geriatric depression rating scale: A preliminary report. *Journal of Psychiatric Research, 17,* 37–49.

Zec, R. F. (1993). Neuropsychological functioning in Alzheimer's disease. In R. W. Parks, R. F. Zec, & R. S. Wilson (Eds.), *Neuropsychology of Alzheimer's disease and related disorders.* New York: Oxford University Press.

Zisook, S., DeVaul, R. A., & Glick, M. A. (1982). Measuring symptoms of grief and bereavement. *American Journal of Psychiatry, 139,* 1590–93.

Zisook, S., & Schucter, S. R. (1986). The first four years of widowhood. *Psychiatric Annals, 15,* 288–294.

Zubenko, G. S., Sullivan, P., Nelson J. P., et al. (1990). Brain imaging abnormalities in mental disorders of late life. *Archives of Neurology, 47,* 1107–1111.

Author Index

This Page Intentionally Left Blank

Subject Index

ABOUT THE EDITORS AND CONTRIBUTORS

THE EDITORS

Gerald Goldstein (Ph.D., University of Kansas, 1962) is Director of the Neuropsychology Research Program at the Highland Drive VA Medical Center in Pittsburgh, and Professor of Psychiatry and Psychology at the University of Pittsburgh. He has authored and coauthored numerous articles, chapters, and books in the area of clinical neuropsychology, which is his major research interest. He is Founding Editor of *Neuropsychology Review*, and has served on the editorial board of the *Journal of Clinical and Experimental Neuropsychology*, the *Journal of Psychopathology and Behavioral Assessment*, the *Archives of Clinical Neuropsychology* and *The Clinical Neuropsychologist*. He is a member of the American Board of Clinical Neuropsychology and Past President of the Division of Clinical Neuropsychology of the American Psychological Association.

Michel Hersen (Ph.D., State University of New York at Buffalo, 1966) is Professor and Dean, School of Professional Psychology, Pacific University, Forest Grove, Oregon. He is Past President of the Association for Advancement of Behavior Therapy. He has written four books, coauthored and coedited 114 books, including the *Handbook of Prescriptive Treatments for Adults* and *Single Case Experimental Designs*. He has also published more than 220 scientific journal articles and is coeditor of several psychological journals, including *Behavior Modification, Clinical Psychology Review, Journal of Anxiety Disorders, Journal of Family Violence, Journal of Developmental and Physical Disabilities, Journal of Clinical Geropsychology, and Aggression and Violent Behavior: A Review Journal*. With Alan S. Bellack, he is coeditor of the recently published 11-volume work: *Comprehensive Clinical Psychology*. Dr. Hersen has been the recipient of numerous grants from the National Institute of Mental Health, the Department of Education, the National Institute of Disabilities and Rehabilitation Research, and the March of Dimes Birth Defects Foundation. He is a Diplomate of the American Board of Professional Psychology, Distinguished Practitioner and Member of the National Academy of Practice in Psychology, and recipient of the Distinguished Career Achievement Award in 1996 from the American Board of Medical Psychotherapists and Psychodiagnosticians. Dr. Hersen has written and edited numerous articles, chapters and books on clinical assessment.

THE CONTRIBUTORS

Teresa A. Ashman (Ph.D., is Psychologist at Mount Sinai School of Medicine in the Department of Rehabilitation Medicine and the Project Coordinator of the Research and Training Center on Community Integration of Individuals with Traumatic Brain Injury. She is involved in ongoing studies of emotional and neurocognitive changes following traumatic brain injury as well as supervising both research and clinical postdoctoral interns and fellows. Dr. Ashman received her doctorate in clinical psychology from the New School for Social Research and completed a postdoctoral fellowship in health psychology at Memorial Sloan-Kettering Cancer Center. Prior to joining this department, Dr. Ashman was an Assistant Professor in the Department of Psychiatry at Mount Sinai School of Medicine. Her areas of specialization include neuropsychology of TBI and gerontology, PTSD, and other Axis I diagnoses following

health problems (specifically HIV, cancer, and TBI), along with clinical and research supervision.

Kristee A. Beres (Ph.D., Pacific Graduate School of Psychology, 1994) is a Clinical Psychologist in private practice in San Diego, California. She specializes in psychodiagnostic assessment and individual theraphy with children and adolescents. Dr. Beres is an independent consultant to Sharp Mesa Vista Hospital's APA accredited predoctoral internship for psychologist. She is responsible for supervising all of the psychological assessments performed by the interns.

Jane Holmes Bernstein (Ph.D., University of Edinburgh, 1973) is Director of the Neuropsychology Program in the Department of Psychiatry, Children's Hospital, Boston/Harvard Medical School and Associate Director (Clinical) of the Learning Disabilities Research Center at the Children's Hospital (Deborah P. Waber, Ph.D., Principal Investigator). Her primary professional interests are in the development of assessment models in developmental, as contrasted with child, neuropsychology and the training of clinicians in this area.

Amy Bohnert (Ph.D., Penn State University, 1999) recently completed her doctorate in clinical child psychology and is currently a postdoctoral fellow in the Developmental Psychopathology Research Training Program of the John F. Kennedy Center for Research on Human Development at Vanderbilt University. Her interests include emotional development in childhood and its relationship to psychopathology.

Patricia J. Brazil (Ph.D.), is Adjunct Assistant Professor in the Department of Medical Psychology at Oregon Health Sciences University, Portland, Oregon. In addition to the Structured Clinical Interview, Professor Brazil's publications and research interests focus on neuropsychology, chronic pain, substance abuse, and somatoform disorders.

James N. Butcher (Ph.D.) is currently Professor of Psychology in the Department of Psychology at the University of Minnesota. He graduated from Guilford College in North Carolina with a B.A. in psychology in 1960 and received an M.A. in experimental psychology from the University of North Carolina at Chapel Hill in 1962. He received a Ph.D. in clinical psychology at the University of North Carolina at Chapel Hill in 1964 and was awarded Doctor Honors Cause at Free University of Brussels 1990. He has maintained an active research program in the areas of: personality assessment, abnormal psychology, cross-cultural personality factors, and computer-based personality assessment. Dr. Butcher is a member of the University of Minnesota Press' MMPI Consultative Committee. Since 1982 the committee has been actively engaged in a large-scale project to revise and restandardize the MMPI. He is former Editor of *Psychological Assessment* and serves as consulting editor for numerous other journals in psychology and psychiatry. He has served on the Board of Trustees of the Society for Personality Assessment and the Executive Committee of Division 5 (Division of Measurement and Evaluation) of the American Psychological Association. Dr. Butcher founded the *Symposium on Recent Developments in the Use of the MMPI* in 1965 to promote and disseminate research information on the MMPI and has organized this conference series for the past 33 years. He also founded the *International Conference on Personality Assessment*, a program devoted to facilitating international research on personality assessment. Eleven international conferences have been held (Belgium, Denmark, Italy, Israel, Japan, Mexico, United States, and Norway). Dr. Butcher has been actively involved in developing and organizing disaster-response programs for dealing with human problems following airline disasters. He organized a model crisis intervention disaster response for the Minneapolis-St. Paul Airport; and organized supervised psychological services following two recent airline disasters: Northwest Flight 255 and Aloha Airlines' Maui incident.

Karen L. Dahlman (Ph.D., New School for Social Research, 1995) is currently Director of Psychology Training at The Mount Sinai School of Medicine where she is a Clinical Assistant Professor of Psychiatry. She directs the Neuropsychological Testing Service in the Geriatric Psychiatry and Geriatric Medicine divisions at Mount Sinai. She is also a member of the Extended Faculty at the New School for Social Research and Affiliate Professor of Psychology at St. John's University. Dr. Dahlman's primary area of research and practice is psychological assessment, with special interest in the neuropsychological assessment of memory disorders.

Craig Edelbrock (Ph.D., Oregon State University, 1976) is Professor of Human Development and Family Studies at Penn State University. His research and scholarly interests are developmental psychopathology and assessment and classification of child adjustement problems. He currently serves on the editorial boards of the *Journal of Child Psychology and Psychiatry*, and is Associate Editor of the *Journal of Abnormal Child Psychology*.

Philip Erdberg (Ph.D., University of Alabama, 1969), is a diplomate in clinical psychology of the American Board of Professional Psychology. He has served as a clinical and research consultant for a variety of therapeutic, educational, and correctional settings in northern California since 1971. He is a past president of the Society for Personality Asssessment and the 1995 recipient of the Society's Distinguished Contribution Award. Dr. Erdberg is driector of research at the Boyer House Foundation.

Miguel Perez Garcia (Ph.D.) received his doctorate degree from the Universidad of Granada (Spain) where he is currently Assistant Professor in the Departamento de Personalidad, Evaluacion, y Tratamiento Psicologico. He compldted a postdoctoral fellowship at the University of North Carolina at Wilmington. His research interests are cross-cultural neuropsychology and ecological validity in neuropsychological assessment.

Robert D. Gatewood (Ph.D., Purdue University) is Associate Dean for Academic Programs, Terry College of Business, University of Georgia. He has authored five textbooks and approximately 80 journal articles and conference papers on the topics of selection, organizational communication, individual reaction to job loss, and quality management.

Ross W. Greene (Ph.D., Virginia Ted, 1989) is Director of Cognitive-Behavioral Psychology at the Clinical and Research Program in Pediatric Psychopharmacology at Massachusetts General Hospital, where he specializes in the treatment of explosive/noncompliant children and adolescents and their families. He is also Assistant Professor of Psychology in the Department of Psychiatry at Harvard Medical School. His research interests include the classification, longitudinal study, and treatment of childhood disruptive behavior disorders and severe social impairment, and student-teacher compatibility. He is on the editorial boards of *Journal of Clinical Child Psychology* and *Journal of Psychoeducational Assessment.*

Jo-Ida C. Hansen (Ph.D., University of Minnesota, 1974) is Professor of Psychology, Director of the Center for Interest Measurement Research, Director of the Vocational Assessment Clinic, and Director of the Counseling Psychology Program at the University of Minnesota. Her current research interests include occupational health psychology, the role of serendipity in career development, the measurement of leisure interests, and the intersection of mid-life personality and vocational interests. She has authored many articles in Journals and has presented papers on vocational interest measurement at national and international meetings. She is coauthor of the Strong Interest Inventory.

Stephen N. Haynes (Ph.D., University of Colorado, 1971) is Professor at the University of Hawaii. Dr. Haynes has published numerous books and articles in the areas of behavioral assessment, psychophysiological disorders and psychopathology.

Jamie M. Joseph (Ph.D., Hofstra University, 1998) is a Licensed Clinical and Certified School Psychologist in New York. She completed a postdoctoral Clinical Fellowship in rational emotive behavior therapy at the Albert Ellis Institute where she is currently a Staff Psychologist and Clinical Supervisor. She works at the Smithtown Central School District where she conducts Psychological Evaluations and provides counseling to students and families. She is a Supervisor at the Psychological and Evaluation Research Clinic at Hofstra University and she provides cognitive behavioral psychotherapy in a private-practice setting on Long Island, New York. Historically, she has served as the Director or Psychological Services at the Whitestone School for Child Development, a school for special-needs children and she has taught as an Adjunct Instructor at Hofstra University. Her specific areas of research are in posttraumatic stress disorder, secondary traumatization, and the psychological effects that treating clients with HIV/AIDS has on mental health providers.

Lynda J. Katz (Ph.D. University of Pittsburgh) assumed the presidency of Landmark College in

July, 1994. Prior to that she held dual appointments at the University of Pittsburgh as Associate Professor of Psychiatry and Education in the School of Medicine, and Associate Professor of Health and Rehabilitation Sciences where she remains an Adjunct Professor. Dr. Katz obtained her Ph.D. in Rehabilitation Counseling/Psychology, as well as an M.Ed. in special education and rehabilitation counseling, and an M.S.W. in psychiatric social work, all from the University of Pittsiburgh. She has authored and coauthored countless reference articles, book chapters, and other publications in the areas of psychiatric rehabilitation, mental retardation, rights of the developmentally disabled, vocational assessment, achievement testing, learning disabilities, and attention deficit hyperactivity disorder. Dr. Katz has presented her research findings at seminars internationally.

Alan S. Kaufman (Ph.D., Columbia University, 1970) is a Clinical Professor of Psychology at Yale University School of Medicine.

Glenn J. Larrabee (Ph.D., Bowling Green State University, 1981), a Diplomate in Clinical Neuropsychology, American Board of Professional Psychology, is currently engaged in the full-time independent practice of clinical neuropsychology. He is a faculty associate of the Center for Neuropsychological Studies at the University of Florida, and a member of the medical staff at Sarasota Memorial Hospital. He is a coauthor of the Continuous Visual Memory Test and has published extensively in the areas of memory assessment, memory disorders of aging, and clinical and forensic neuropsychological assessment. A member of the editorial board of several journals, he also consults with major test-publishing companies.

Richard C. Mohs (Ph.D.) is Professor and Vice Chairman of the Department of Psychiatry at the Mount Sinai School of Medicine in New York City, and Asssociate Chief of Staff for Research at the Bronx Veterans Affairs Medical Center. Dr. Mohs received his doctoral degree in psychology from Stanford University and completed fellowship training in psychopharmacology at the Stanford University School of Medicine. His research has been supported by the National Institute on Aging, the Department of Veterans Affairs, the John D. and Catherine T. MacArthur Foundation, the Charles A. Dana Foundation, and the Institute for the Study of Aging.

Robert W. Motta (Ph.D.) is Professor of Psychology and received his Doctorate in Psychology from Hofstra University in 1975. He possesses a Diplomate from APA's American Board of Professional Psychology (ABPP), with a specialization in Behavioral Psychotherapy. He is presently on the examining Board of ABPP, is a licensed clinical psychologist, a certified school psychologist, a certified child abuse identification trainer, and is listed in the National Register of Health Service Providers in Psychology. Dr. Motta is past president and secretary treasurer of the School Psychology Division of the New York State Psychological Association. He has published over 70 papers and book chapters on adjustment disorders of children and is on the Editorial Advisory Boards of a number of professional journals. His specific areas of research are in Posttraumatic Stress Disorder and in psychological effects of physical exercise. Dr. Motta is the Director of the Psy.D. Program in School-Community Psychology and is the former director of another Psychology Ph.D. Program at Hofstra. The Psy.D. is a doctoral degree for psychological practitioners. Graduates of the Psy.D. program are trained to work in facilities such as schools, community mental-health centers, drug-treatment programs, programs for the physically challenged, psychiatric facilities, homes for the aged, and so on. Program graduates are also trained to provide individual psychological services.

Elahe Nezami received her B.A. from the University of Massachusetts in 1981. She has an M.A. degree in clinical and counseling psychology from the University of Houston. Her second M.A. degree and her Ph.D. in clinical psychology, completed in 1993, are from the University of Southern California.

Thomas H. Ollendick (Ph.D., Purdue University, 1971) is University Distinguished Professor and Director of the Child Study Center at Virginia Tech. His research and teaching interests center on social-learning theory and the assessment, treatment, and prevention of child behavior disorders. An editorial board member of 13 journals, Dr. Ollendick is editor of the *Journal of Clinical Child Psychology* (1997–2001), and is coeditor of *Clinical Child and Family Psychology Review*. He is the author of numerous research publications and the

coauthor or coeditor of 21 books including *Clinical Behavior Therapy with Children and Child Behavioral Assessment.* A Fellow of the American Psychological Association, he is Past President of the Clinical Child Psychology Section (Division 12, Section I), and is President of Division 12 (Clinical Psychology). He is also a Past President of the Association for Advancement of Behavior Therapy.

Mitchel D. Perlman, Ph.D., received his doctoral degree (Clinical Psychology) in 1986 from California School of Professional Psychology–San Diego. His teaching and professional interest lay in psychodiagnostic assessment encompassing cognitive, emotional, and forensic domains. Integrating his research experience with his therapeutic expertise, Dr. Perlman sits on the advisory board of Alvarado Parkway Institute Psychiatric Hospital for Children & Adolescents.

Evelyn Perloff has done research, writing, consulting, and teaching in testing, evaulation, and measurement for the U.S. government, research firms, organizations, and academia. Director of Behavioral Measurement Database Services, she provides on-line and CD-ROM access, via Ovid Technologies, to 60,00 records on instruments in the health and psychosocial sciences.

Robert Perloff (Ph.D.) is Distinguished Service Professor Emeritus of Business Administration and of Psychology at the University of Pittsburgh, and is a former President of the American Psychological Association and of the American Evaluation Association. He has published, taught, consulted, and conducted research in organizational behavior, consumer psychology, measurement, and evaluation.

Aurelio Prifitera (Ph.D.) is Vice President in charge of the development of the clinical measurement scales. He has been with The Psychological Corporation since 1985. Formerly, he was a staff psychologist and faculty member at Northwestern University Medical Center. He also served as staff psychologist at Highland Hospital, Asheville, North Carolina, and was appointed adjunct faculty at Duke University Medical School. Dr. Prifitera received his Doctorate in clinical psychology from Loyola University of Chicago and Master's degrees from the University of Chicago and the University of Illinois, Champaign.

Antonio E. Puente (Ph.D.) received his doctoral degree from the University of Georgia. He is currently Professor of Psychology at the University of North Carolina at Wilmington. He maintains an active practice in clinical neuropsychology. Dr. Puente, a native Cuban, is Past President of the National Academy of Neuropsychology, represents APA's neuropsychology Division on the APA Council of Representatives, and is Editor-in-Chief of *Neuropsychology Review.*

Michael C. Ramsay (Ph.D. candidate) majors in Research, Measurement, and Statistics in the Department of Educational Psychology at Texas A&M University, College Station. He maintains an ongoing research program at the university's Department of Educational Curriculum and Instruction, where he develops innovative methodologies and data collection techniques. He has co-directed large-scale research projects at the university's Educational Research and Evaluation Laboratory and Measurement and Research Services. He has consulted with numerous clients, taught statistics and educational psychology, reviewed for several journals and conferences, conducted workshops on advanced statistical techniques and computer applications, and authored or coauthored many scholarly publications. Research interests include bilingual assessment, analysis of psyhometric tests, multivariate evaluation of medical research, and intellectual characteristics of ethnic groups.

Mark D. Reckase (Ph.D., Syracuse University, 1972) is Professor of Education in the Measurement and Quantitative Methods Area within the Counseling, Educational Psychology, and Special Education Department in the College of Education at Michigan State University. Prior to joining the faculty of the Michigan State University, he was the Assistant Vice President for Assessment Innovations at ACT, Inc., where he was responsible for designing and implementing new assessment procedures. He has authored and coauthored many articles and book chapters on psychometric theory and the development of assessment procedures for educational applications.

Daniel J. Reschly (Ph.D., University of Oregon, 1971) is Professor of Education and Psychology, Peabody College, Vanderbilt University. Dr. Reschly has authored numerous articles on psychoeducational assessment, disproportionate minor-

ity placement in education programs, school psychology professional practices, and legal issues. He edited the *School Psychology Review* and served as President of the National Association of School Psychologist. Dr. Reschly's recent work has focused on assessment and interventions with children and youth with disabilities, particularly assessment with minority children and youth.

Cecil R. Reynolds (Ph.D., ABPN, ABPP) earned his doctoral degree from the University of Georgia in 1978 under the tutelage of Dr. Alan S. Kaufman, with a major in school psychology and minors in statistics and clinical neuropsychology. He served an internship divided between the Medical Collage of Georgia and the Rutland Center for Severely Emotional Disturbed Children. Prior to joining the Texas A&M University School Psychology faculty in 1981, Dr. Reynolds was a faculty member at the University of Nebraska-Lincoln, where he served as Associate Director and Acting Director of the Buros Institute of Mental Measurement, after writing the grants and proposals to move the Institute to Nebraska following the death of its founder, Oscar Buros. His primary reserach interests are in all aspects of psychological assessment with particular emphasis on assessment of memory, emotional and affective states and traits, and issues of cultural bias in testing. He is the author of more than 300 scholarly publications and author or editor of 24 books including the *Handbook of School Psychology,* the *Encyclopedia of Special Education,* and the *Handbook of Clinical Child Neuropsychology.* He is the author of several widely used tests of personality and behavior including the Behavior Assessment System for Children and the Revised Children's Manifest Anxiety Scale. He is also senior author of the Test of Memory and Learning and coauthor of several computerized test-interpretation systems.

Dr. Reynolds holds a diplomate in Clinical Neuropsychology from the American Board of Professional Neuropsyhology, of which he is also past president. He is a diplomate in School Psychology of the American Board of Professional Psychology and is a diplomate of the American Board of Forensic Examiners. He is a past president of the National Academy of Neuropsychology and of APA Division 5 (Evaluation, Measurement, and Statistics). He is the current Immediate Past president of the APA Division of Clinical Neuropsychology (40). Dr. Reynolds teaches courses primarily in the areas of psychological testing and diagnosis and neuropsychology in addition to supervising clinical practica in testing and assessment. He is Editor in Chief of *Archives of Clinical Neuropsychology,* the official journal of the National Academy of Neuropsychology, and serves on the editorial boards of 11 other journals in the field. Dr. Reynolds has received multiple national awards recognizing him for excellence in research including the Lightner Witmer Award and the early career awards from Division 5 and 15. In 1999, he received the Div. 16 Senior Scientist Award. He is a co-recipient of the Society for the Psychological Study of Social Issues Robert Chin Award and a MENSA best research article award. His service to the profession and to society has been recognized as well through the President's Gold Medal for Service to the National Academy of Neuropsychology and the University of N. C. at Wilmington Razor Walker Award (received in 1999). He is a Professor of Educational Psychology and Distinguished Research Scholar in the College of Education at Texas A&M University and a charter member of the TAMU Faculty of Neuroscience.

Carol Robinson-Zañartu (Ph.D., University of Pittsburgh, 1981) is Professor of School Psychology, Department of Counseling and School Psychology, San Diego State University. Dr. Robinson-Zañartu has authored numerous articles and professional papers on educational equity, culturally appropriate assessment, and Native American educational issues. She served as president of the California Association for Mediated Learning. Dynamic assessment and mediated learning interventions, especially across cultures, are specific areas of her research.

Shawn Christopher Shea (M.D.) is the former Director of the Diagnostic and Evaluation Center at the Western Psychiatric Institute and Clinic, Pittsburgh, Pennsylvania. A nationally acclaimed workshop leader, Dr. Shea has been a recipient of an Outstanding Course Award presented by the American Psychiatric Association for his presentations at their annual meetings. He has presented at the Cape Cod Symposia, the Santa Fe Symposia, and the McMaster Muskoka Seminars. He is the author of *Psychiatric Interviewing: the Art of Understanding,* 2nd edition, and *The Practical Art of Suicide Assessment.* Dr. Shea is the Director of The Training Institute for Suicide Assessment and Clinical Interviewing, a training and consultation

service providing workshops, consultations, and quality assurance design in mental health assessments. He is also an Adjunct Assistant Professor of Psychiatry at the Dartmouth Medical School and in private practice.

Gregory T. Slomka (Ph.D., University of Pittsburgh, 1986) is an Assistant Professor of Psychiatry at the University of Pittsburgh, School of Medicine in Pittsburgh, Pennslyvania. He was previously Clinical Coordinator of Assessment Services within Neuropsychological Assessment and Rehabilitation Service at Western Psychiatric Institute and Clinic, Pittsburgh, Pennsylvania. His major interests lie in developmental neuropsychology and the application of neuropsychology testing in multiply disabled populations. He has authored a number of chapters on the neuropsychological assessment of children and the developmentally disabled.

David S. Tulsky (Ph.D.) received his doctoral degree in clinical psychology from the University of Illionis at Chicago, where he specialized in clinical psychology and psychometrics. He joined The Psychological Corporation in 1992 and lead the revision of the Wechsler Adult Intelligence Scale, Third Edition. Currently, he is a Senior Project Director in the Psychological Measurement Group and is the manager of the team that is responsible for developing intelligence scales. He is also an instructor in the graduate program in school psychology at Trinity University. Previous to joining the Psychological Corporation, Dr. Tulsky worked as a reseracher studying quality of life in cancer patients and as a practicing psychologist at the Rush Cancer Center of Rush Presbyterian St. Luke's Medical Center in Chicago.

Michael D. Weiler (Ph.D., University of Rhode Island, 1992) is Senior Neuropsychologist with the Learning Disabilities Research Center, Children's Hospital, Boston and an Instructor in Psychology, Department of Psychiatry, Harvard Medical School. His research interests focus on disorders of attention and learning, and their remediation, in children.

Arthur N. Wiens (Ph.D.) is Professor Emeritus of Medical Psychology and Head, Division of Clinical Psychology, Oregon Health Sciences University. He has a long-time interest in research on the interview and in teaching interviewing skills. He coauthored *The Interview: Research on its Anatomy and Structure* and *Non-Verbal Communication: The State of the Art.*

Jianjun Zhu (Ph.D.) is currently a project director in the Psychological Corporation and teaches at Trinity University. He received his bachelor's and master's degrees from Beijing. Normal University, and his doctoral degree from the University of Texas at Austin. He was one of the project directors directing the WAIS-III revision.

This Page Intentionally Left Blank